Étude de la dynamique d'encrassement d'un réseau d'assainissement : méthodologie de traitement et d'analyse de données de capteurs

Ali Shakil

Centrale Méditerranée

NNT : 2023AIXM0001

THÈSE DE DOCTORAT

Soutenue à Centrale Méditerranée
le 06/02/2024 par

Ali Muhammad SHAKIL

Étude de la dynamique d'encrassement d'un réseau d'assainissement : méthodologie de traitement et d'analyse de données de capteurs

Discipline
Physique et Sciences de la Matière

Spécialité
Optique, Photonique et Traitement d'Image

École doctorale
ED 352 Physique et Sciences de la Matière

Laboratoire/Partenaires de recherche
Institut Fresnel
Institut de Mathématiques de Marseille
SUEZ
Seramm

Composition du jury

Pr. Pascal CHARGÉ Polytech Nantes	Rapporteur
Pr. Éric MATZNER-LOBER Université Rennes 2	Rapporteur
Pr. Gilles FAY Centrale Supelec	Président du jury
Dr. Charlotte SAKAROVITCH SUEZ EAU FRANCE	Examinatrice
Cyril Leclerc SUEZ EAU FRANCE	Examinateur
Dr. Mohammad-Ali KHALIGHI Centrale Méditerranée	Directeur de thèse
Pr. Pierre PUDLO Aix-Marseille Université	Co-directeur de thèse

INSTITUT FRESNEL MARSEILLE
INSTITUT de MATHÉMATIQUES de MARSEILLE
suez
Seramm

Affidavit

Je soussigné, Ali Muhammad Shakil, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique Mohammad-Ali Khalighi et de Pierre Pudlo, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 27/11/2023

Liste de publications et participation aux conférences

Liste des publications et/ou brevets réalisées dans le cadre du projet de thèse :

- Ali Shakil, Mohammad-Ali Khalighi, Pierre Pudlo et al. « Outlier detection in non-stationary time series applied to sewer network monitoring ». Dans : *Elsevier - Internet of Things* 21 (2023), p. 100654.
- Ali Shakil, Pierre Pudlo, Mohammad-Ali Khalighi et al. « Efficient Low-Complexity Data Clustering Based on Unsupervised Machine Learning for a Sewer Network ». En cours de préparation.

Participation aux conférences et écoles d'été au cours de la période de thèse :

- Ali Shakil, Charlotte Sakarovitch, Cyril Leclerc et al. « Une solution d'optimisation de la performance environnementale par le digital ». Congrès ASTEE - Association Scientifique et Technique pour l'Eau et l'Environnement, Nice, France, 5 juin au 8 juin 2023.
- Ali Shakil, Charlotte Sakarovitch, Cyril Leclerc et al. « Smart Sewer Network Monitoring in Marseille (France) : Drain Waste Accumulation Modelling and Analysis ». 11th IWA - International Water Association - Conference Efficient, Bordeaux, France, 13 septembre au 15 septembre 2023.

Résumé

Cette thèse s'inscrit dans le cadre d'un projet mené par le service d'assainissement de Marseille Métropole (SERAMM), une filiale de Suez, sur la digitalisation du réseau d'assainissement de la ville de Marseille. Les éléments essentiels de ce réseau sont les "avaloirs" qui ont le rôle d'absorber les eaux pluviales. Ces avaloirs ont besoin d'une maintenance permanente en raison des problèmes d'encrassement qui peuvent entraîner des inondations, des dommages aux équipements et la pollution de l'environnement. Ce travail a pour objectif d'étudier un ensemble d'avaloirs connectés grâce à des capteurs de mesure du niveau d'encrassement, afin de rendre l'opération de maintenance plus efficace. Pour ce faire, nous souhaitons comprendre la dynamique globale d'accumulation des déchets dans les avaloirs en analysant les données recueillies par ce réseau de capteurs.

Cette tâche, au premier regard simple, s'est avérée très complexe. En effet, les premières analyses des données révèlent une diversité importante dans la dynamique d'encrassement, avec des augmentations ou des diminutions progressives ou brusques. De plus, cette dynamique est influencée par des éléments contextuels tels que la proximité d'arbres ou de certains commerces, ou encore la pluie.

Dans un premier temps, notre étude a consisté en un prétraitement des données collectées, incluant notamment l'élimination des redondances et la détection des anomalies. Ces dernières se manifestent sous forme de pics dans les données qu'il est nécessaire de détecter et de supprimer. Dans un second temps, nous avons ensuite utilisé des algorithmes d'intelligence artificielle afin de regrouper les avaloirs selon leurs comportements et d'identifier des clusters d'avaloirs ayant des dynamiques distinctes.

Pour approfondir notre compréhension de la dynamique d'encrassement, nous avons ensuite examiné l'impact des éléments contextuels sur les comportements d'encrassement établis. Après avoir identifié ces éléments, ainsi que les données qui leur sont associées, nous avons analysé leur influence sur différentes catégories d'avaloirs en nous basant sur des méthodes statistiques, soit de manière bivariée, en étudiant chaque facteur individuellement, soit de manière multivariée, en tenant compte de l'ensemble des facteurs contextuels.

Mots clés : Réseau d'assainissement, Intelligence Artificielle, Analyse de données, Détection d'anomalies, Clustering, Capteurs.

Abstract

This thesis is part of a project led by the Sanitation Service of Marseille Métropole (SERAMM), a subsidiary of Suez, focusing on the digitalization of Marseille's sewer network. An essential element of this network is the "storm drain" designed to absorb rainwater. These drains require constant maintenance due to concerns over waste accumulation, which can lead to flooding, equipment damage, and environmental pollution. The aim of this study is to examine a group of storm drains equipped with sensors that measure waste levels to enhance maintenance efficiency. The objective is to understand the overall dynamics of waste accumulation in these drains by analysing the data collected from this network of sensors.

This task, initially seeming straightforward, turned out to be highly complex. Indeed, our initial data analysis uncovered a significant variety in the dynamics of waste accumulation, characterized by gradual or sudden increases and decreases. Moreover, this dynamic is impacted by contextual factors such as the proximity to trees or various commercial establishments, as well as rainfall.

We started our study by preprocessing the collected data, which included eliminating redundancies and detecting anomalies. These anomalies, which appeared as peaks in the data, were identified and removed. Subsequently, we used artificial intelligence algorithms to categorize the storm drains based on their behavior and to identify clusters of drains with distinct dynamics.

To deepen our understanding of these dynamics, we investigate the impact of exogenous factors on the established waste accumulation categories. After identifying these factors and the relevant data, we analyzed their influence on different storm drain categories using statistical methods. This includes bivariate analysis for assessing each factor individually, and multivariate analysis, by considering all contextual factors collectively.

Keywords: Sanitation Network, Artificial Intelligence, Data Analysis, Anomaly Detection, Clustering, Sensors.

Remerciements

Je remercie tout d'abord les membres du jury d'avoir fait le déplacement pour assister à ma soutenance de thèse, pour les échanges pertinents ainsi que pour leur engagement. Un remerciement particulier à Pascal Chargé et Éric Matzner-Lober pour la qualité de leurs retours sur le manuscrit. Un grand merci également à Gilles Fay pour avoir accepté de présider le jury.

Je remercie chaleureusement Mohammad-Ali Khalighi et Pierre Pudlo pour leur encadrement et leurs conseils tout au long de cette aventure académique.

Je souhaite exprimer ma plus profonde reconnaissance envers le LyRE, et plus particulièrement Charlotte Sakarovitch et Cyril Leclerc, qui m'ont accompagné et soutenu quotidiennement tout au long de mes trois années de thèse.

Je tiens ensuite à remercier en particulier Dominique Laplace, qui a porté le projet des avaloirs connectés et qui a été pour moi un mentor. À la suite de son départ à la retraite, je tiens également à exprimer ma gratitude envers Laura Pischedda qui a assuré la continuité.

Je remercie bien sûr la métropole de Marseille qui a financé tout le projet et sans qui cette thèse n'existerait pas.

J'adresse ensuite des remerciements plus personnels à mes amis Timothée Justel, Chiheb Daaloul, pour leur support précieux, à mes collègues du SERAMM, de l'Institut Fresnel et de l'équipe DATA du LyRE, que je ne peux pas tous citer.

Finalement, je tiens à exprimer ma gratitude envers ma famille, qui m'a constamment soutenu et encouragé, en particulier mes parents. Ils ont insufflé en ma sœur et moi l'importance du travail et ont déployé tous les efforts nécessaires pour nous offrir les meilleures chances. Je leur serai toujours reconnaissant.

Table des matières

Table des figures

Acronymes

1SE One Standard Error.

ABA Adjusted Balanced Accuracy.
ANOVA Analysis Of Variance.
ARI Adjusted Rand Index.
ARIMA Auto-Regressive Integrated Moving Average.

BA Balanced Accuracy.

DBSCAN Density-Based Spatial Clustering of Application with Noise.

EM Expectation Maximization.

FPR False Positive Rate.

GMM Gaussian Mixture Model.

I2M Institut de Mathématique de Marseille.
IA Intelligence artificielle.
IdO Internet des Objets.
IRIS Îlots Regroupés pour l'Information Statistique.

LASSO Least Shrinkage and Selection Operator.
LPWAN Low Power Wide Area Network.

.

OVD Opposite Variation Detection.

PPZ Peak Pattern based Z-score.

RIDGE Regression Isotropic Distributed Gaussian Estimate.

ROC Receiver Operating Characteristic.

RR Risque Relatif.

SERAMM Service d'Assainissement de Marseille Métropole.

SIG Système d'Information Géographique.

TPR True Positive Rate.

WCSS Within-Cluster Sum of Squares.

Introduction

À l'ère de l'urbanisation rapide, les infrastructures urbaines comme les réseaux d'assainissement deviennent des composants critiques, et sont confrontées à des enjeux majeurs, que ce soit pour la gestion de la ressource en eau, la valorisation des déchets ou bien protection de l'environnement [1]. Ces réseaux d'assainissement sont particulièrement complexes et nécessitent un entretien et une surveillance constants, afin d'assurer le bien-être et la sécurité des populations urbaines. Il est donc particulièrement important de développer de nouvelles solutions afin de garantir la qualité de vie des usagers et de protéger l'environnement. Une approche efficace est d'utiliser le concept d'Internet des Objets (IdO, ou Internet of Things, IoT), qui consiste en un réseau d'objets physiques connectés, dotés de capacités de communication et d'échange de données [2]. En utilisant des technologies de connexion sans-fil, ces objets peuvent collecter, analyser et partager des informations en temps réel [3]. L'IdO et le déploiement des objets connectés offre de nombreuses perspectives pour améliorer considérablement l'efficacité des systèmes mais aussi réduire leur consommation énergétique. À titre d'exemple, on peut citer l'utilisation du concept d'IdO dans la gestion du trafic routier dans les villes intelligentes [4, 5] ; la surveillance médicale [6] ; et le suivi des chaînes d'approvisionnement et de production dans l'industrie [7]. L'IdO a également été utilisé pour les réseaux d'eau et d'assainissement, par exemple pour la surveillance de la qualité de l'eau [8], la détection des gaz dangereux dans les réseaux [9], ou la détection des conduits endommagés pour prévenir les fuites d'eau [10]. L'IdO peut également contribuer à la protection de l'environnement en permettant la surveillance des milieux naturels, ou une gestion plus efficace des ressources naturelles. Dans un contexte de changement climatique, l'optimisation de la consommation de ressources et le traitement des déchets revêtent une importance accrue [11].

En combinant l'IdO et l'intelligence artificielle (IA), de nouveaux systèmes dits intelligents peuvent être déployés, capables de s'adapter aux conditions environnementales réelles, permettant ainsi une gestion plus efficace [12, 13]. Des exemples de ces systèmes intelligents peuvent être retrouvés pour la gestion des réseaux électriques [14] ou la surveillance des nappes phréatiques [15, 16]. Dans le cadre de la maintenance des réseaux d'assainissement, on peut citer la détection et la

prévision des inondations urbaines après de fortes pluies [17].

Aujourd'hui, la maintenance des réseaux d'eaux urbains se complexifie dans les métropoles en raison de la croissance des populations et de l'expansion territoriale qui en découle. Ces métropoles doivent faire face à une demande croissante en matière d'eau potable et de services d'assainissement pour répondre aux besoins de leur population, tout en gérant des infrastructures d'approvisionnement et de traitement de l'eau vieillissantes. Cette pression croissante sur les infrastructures hydrauliques urbaines exige des solutions innovantes pour garantir une gestion efficace et durable. Dans ce contexte, l'optimisation de la maintenance des systèmes d'assainissement devient une priorité. Les avaloirs, plus communément appelés "bouche d'égout", sont des éléments clés de cette infrastructure car ils permettent l'absorption des eaux pluviales. La maintenance de ces avaloirs est essentielle car ils sont constamment exposés à l'accumulation de déchets et de feuillages. De plus, certains facteurs comme la pluie ou la présence d'éléments sources de déchets exacerbent ce phénomène en entraînant des débris variés dans les avaloirs tels que des mégots, des emballages, des canettes, etc. Ces facteurs contribuent non seulement à la saturation rapide des avaloirs mais aussi à l'augmentation des défis liés à leur maintenance.

La métropole de Marseille, cas d'étude de cette thèse, est particulièrement concernée par ce problème d'encrassement des avaloirs, pouvant engendrer plusieurs types de nuisances. Parmi celles-ci, on compte les nuisances visuelles (lorsque les déchets débordent des avaloirs), les nuisances hydrauliques (qui peuvent conduire à des inondations locales en cas de fortes pluies), ainsi que les nuisances matérielles et environnementales résultant de l'endommagement des équipements d'assainissement et du rejet de déchets en mer. Historiquement, la maintenance des avaloirs à Marseille a été gérée sous forme d'inspections visuelles et d'interventions de nettoyage basées sur des quotas annuels. Évidemment, cette approche présentait des limites en termes d'efficacité car elle entraînait des déplacements inutiles.

Dans une démarche d'optimisation et d'efficacité accrue, une nouvelle approche a été adoptée très récemment, qui consiste à installer des capteurs pour surveiller à distance et en temps réel le niveau d'encrassement des avaloirs. Cette innovation vise à permettre des interventions plus ciblées et plus efficaces, en intervenant au bon endroit et au bon moment. Le projet d'installation de capteurs pour la gestion des avaloirs à Marseille a débuté en 2019 et s'est accéléré à partir de 2020 avec le déploiement de 1200 nouveaux capteurs. L'objectif du projet est d'atteindre un réseau de 5000 avaloirs équipés de ces dispositifs d'ici 2024, année de l'organisation des Jeux olympiques où Marseille accueillera entre autres des épreuves nautiques.

L'objectif de cette thèse est d'aller au-delà du simple suivi de mesures du niveau d'encrassement d'un avaloir. L'ambition est d'analyser les données mesurées par ces capteurs afin de mieux comprendre, de manière plus globale, la dynamique

d'encrassement des avaloirs. Il s'agit d'identifier les principaux comportements, de détecter des motifs de remplissage des avaloirs, et d'identifier les facteurs contextuels exogènes pouvant influencer ou causer cette dynamique.

Ce manuscrit de thèse est structuré en quatre parties :

- Le premier chapitre établit les bases de l'étude, présentant le contexte global du réseau d'assainissement de Marseille, les avaloirs qui le composent, et plus globalement, le dispositif étudié.
- Le chapitre 2 se concentre sur les données collectées et leur prétraitement. Il donne une description détaillée des problématiques liées à ces données et décrit les méthodologies adoptées pour les traiter avec notamment la mise en place d'un algorithme de détection d'anomalies spécifiquement conçu pour identifier des "pics" dans les données.
- Le troisième chapitre se consacre à l'analyse de la dynamique d'encrassement des avaloirs, utilisant des techniques de classification non supervisées pour regrouper les avaloirs selon leurs comportements et les proportions qu'ils représentent. La méthodologie employée dans ce chapitre permet d'explorer les choix d'algorithmes, d'attributs et d'hyperparamètres, tout en minimisant le nombre de combinaisons pour faciliter la visualisation et l'interprétation des résultats.
- Enfin, le quatrième chapitre explore l'impact des éléments contextuels sur la dynamique d'encrassement. Après avoir identifié les éléments d'influence et collecté les données associées, nous analysons comment ces facteurs affectent les différentes catégories d'avaloirs établies précédemment, en se basant sur des méthodes statistiques, soit de manière bivariée, en étudiant chaque facteur individuellement, soit de manière multivariée, en tenant compte de l'ensemble des facteurs contextuels, pour prédire l'appartenance d'un avaloir à un groupe spécifique.

Chapitre 1

Réseau d'assainissement et avaloir connecté

Sommaire

FIGURE 1.1 – Exemples d'avaloirs

1.1 Contexte

L'*assainissement* désigne les activités liées au traitement des eaux usées (eaux résiduelles issues des activités humaines : domestiques, industrielles, etc.) et à l'évacuation des eaux pluviales. Ceci ne comprend pas l'approvisionnement en eau potable. Le réseau d'assainissement de Marseille est géré par SUEZ à travers sa filiale SERAMM (Service d'Assainissement de Marseille Métropole). SUEZ travaille sur la digitalisation du réseau afin de réduire la quantité de déchets urbains rejetés en mer. Plus précisément, ce sont ses filiales SERAMM et LyRE (centre de Recherche et d'Expertise de SUEZ) qui sont responsables de ce projet.

Le sujet de cette thèse découle du SERAMM, responsable du réseau d'assainissement de Marseille. La maintenance de ce réseau inclut en particulier l'entretien des *avaloirs*, plus communément appelés "bouches d'égout". Ces ouvertures sont généralement le long du trottoir et servent à absorber les eaux de pluie. Quelques exemples d'avaloirs sont présentés en Figure 1.1. La maintenance du réseau d'avaloirs est essentielle car ils absorbent régulièrement les divers déchets et feuillages présents en surface. Cet "encrassement" entraîne les divers types de nuisances suivants :

- les nuisances visuelles, lorsque les déchets présents dans l'avaloir débordent en surface ;
- les nuisances hydrauliques, quand l'évacuation est obstruée, empêchant le bon écoulement des eaux de pluie et entraînant des inondations locales en cas de fortes pluies ;
- les nuisances matérielles, lorsque les déchets engouffrés dans le réseau endommagent les équipements qui s'y trouvent (pompes, dégrilleurs [1], etc.) et entraînent une surcharge pour la station d'épuration ;
- les nuisances environnementales, quand les déchets sont rejetés en mer, que ce soit via le réseau lui-même (lorsque l'avaloir est directement relié à la mer) ou par débordement de l'avaloir en période de pluie, qui cause un déversement des déchets en milieu urbain et *in fine* en milieu naturel.

1. Système destiné à filtrer les déchets dans un ouvrage hydraulique.

La Figure 1.2 illustre l'encrassement des avaloirs et les différentes nuisances qui en résultent à travers des photographies prises à Marseille.

FIGURE 1.2 – Exemples d'encrassement subi des avaloirs et des nuisances qui en résultent.

Ces nuisances sont particulièrement délétères pour la préservation de la rade de Marseille et de son littoral. Ce projet s'inscrit dans le contexte des Jeux olympiques de 2024 pour lesquels la ville de Marseille accueillera une partie des épreuves (dont les épreuves nautiques comme la voile).

1.2 Problématique

Jusqu'en 2020, la maintenance des avaloirs se déroulait de la manière suivante : un opérateur se rendait sur le terrain afin de vérifier visuellement l'état d'encrassement. Cette vérification est appelée *visite*. Si l'opérateur constatait que l'avaloir est encrassé, il le nettoyait, ce que l'on appelle *curage*. La maintenance du réseau était effectuée selon des quotas de 50000 visites et 25000 curages annuels.

Pour éviter les déplacements inutiles et intervenir au bon endroit et au bon moment, la solution retenue consiste à équiper chaque avaloir d'un capteur permettant de suivre à distance et en temps réel son niveau d'encrassement. Ainsi, en juin 2023, environ 4500 capteurs mesurent le niveau d'encrassement d'une partie des avaloirs du réseau de la ville. Le suivi en temps réel des déchets au sein des avaloirs permet une maintenance plus efficace.

L'objectif de cette thèse est de dépasser la surveillance individuelle de ces niveaux d'encrassement au sein des avaloirs et de **comprendre les dynamiques**

globales d'accumulation des déchets dans le but d'améliorer la maintenance du réseau d'avaloirs et d'éviter toutes sortes de nuisances assimilés.

1.3 Acteurs du projet

SUEZ est l'entreprise finançant cette thèse à travers le dispositif CIFRE (Conventions Industrielles de Formation par la Recherche), est spécialisée dans les services de gestion de l'eau, des déchets et de l'énergie. L'Annexe A présente plus en détails ses chiffres clefs.

Le **LyRE** est le centre de recherche de SUEZ, basé à Bordeaux. Son objectif principal est d'innover et de développer des solutions technologiques avancées pour répondre aux défis environnementaux et améliorer la gestion des ressources. Il se concentre sur plusieurs domaines de recherche liés à l'environnement, tels que la gestion de l'eau, la gestion des déchets, l'efficacité énergétique et les énergies renouvelables.

Le **SERAMM** est l'organisme chargé de la gestion et de l'entretien du réseau d'assainissement dans la métropole de Marseille. Il est responsable de la collecte, du traitement et de la distribution de l'eau potable, ainsi que de la collecte et du traitement des eaux usées. Son objectif principal est de préserver la qualité de l'eau et de protéger l'environnement tout en répondant aux besoins de la population en matières d'assainissement et d'approvisionnement en eau potable. L'Annexe A présente plus en détails ses chiffres clefs.

L'Institut Fresnel est un laboratoire de recherche basé à Marseille avec des domaines de recherche principaux sont l'optique, la photonique, et le traitement du signal et des images.

L'Institut de Mathématique de Marseille (I2M) est un laboratoire de mathématiques pures et appliquées, également basé à Marseille.

1.4 Réseau d'assainissement

1.4.1 Présentation du réseau

Comme toutes les métropoles, Marseille possède un réseau d'assainissement conséquent et complexe (Voir Annexe A), qui a la particularité d'être divisé en trois sous-réseaux, illustrés sur la Figure 1.3, à savoir : le réseau d'eaux usées, le réseau pluvial relié aux avaloirs de la ville (afin d'absorber et d'évacuer les eaux de ruissellement), et le réseau dit *unitaire* qui combine eaux usées et eaux pluviales.

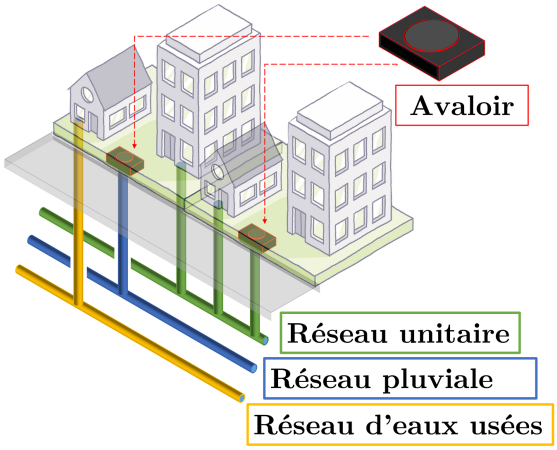

FIGURE 1.3 – Illustration des différents réseaux : eaux usées, pluviales et unitaire (mélange les eaux pluviales et les eaux usées).

Cette répartition en sous-réseaux est historique : le premier réseau d'assainissement qui a été construit à Marseille est le réseau unitaire, créé à la fin du XIX^{e} siècle. Cette ancienne infrastructure se trouve principalement dans l'hypercentre de la ville et ses alentours. Le réseau construit par la suite, au milieu du XX^{e} siècle, est appelé *séparatif* : on sépare les eaux usées des eaux de pluie afin d'éviter la remontée des mauvaises odeurs des eaux usées à travers les avaloirs. La Figure 1.4 présente une cartographie de ces différents sous-réseaux.

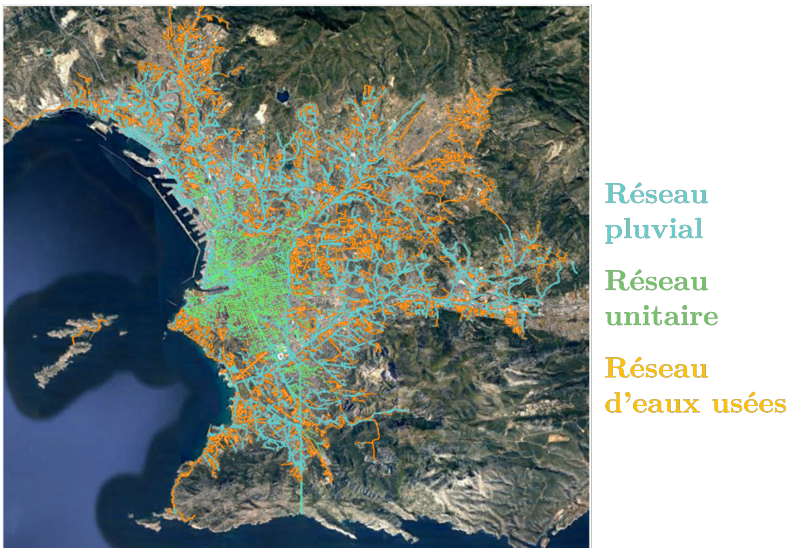

FIGURE 1.4 – Répartition des différents sous-réseaux d'assainissement de Marseille. Le réseau unitaire, en vert, est principalement situé dans les quartiers les plus historiques de la ville et ses alentours.

La problématique traitée dans cete thèse concerne l'encrassement des avaloirs correspondant au réseau pluvial et au réseau unitaire (le réseau d'eaux usées n'étant pas relié aux avaloirs).

1.4.2 L'avaloir

L'avaloir est le premier dispositif d'absorption des eaux de pluie. Un avaloir est un ensemble d'ouvertures, c'est-à-dire un mélange de marquises et/ou de grilles qui mènent vers un même engouffrement, appelé *fosse*. Par exemple, on retrouve un avaloir avec deux marquises sur la Figure 1.5.(a) ; une succession de grilles sur la Figure 1.5.(b) ; une marquise seule sur la Figure 1.5.(c) et un avaloir avec une marquise et une grille sur la Figure 1.5.(d). Actuellement, on compte environ 16 000 avaloirs à Marseille.

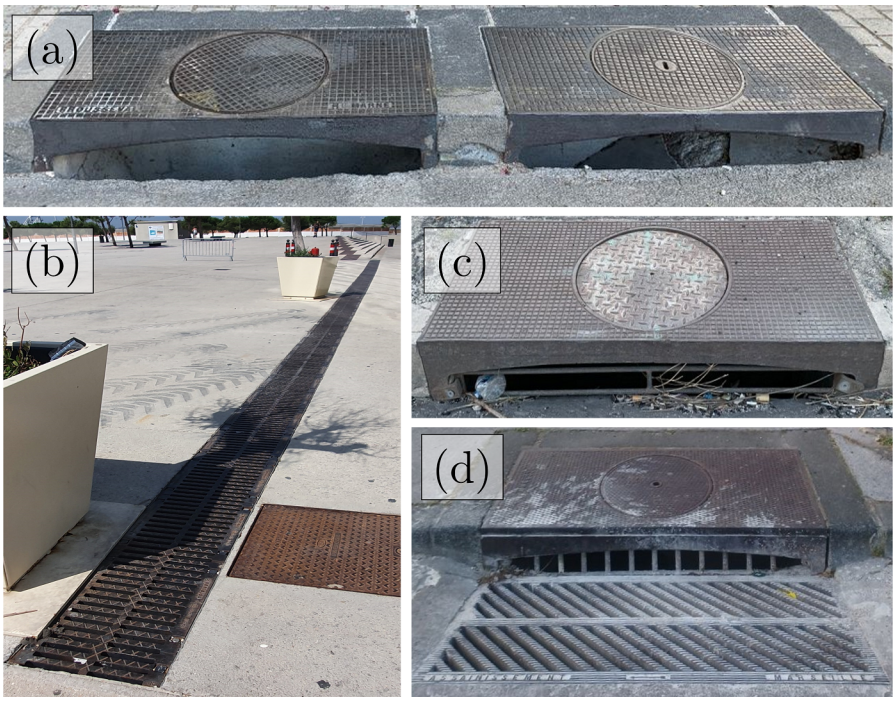

FIGURE 1.5 – Illustrations de différents avaloirs : (a) avaloir avec deux marquises sans barreaudage ; (b) une succession de grilles ; (c) une marquise seule avec un barreaudage horizontal ; et (d) un avaloir composé d'une marquise avec un barreaudage vertical et d'une grille.

Dans notre étude, le dispositif de suivi du niveau d'encrassement (présenté plus en détail en Section 1.5) est installé en priorité sur une marquise. Les avaloirs composés uniquement de grilles n'ont donc pas été équipés de capteur durant les

premières séries d'installations. Toutefois, en août 2023, une cinquantaine de grilles sont équipées de tels capteurs.

La forme des avaloirs

La Figure 1.6 décrit les divers éléments d'un avaloir à marquise unique. On y retrouve notamment le *tampon*, qui correspond à l'ouverture supérieure de l'avaloir (utilisé pour la maintenance), et *l'exutoire*, qui est l'évacuation vers le réseau d'assainissement.

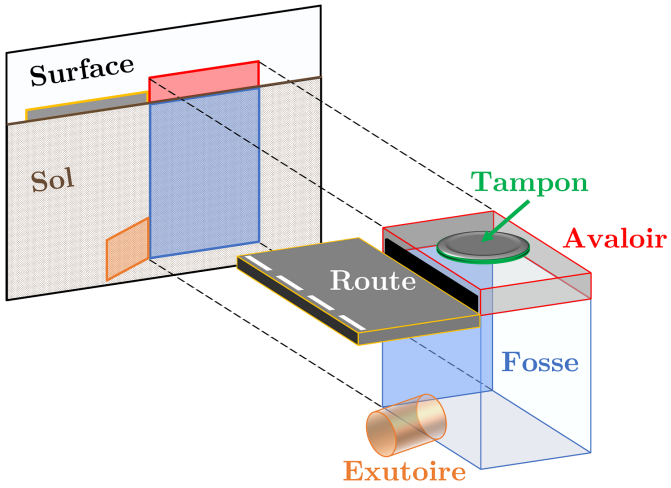

FIGURE 1.6 – Illustration d'un avaloir à marquise, de son couvercle supérieur (tampon) utilisé pour la maintenance, et du tuyau d'évacuation des eaux (exutoire). La fosse de l'avaloir correspond au volume sensible à l'encrassement. Dans cette illustration, la fosse a une forme nommée "Pavé droit - Rectangle".

Les fosses des avaloirs présentent des formes variées. Elles peuvent être rectangulaires, comme présenté en Figure 1.6 ; trapézoïdales (cf. Figure 1.7.(a)) ; cylindriques (cf. Figure 1.7.(b)) ; ou quelconques (cf. Figure 1.7.(c)). Par ailleurs, on distingue plus spécifiquement la forme du fond de l'avaloir. Les illustrations mentionnées précédemment représentent toutes des avaloirs ayant un fond *plat*. Cependant, ce fond peut aussi être *incliné*, comme illustré en Figure 1.7.(d). La forme de l'avaloir, et notamment celle de son fond, sont des caractéristiques cruciales pour le dispositif. Nous approfondirons ce sujet en Section 1.5.

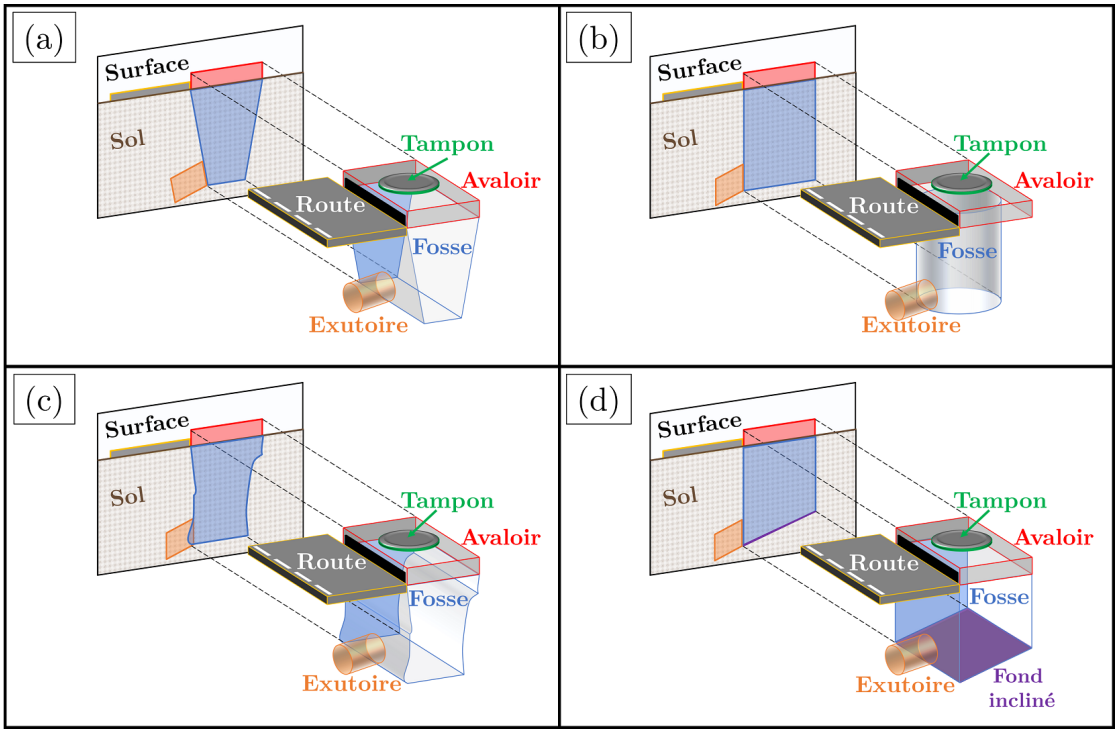

FIGURE 1.7 – Illustrations des différentes formes d'avaloir : (a) le "Trapèze" ; (b) le "cylindrique" ; (c) forme quelconque ; et (d) avaloir ayant un fond incliné.

Équipements des avaloirs

Les avaloirs peuvent être équipés de divers systèmes influençant leur géométrie ou leur sensibilité aux déchets. Deux exemples classiques d'équipements sont décrits ci-dessous.

- **Le barreaudage** : certaines marquises sont équipées de barreaux afin d'empêcher les macro-déchets (d'une dizaine de centimètres ou plus, comme des canettes, des bouteilles, etc.) de rentrer dans l'avaloir. Ce barreaudage peut être horizontal ou vertical (voir respectivement les Figures 1.5.(c) et 1.5(d)). Environ 2500 avaloirs du réseau sont équipés de barreaux.
- **Le panier** : il s'agit d'une structure (généralement en forme de seau) qui a pour fonction de retenir les macro-déchets afin d'éviter qu'ils ne pénètrent dans les canalisations. Cet équipement est présenté en Figure 1.8. Le panier est un équipement amovible qui facilite le curage. Ces paniers sont parfois indispensables pour équiper les avaloirs de forme quelconque d'un capteur. La Section 1.5 décrit plus en détail les problèmes liés à cette forme. En juin 2023, le SERAMM a conduit une phase d'expérimentation avec une quarantaine de paniers déployés.

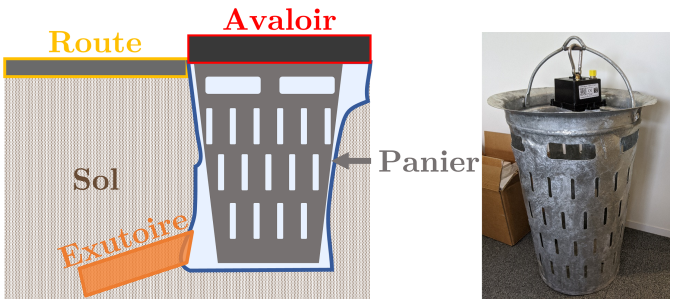

FIGURE 1.8 – Illustration d'un panier d'avaloir en forme de seau.

1.5 L'avaloir connecté

Présentation du dispositif

Afin de pouvoir suivre le niveau d'encrassement des avaloirs, la solution qui a été choisie par le SERAMM est de les équiper d'un capteur de mesure du niveau d'encrassement *US Hummbox GreenCityZen*, comme présenté sur la Figure 1.9. La description et les spécificités du capteur peuvent être trouvées en Annexe B. Aussi, la Figure 1.10 présente un schéma du dispositif complet. Le capteur est installé dans la partie supérieure de l'avaloir, pointant vers le bas. Il calcule le niveau d'encrassement présent dans l'avaloir en mesurant le temps entre l'émission d'une impulsion ultrasonore et la réception de son écho. Ces capteurs ont une fréquence nominale de 2 mesures par jour, avec une autonomie théorique de 10 ans. La fréquence de mesure peut être augmentée au détriment de la consommation énergétique et donc d'une baisse de l'autonomie. La mesure est ensuite transmise vers un *cloud* (centre de stockage et de calcul) par un réseau sans-fil dédié, qui sera décrit en Section 1.6.

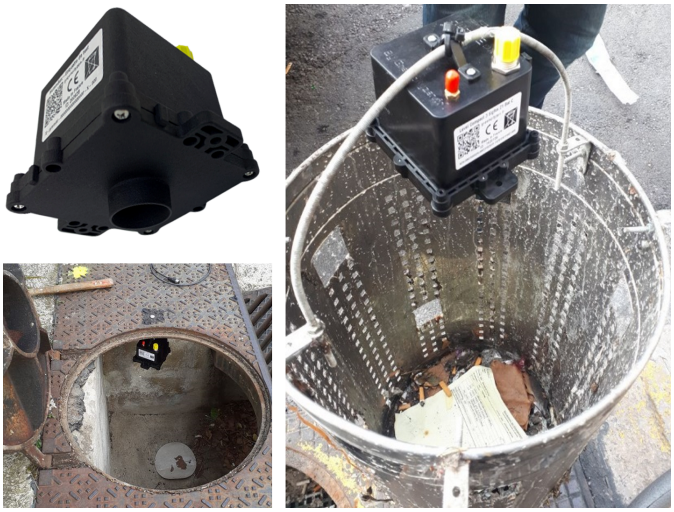

FIGURE 1.9 – Capteur utilisé pour mesurer le niveau d'encrassement d'un avaloir.

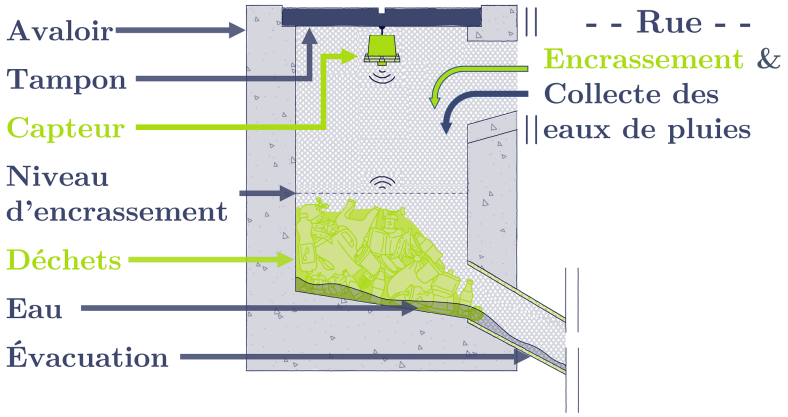

FIGURE 1.10 – Avaloir connecté : le capteur est placé sur la partie supérieure, pointant vers le bas et calcule le niveau d'encrassement en mesurant le délai entre l'émission d'une impulsion ultrasonore et la réception de son écho.

Contraintes et Aspects pratique de l'installation du capteur

Les capteurs utilisés ont une "zone morte", c'est-à-dire qu'ils ne peuvent pas mesurer des distances inférieures à 20-25 cm. Il n'est donc pas pertinent d'installer ces capteurs dans les avaloirs ayant une fosse peu profonde. Dans notre cas, les avaloirs étudiés ont une profondeur allant de 40 à 450 cm. Aussi, pour un fonctionnement optimal, la surface de reflexion doit être la plus plane et perpendiculaire possible vis-à-vis de la trajectoire de l'onde ultrasonore, ce qui empêche l'installation du capteur dans certains avaloirs. Par exemple, la Figure 1.11 illustre un avaloir dont le

fond est trop incliné et la Figure 1.12 montre un avaloir dont le fond n'est pas une surface plane. De plus, l'onde émise par le capteur peut se refléter dans certains cas sur les parois de l'avaloir plutôt que sur le fond, comme illustré sur la Figure 1.13. De ce fait, le capteur doit être installé aussi loin que possible des bords de la fosse, ce qui peut rendre son installation difficile (voire impossible) pour certains avaloirs de forme irrégulière, comme illustré sur la Figure 1.14. En pratique, c'est l'opérateur qui détermine visuellement si un avaloir respecte ces conditions de dimensions et de forme avant de procéder à l'installation du capteur. Nous verrons dans la Section 2.1.6 que ce problème de planéité de surface se pose également lors de la mesure du niveau d'encrassement.

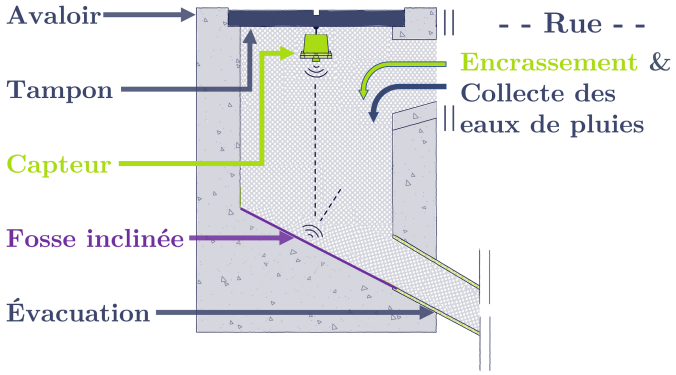

FIGURE 1.11 – Avaloir non adapté à l'installation d'un capteur car le fond de la fosse est particulièrement incliné.

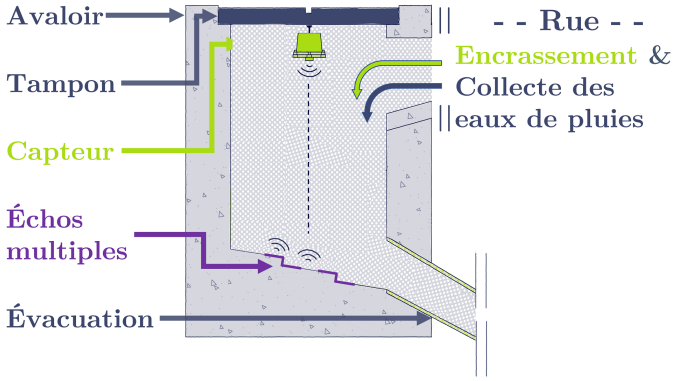

FIGURE 1.12 – Avaloir non adapté à l'installation d'un capteur car le fond de la fosse est une surface non plane.

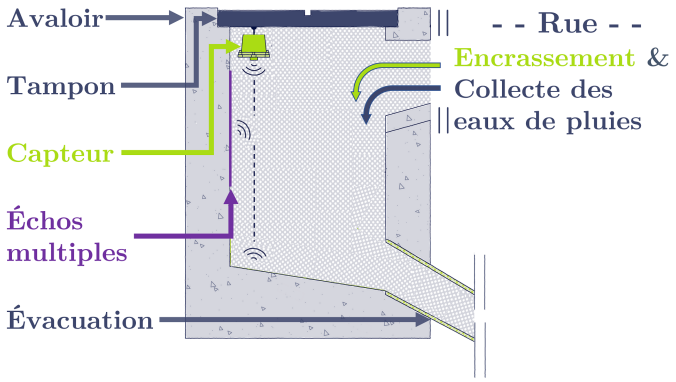

FIGURE 1.13 – Capteur installé trop proche des parois, l'onde émise est alors réfléchie par la paroi, rendant la mesure du niveau erronée.

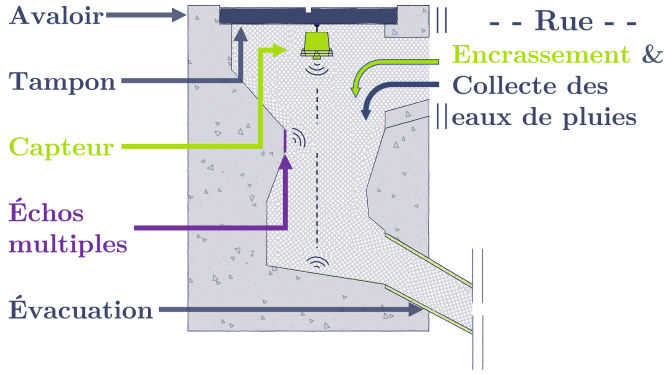

FIGURE 1.14 – Avaloir dont la forme n'est pas adaptée à l'installation d'un capteur.

1.6 Dimensions du réseau étudié

Le réseau de capteurs et son déploiement

Les capteurs ultrasonores sont connectés via un réseau étendu à faible consommation LPWAN (*Low Power Wide Area Network*) [18, 19], plus précisément un réseau Sigfox. Les mesures effectuées par les capteurs sont rassemblées et traitées dans un cloud en passant par des passerelles (*gateways*) Sigfox. Pour donner une vision plus concrète du réseau, la Figure 1.15 présente une illustration du réseau réel déployé dans le centre-ville de Marseille (quartier du Vieux Port) ainsi qu'un schéma de l'architecture du réseau. En juin 2023, environ 4500 capteurs avaient été déployés.

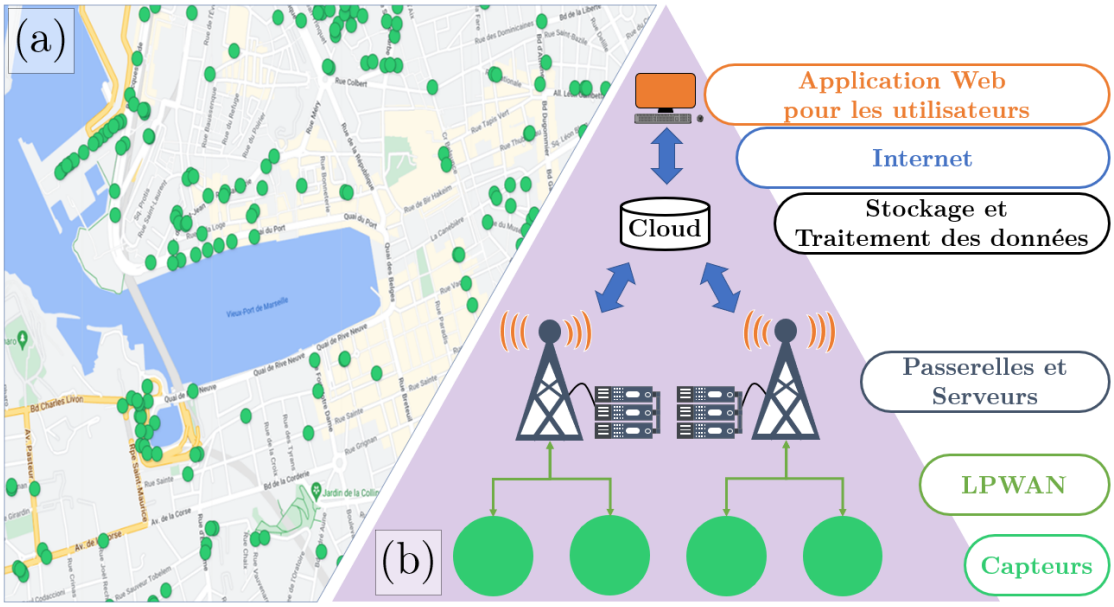

FIGURE 1.15 – Illustration du réseau LPWAN : (a) une cartographie du réseau déployé dans le quartier du Vieux-Port, les points verts représentent la position des avaloirs connectés ; (b) architecture réseau.

Le déploiement de ce réseau s'est fait progressivement. Les premiers essais ont été réalisés en 2019 avec une cinquantaine de capteurs. Une fois le dispositif validé, l'installation du réseau a débuté en 2020 avec le déploiement de 1200 nouveaux capteurs. Ces capteurs ont été déployés en priorité dans des zones sensibles telles que le littoral, l'hypercentre, les zones de marchés ou près du fleuve Huveaune qui se jette dans la mer à Marseille. Le déploiement s'est ensuite poursuivi avec l'installation d'environ 3300 capteurs supplémentaires entre début 2021 et juin 2023. Le déploiement se poursuit avec l'objectif à terme d'avoir un réseau opérationnel de 5000 capteurs d'ici 2024, année de lancement des jeux olympiques.

Représentativité du réseau étudié

Compte tenu des aspects pratiques expliqués dans la section précédente, le capteur est installé en priorité sur une marquise (pour des raisons liées à la profondeur de la fosse), excluant les avaloirs composés uniquement de grilles. Parmi les $\sim$ 16 000 avaloirs du réseau d'assainissement de Marseille, $\sim$ 9500 possèdent au moins une marquise. De plus, parmi ces 9500, seuls $\sim$ 6100 avaloirs sont considérés comme *équipables* (possèdant une forme et des dimensions adaptées au bon fonctionnement du capteur). Un réseau de 5000 avaloirs connectés représente environ 31% des avaloirs de Marseille, 53% des avaloirs avec marquise, et 82% des avaloirs équipables avec la technologie utilisée.

1.7 Description des données collectées

1.7.1 Mesure du niveau d'encrassement

Chaque capteur mesure la distance jusqu'à la surface de réflexion qui peut correspondre au niveau d'encrassement, au fond de l'avaloir (s'il est vide) ou encore au niveau d'eau (dans le cas où l'avaloir est obstrué et que l'eau s'y est accumulée). Comme illustré sur la Figure 1.16, le niveau d'encrassement N_e correspond à :

$$N_e = D_{\max} - D\,, \tag{1.7.1}$$

avec D la distance mesurée par le capteur et $D_{\max}$ la distance maximale que le capteur peut mesurer, c'est-à-dire la distance jusqu'au fond de l'avaloir. Cette dernière est étalonnée lors de l'installation. La distance minimale mesurée dans notre cas est de l'ordre d'une vingtaine de centimètres (ce que nous avions appelé *zone morte* du capteur auparavant).

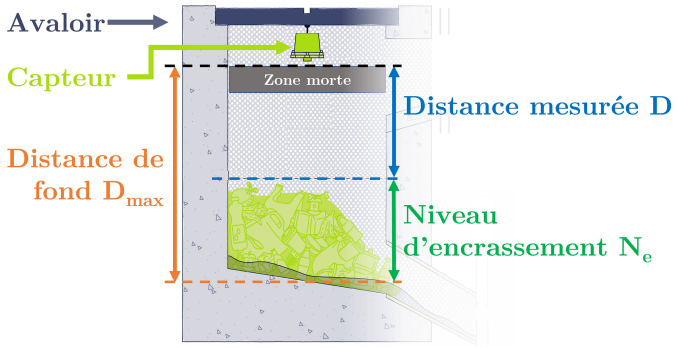

FIGURE 1.16 – Illustration de la mesure du niveau d'encrassement et de la *zone morte*.

1.7.2 Fréquence de mesure

La fréquence de mesure de chaque capteur est ajustable entre 1 et 144 mesures/jour. Lorsque l'opérateur reconfigure cette fréquence de mesure, la modification est prise en compte lors de la prochaine connexion du capteur au réseau LPWAN. L'historique des données collectées contient donc des mesures à période variable. Par exemple, sur 1 an d'historique pour un capteur donné, on peut constater 24 mesures/jour pendant 2 mois ; 2 mesures/jour pendant 6 mois puis 4 mesures/jour pendant 4 mois. Enfin, à une fréquence de mesure fixée, on peut également observer une dérive de la période entre deux mesures. Cette dérive varie d'une mesure à l'autre et peut aller jusqu'à quelques dizaines de secondes.

1.7.3 Dynamique d'encrassement

Les données résultantes décrivent l'évolution dans le temps du niveau d'encrassement présent dans l'avaloir. Cette évolution est influencée par de nombreux facteurs exogènes et environnementaux tels que la pluie, le vent, la topographie, etc. Ces facteurs sont propres à chaque avaloir (voir Section 4.1 pour plus de détails).

En visualisant les données collectées, on constate que les avaloirs ont des dynamiques différentes. Quelques exemples de dynamique sont présentés sur les Figures 1.17 et 1.18 (données réelles). On peut constater sur ces exemples que le niveau d'encrassement peut varier progressivement ou soudainement. On constate également que le niveau d'encrassement peut diminuer sans qu'il n'y ait eu de curage. Dans ce cas, on suppose que l'encrassement s'est engouffré dans le réseau, ou est ressorti de l'avaloir par débordement (en cas de pluie par exemple). Nous appelons ces pertes des *lessivages* (cf. Figure 1.19). Dans le cas où les déchets présents dans l'avaloir se "tassent", une baisse du niveau d'encrassement sera également observée. Ne pouvant distinguer ce phénomène de "tassage" des rejets réels de déchets, nous appellerons par la suite *lessivages* toute baisse de niveau d'encrassement non due à un curage.

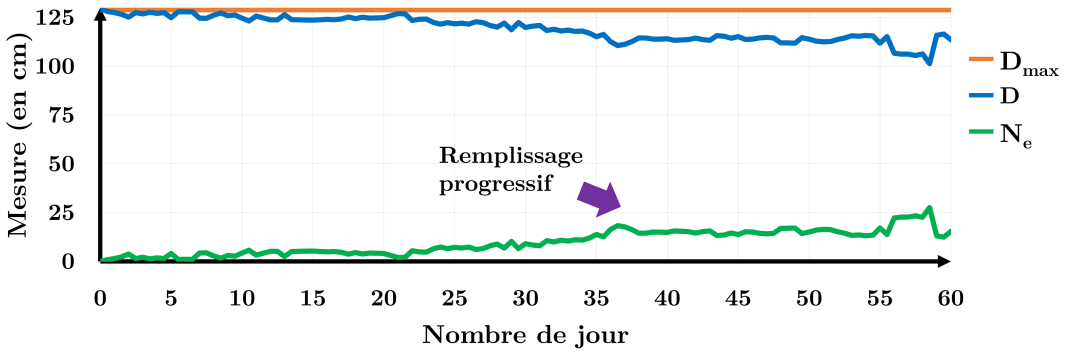

FIGURE 1.17 – Données réelles d'un avaloir avec un remplissage progressif. On constate que l'avaloir s'est rempli d'environ 25 cm de déchets en deux mois.

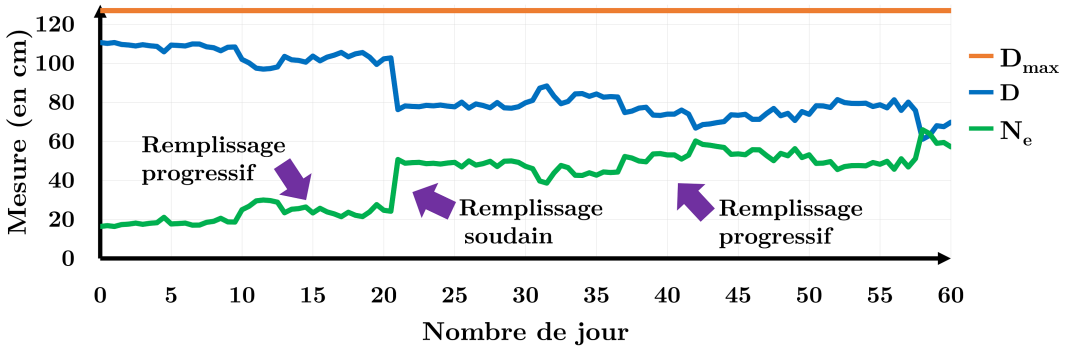

FIGURE 1.18 – Données réelles d'un avaloir avec un remplissage qui peut être progressif ou soudain (comme celui du jour 21).

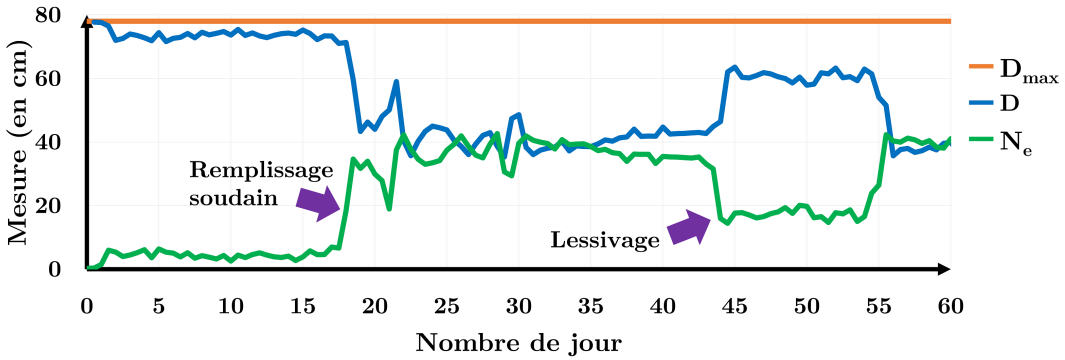

FIGURE 1.19 – Données réelles d'un avaloir sur lesquelles on peut constater deux remplissages soudains (jour 18 et 21) ainsi qu'une perte des déchets (lessivage) jour 44.

1.8 Synthèse du chapitre

Dans ce chapitre, nous avons décrit l'initiative du SERAMM, une filiale de Suez, visant à optimiser la maintenance des 16 000 avaloirs du réseau d'assainissement en les dotant de capteurs connectés. Le déploiement de ce réseau a débuté par des phases de test en 2019, et sa mise à échelle est prévue jusqu'en 2024 afin de disposer d'un réseau opérationnel de 5000 capteurs. Ces capteurs, installés dans la partie supérieure de l'avaloir, mesurent le niveau d'encrassement en calculant le temps écoulé entre l'émission d'une impulsion ultrasonore et la réception de son écho. Ils possèdent une fréquence de mesure modulable et peuvent être reconfigurés par l'opérateur. À première vue, les données collectées par ces capteurs révèlent une diversité de comportements possibles : on peut observer, par exemple, que l'encrassement peut augmenter progressivement ou brusquement, et qu'il peut

diminuer même sans curage, rendant éventuellement unique la dynamique de chaque avaloir. Du point de vu opérationnel, on observe que certains éléments exogènes influencent ou bien sont responsable d'une partie de cette dynamique d'encrassement.

Cette thèse a pour but de décrypter la dynamique d'encrassement globale en s'appuyant sur les données des capteurs. Cependant, avant d'entamer une analyse approfondie, il est essentiel de procéder à un nettoyage des données. Le prochain chapitre traitera de ce sujet.

Chapitre 2

Prétraitement des données

Sommaire

Le prétraitement vise à éliminer autant que possible les informations erronées. Ce chapitre décrit les problématiques rencontrées concernant les données et les solutions mises en œuvre pour y répondre. L'une de ces problématiques concerne notamment le traitement des pics de mesure. Les travaux de prétraitement réalisés pour éliminer ces pics ont fait l'objet d'une publication dans la revue Internet of Things de l'éditeur Elsevier [20].

2.1 Spécificités des données collectées

2.1.1 Données manquantes

Avant tout, rappelons que les données étudiés ici sont des mesures du niveau d'encrassement dans les avaloirs équipés du capteur ultrasonnore, comme décrit en Section 1.7. Premièrement, on peut observer certaines valeurs manquantes dans les données collectées. Ce problème concerne à la fois le référencement du réseau d'avaloirs et les mesures effectuées par les capteurs. Ces deux cas sont décrits plus en détail ci-dessous.

Problèmes de référencement

Le réseau d'assainissement de Marseille étant ancien, sa digitalisation passe par le référencement du patrimoine, et plus précisément, celui des avaloirs. Aujourd'hui, un *Système d'Information Géographique* (SIG) regroupe l'ensemble des données des avaloirs : leur localisation GPS, leurs dimensions, leurs formes, les équipements dont ils disposent, etc. Cependant, ces travaux de référencement sont encore en cours et présentent des lacunes qui peuvent résulter de divers facteurs : erreurs lors du transfert des anciennes informations (anciennes bases de données, documents papiers, etc.) à la nouvelle base, erreurs de mise à jour de la base (lors de l'installation de nouveaux équipements par exemple) ou à des contraintes opérationnelles lorsqu'un opérateur chargé de référencer l'avaloir ne peut pas y accéder (véhicule stationné dessus, travaux de voirie, etc.).

Mesures effectuées par les capteurs

Dans certains cas, le capteur peut échouer à transmettre la mesure, entraînant ainsi des mesures manquantes. Ce problème peut être plus ou moins important en fonction de la distance entre le capteur et le *gateway* LPWAN, du positionnement du capteur dans l'avaloir, etc. Cela peut être de courte durée (quelques heures ou jours), par exemple, en cas d'inondation locale où le capteur noyé ne parvient plus à envoyer ses mesures, ou bien de plus longue durée, comme lors d'un dysfonctionnement du capteur en lui-même, de son installation (le capteur tombe dans l'avaloir) ou suite

à un vol de capteur. Dans ce cas, la résolution de ce problème est plus complexe et nécessite une intervention humaine sur le terrain.

2.1.2 Fréquence de mesure et quantité de données

Aujourd'hui, la fréquence nominale de mesure des capteurs est fixée à 2 mesures/jour. Cette fréquence a été choisie afin de garantir un suivi suffisamment régulier du niveau d'encrassement pour effectuer des actions de nettoyage ciblés. De plus, ce choix de fréquence permet de garantir une autonomie d'environ 10 ans pour les capteurs. On supposera que cette fréquence est suffisante pour identifier et comprendre la dynamique globale d'accumulation des déchets. Bien évidemment, pour observer les dynamiques ayant un temps caractéristique plus réduit (heure ou minute), il faudrait augmenter cette fréquence de mesure. Ainsi, la précision avec laquelle les mesures décrivent la dynamique en pratique dépend du capteur et du paramétrage de la fréquence de mesure. On rappelle que ce paramétrage peut varier dans le temps (par exemple, 24 mesures/jour pendant 2 mois, puis 2 mesures/jour pendant 6 mois, etc.). Finalement, le réseau de capteurs ayant été déployé progressivement sur plusieurs années, l'historique des données disponibles pour chaque avaloir dépend de la date d'installation du capteur et de ses éventuelles interventions de maintenance. On choisira plus tard d'uniformiser les informations disponibles en sous-échantillonnant les données (cf Section 3.4.1).

2.1.3 Incohérences par rapport aux dimensions de l'avaloir

Certaines mesures sont incohérentes par rapport à la géométrie de l'avaloir. Il s'agit des cas où la valeur de la mesure est supérieure à la distance entre le capteur et le fond de l'avaloir, appelée ici *distance de fond* $D_{\max}$, ce qui est physiquement impossible (cf. la Figure 1.16). Afin de supprimer ces mesures, il suffit en théorie de simplement supprimer les mesures D supérieures à $D_{\max}$.

En pratique, $D_{\max}$ peut avoir été mal calibrée lors de l'installation du capteur ou ne pas avoir été mise à jour lors d'une modification de l'installation. Une autre solution moins contraignante consiste à prendre comme référence la profondeur de l'avaloir P_{ava} et non $D_{\max}$. Toutefois, celle-ci peut également être erronée car le référencement de l'avaloir n'est pas parfait.

Finalement, le choix a été fait de supprimer les mesures supérieures à $\max(D_{\max} + 20 \text{ cm}, P_{\text{ava}})$. L'idée est de prendre comme référence la valeur la plus grande entre $D_{\max}$ et P_{ava}, en rajoutant une marge d'erreur de 20 cm pour la calibration de $D_{\max}$. Le choix de cette marge de 20 cm repose sur l'hypothèse qu'elle représente une limite supérieure de l'erreur de calibration de $D_{\max}$. Autrement dit, on établit un seuil moins contraignant tout en conservant une marge d'erreur afin d'éviter la suppression de mesures potentiellement valides.

La Figure 2.1 montre un exemple typique de ce problème : un capteur installé dans un avaloir d'une profondeur de $P_{\text{ava}} = 63$ cm (selon le référencement) avec une distance de fond calibrée à $D_{\max} = 38.5$ cm. On constate que la plupart des mesures sont supérieures à $D_{\max}$. En pratique, on interprète ce cas de la manière suivante : les mesures aux alentours de 50 cm semblent assez stables donc la distance de fond $D_{\max}$ réelle doit être d'environ 50 cm et les mesures plus grandes (par exemple, supérieures à 1 m) sont des mesures anormales à supprimer. En appliquant le seuil choisi, les mesures supérieures à $\max(D_{\max} + 20 \text{ cm}, P_{\text{ava}}) = 58.5$ cm seront supprimées.

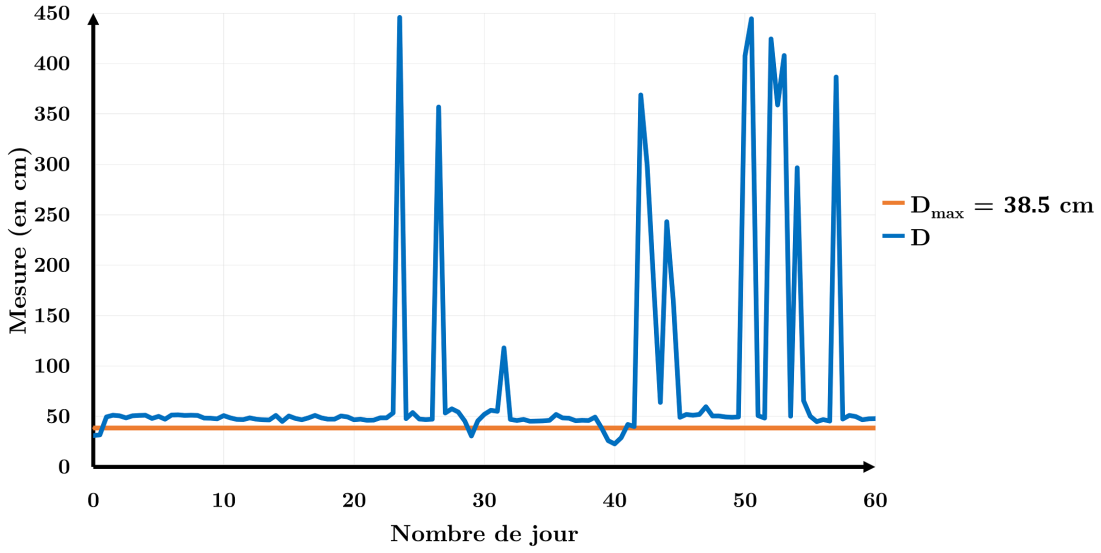

FIGURE 2.1 – Mesure d'un capteur installé dans un avaloir de profondeur 63 cm selon le référencement. La distance de fond a été calibrée à 38.5 cm. On constate des mesures largement supérieures à $D_{\max}$.

2.1.4 Incohérences par rapport à la fréquence de mesure

Par construction, les capteurs n'effectuent pas deux mesures séparées par moins de 10 minutes. Il est cependant possible de retrouver dans l'historique des données des mesures à des intervalles de temps plus courts (par exemple : une mesure relevée à 20h02 et l'autre à 20h04). Ces mesures successives dans un intervalle de temps réduit sont généralement des valeurs redondantes et peuvent être dues à des tests de fonctionnement et de calibration du capteur pendant son installation ou à d'autres problèmes inconnus. Dans de tels cas, notre approche est de supprimer les mesures redondantes.

2.1.5 Incohérences liées à l'installation

Il peut arriver que le capteur envoie des valeurs aberrantes ou incohérentes sur des périodes plus ou moins longues. On conjecture alors que le capteur est mal installé ou que d'autres facteurs externes et inconnus perturbent les mesures. Ces anomalies affectent la fiabilité des données collectées, et nécessitent des ajustements ou des vérifications supplémentaires afin de garantir l'intégrité des mesures. Un exemple de ce problème est présenté dans la Figure 2.2, où les mesures du capteur oscillent considérablement au début, variant autour de 175 cm ou de 75 cm. Ces fluctuations pourraient être causées par un mauvais positionnement du capteur : l'onde ultrasonore émise par le capteur se réfléchirait sur la paroi de l'avaloir plutôt que sur le fond. Dans cet exemple, une intervention a été réalisée autour des jours $24 - 25$ pour repositionner le capteur. Suite à cette intervention, les données semblent s'être stabilisées. Ce problème ne survient pas seulement lors de l'installation du capteur, il peut également être intermittent. La Figure 2.3 illustre le cas d'un capteur dont les mesures semblent incohérentes entre les jours 34 et 82, mais fiables le reste du temps. Cette incohérence pourrait résulter d'une perturbation temporaire inconnue affectant le capteur. Les capteurs présentant ce genre d'anomalies ont été exclus de notre étude dès leur détection.

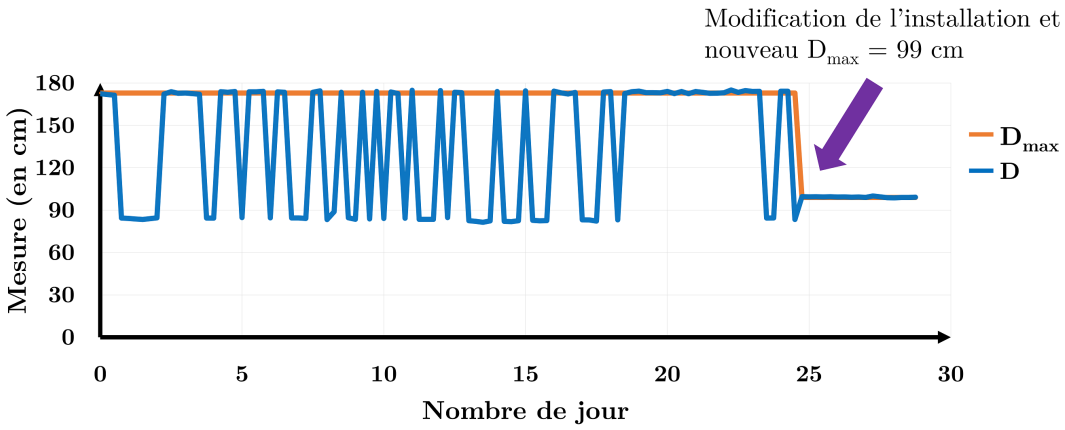

FIGURE 2.2 – Exemple de donnée réelles présentant des mesures incohérentes. Un repositionnement du capteur a permis de résoudre le problème au jour ~ 25.

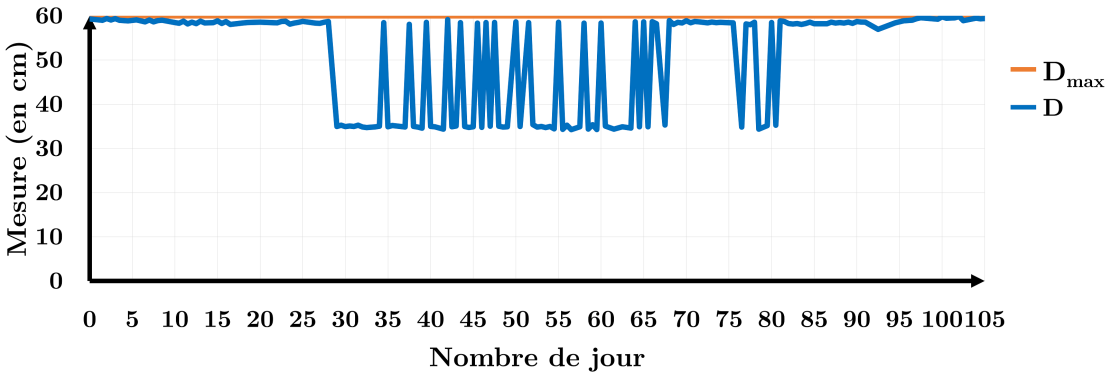

FIGURE 2.3 – Exemple d'un capteur dont les mesures présentent des oscillations importantes entre les jours 34 et 82 mais qui seraient fiables le reste du temps.

2.1.6 Bruits de mesures

Variabilité de la mesure

Selon la documentation technique du capteur, l'exactitude de la mesure est de ±2 cm et sa résolution est de ±1 cm. Néanmoins, en pratique, la variabilité des mesures semble dépendre de l'avaloir et peut évoluer au fil du temps. Actuellement, cette variabilité est interprétée comme étant causée par la nature de la surface de réflexion. Pour rappel, pour un fonctionnement optimal du capteur, la surface mesurée devrait être aussi plane et horizontale que possible (voir Section 1.5).

Pour illustrer cette variabilité, prenons deux exemples : les figures Figures 2.4 et 2.5 montrent un cas où la variabilité de mesure est faible, avec 95% des mesures se situant dans un intervalle de largeur 0.37 cm. À l'inverse, les figures Figures 2.6 et 2.7 montrent un cas avec une grande variabilité, où 95% des mesures se trouvent dans un intervalle de largeur 8.27 cm. On interprète ces observations comme étant le résultat de deux capteurs mesurant un niveau d'encrassement constant (probablement faible ou inexistant puisque les mesures sont proches de $D_{\max}$) mais avec des surfaces de mesure différentes dans chaque avaloir, ce qui expliquerait la variabilité des mesures observées. Dans certains cas extrêmes, la variabilité est si prononcée (comme illustré en Figure 2.8) que les données sont considérées comme anormales et sont donc écartées de notre analyse.

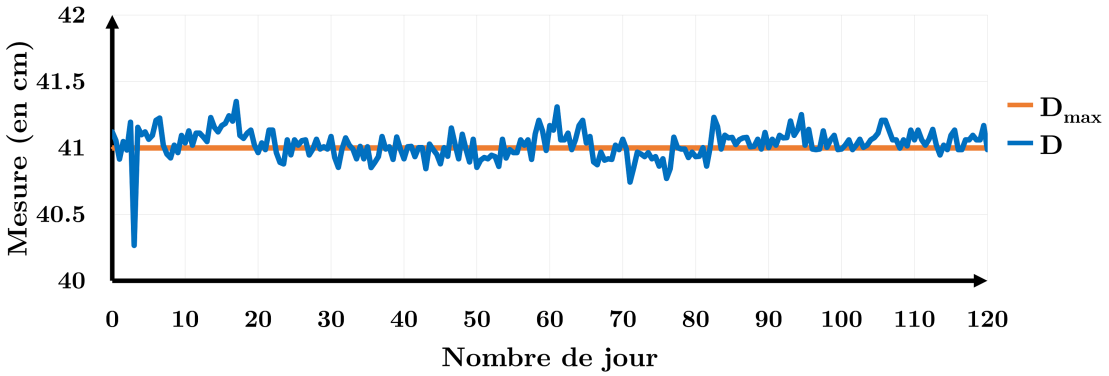

FIGURE 2.4 – Mesures de faible variabilité autour d'un niveau d'encrassement constant et supposé nul.

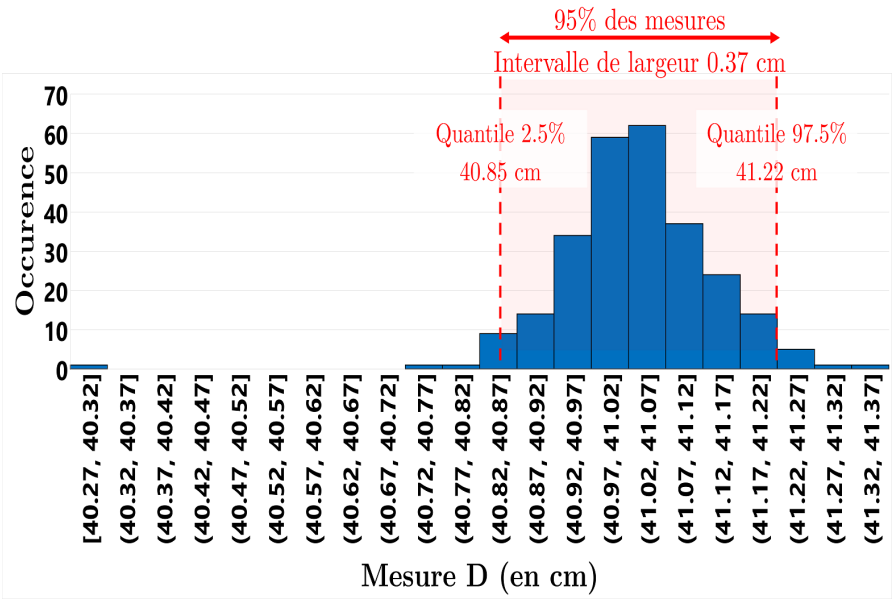

FIGURE 2.5 – Distribution des mesures présentées en Figure 2.4. Dans ce cas, 95% des mesures ce trouvent dans un intervalle de largeur 0.37 cm.

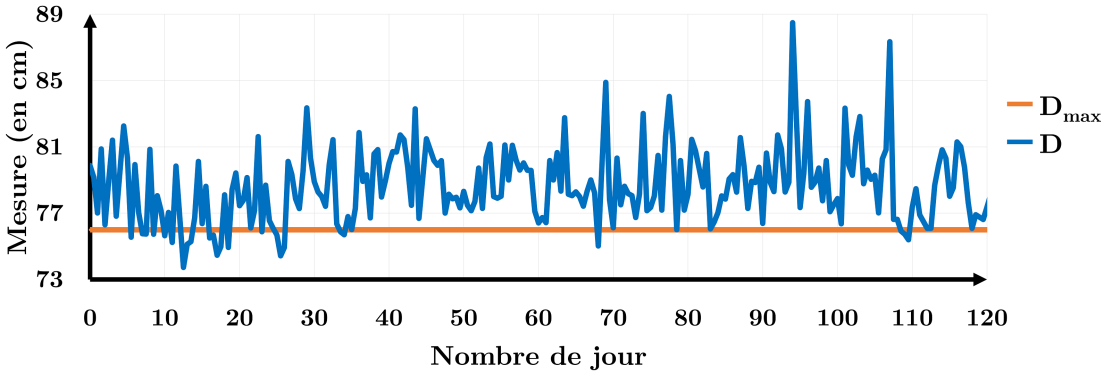

FIGURE 2.6 – Mesures de forte variabilité autour d'un niveau d'encrassement constant et supposé nul.

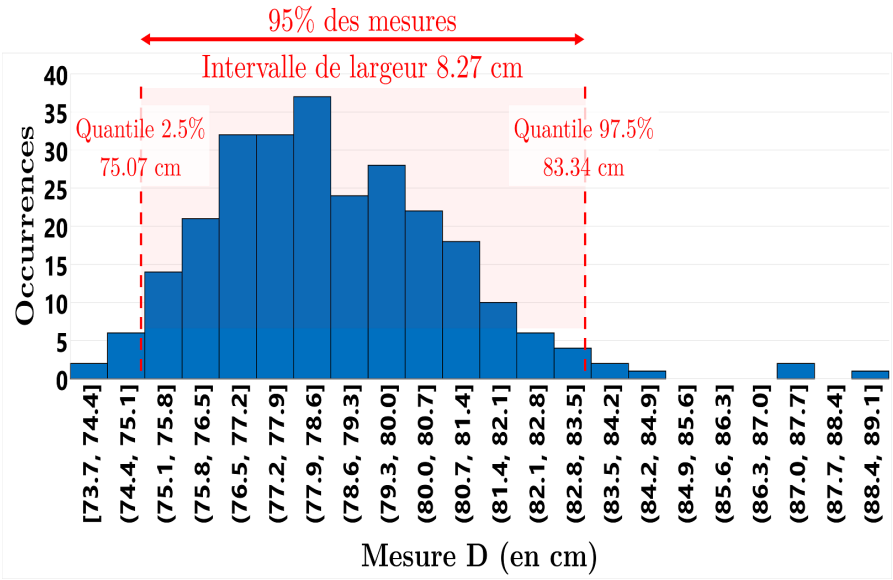

FIGURE 2.7 – Distribution des mesures présentées en Figure 2.6. Dans ce cas, 95% des mesures ce trouvent dans un intervalle de largeur 8.27 cm.

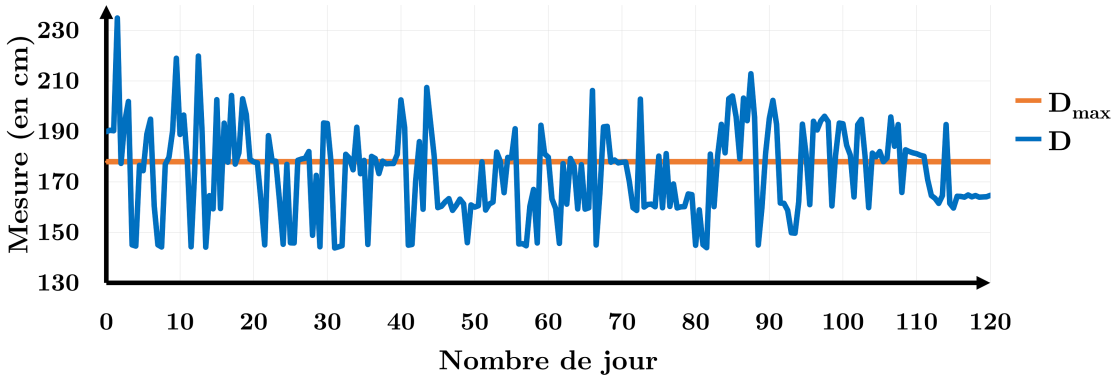

FIGURE 2.8 – Mesures avec une très grande variabilité que l'on considère comme anormales. Lorsque nous identifions un capteur avec ce type de problème, il est écarté de notre étude.

Bruit de type pic

En plus du bruit associé à la variabilité "classique" des mesures, nous observons dans les données collectées, un autre bruit prenant la forme de pics. Ces mesures aberrantes apparaissent localement et montrent des valeurs qui divergent fortement de la tendance générale. Comme mentionné précédemment, on explique ce bruit par la nature de la surface mesurée. Celle-ci peut être influencée par la forme des déchets, leur composition, ou encore leur disposition à l'intérieur de l'avaloir. Le problème lié à une surface de mesure inégale est illustré en Figure 2.9. Un exemple concret de ce bruit sur des données réelles est présenté dans la Figure 2.10, où l'on identifie clairement 4 pics aberrants.

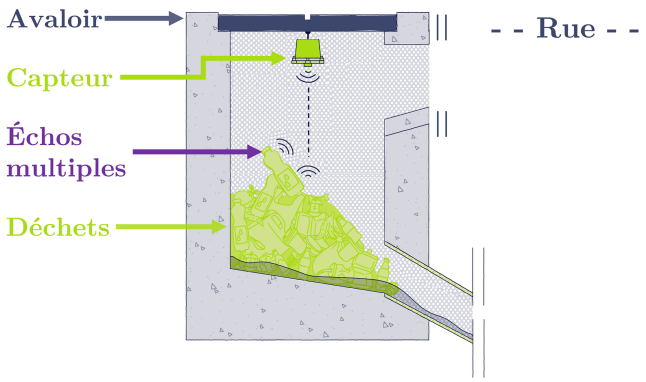

FIGURE 2.9 – Illustration des problèmes de mesure lorsque les déchets présents dans l'avaloir forment une surface de réflexion non plane.

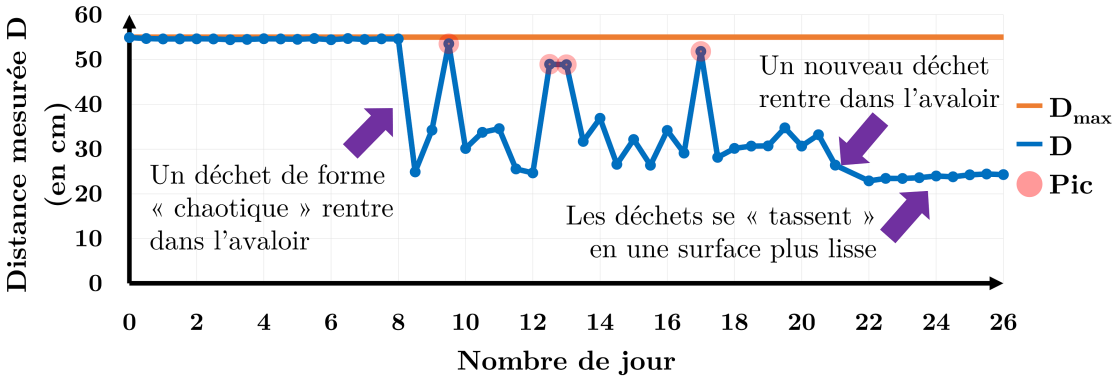

FIGURE 2.10 – Exemple du bruit type "pic" présent dans les données, les points aberrants sont surlignés en rouge.

2.2 Détection et suppression des pics

Comme décrit en Section 2.1.6, les données étudiées peuvent contenir des mesures localement aberrantes dues à la surface de mesure formée par les déchets présents dans l'avaloir. Cette section se consacre à la détection et la suppression de ces anomalies. Nous examinerons d'abord quelques méthodes existantes avant de détailler celle retenue pour nettoyer les données. La performance de cette méthode sera ensuite évaluée sur un échantillon de données étiqueté manuellement.

2.2.1 Détection d'anomalies dans les séries temporelles

De nombreux travaux précédents ont abordé la détection des valeurs aberrantes dans les séries temporelles [21, 22]. Cependant, la plupart de ces solutions ne conviennent pas à notre contexte où nous traitons une grande quantité de données univariées non stationnaires et non étiquetées avec une évolution complexe au fil du temps. La dynamique étudiée ici dépend de nombreux paramètres contextuels, cf. Section 4.1. Pour illustrer la nécessité de développer une nouvelle solution pour l'application envisagée, nous présentons brièvement dans ce qui suit une description de quelques méthodes de la littérature, et expliquons pourquoi elles ne conviennent pas à notre cas.

Une façon simple pour supprimer les pics des données consisterait à appliquer un filtre passe-bas pour éliminer ce "bruit haute fréquence". Cependant, cette solution n'est pas applicable ici en raison de la nature des mesures : l'utilisation de cette approche entraînerait la suppression des discontinuités inhérentes aux données. De plus, même les filtres préservant ces discontinuités, comme le filtre médian [23], entraîneraient une perte d'informations utiles en raison du lissage des données car ils ne peuvent pas distinguer les événements de courte durée présent dans les données

(informations utiles) des pics que l'on souhaite supprimer. De manière plus générale, le lissage des données, tel que le lissage exponentiel [24], biaiserait tout traitement ou analyse des données qui pourrait suivre (tel que classification, prédiction, etc.). Pour des solutions telles que l'*Adaptive Piecewise Constant Approximation* [25], où les valeurs aberrantes sont supprimées indépendamment dans chaque segment de données, les performances dépendent fortement du nombre de segments définis initialement, ce qui est difficile à optimiser dans notre cas car chaque avaloir a un comportement qui lui est propre.

En ce qui concerne les méthodes de détection d'anomalies statistiques proposées dans la littérature, elles reposent principalement sur la prédiction. Les approches classiques basées sur les modèles autorégressifs (AR), à moyenne mobile (MA) ou ARMA nécessitent une stationnarité du second ordre (c'est-à-dire en termes de moyenne et de variance), ce qui ne s'applique pas dans notre cas. De manière similaire, AR intégrée MA (ARIMA) ne convient pas à notre contexte car elle suppose qu'il existe un ordre de différenciation permettant d'obtenir une stationnarité [26]. De plus, l'approche itérative de suppression des valeurs aberrantes basée sur le *Extreme Studentized Deviate Test* nécessite que le nombre de valeurs aberrantes soit connu au préalable [27].

D'autres approches étudiées pour la détection des valeurs aberrantes sont basées sur l'apprentissage automatique. La plupart des algorithmes proposés, tels que le *One Class Support Vector Machine* [28], sont supervisés ou semi-supervisés. Ceci nécessite que des valeurs aberrantes soient étiquetées ou au moins qu'un ensemble de données propres (sans valeurs aberrantes) soit disponible, ce qui ne peut donc pas être appliqué aux données brutes dans notre cas. Quelques algorithmes non supervisés de détection des valeurs aberrantes ont également été proposés pour les séries temporelles, tels que le *Peer-Group Analysis* [29], qui consiste à caractériser un schéma de comportement attendu entre des objets similaires. Cependant, dans notre cas, chaque avaloir a ses propres caractéristiques et peut être considéré comme indépendant des autres. Une autre technique classique, non supervisée, est celle du *Sub-sequence Time Series Clustering* [30], qui consiste à appliquer l'algorithme K-Means aux séries temporelles en utilisant une fenêtre glissante. Encore une fois, chaque avaloir ayant ses caractéristiques spécifiques, le paramètre K (le nombre de clusters dans cet l'algorithme) doit être ajusté de manière personnalisée. Une autre solution consiste à utiliser des méthodes basées sur les auto-encodeurs, qui sont des réseaux de neurones artificiels utilisés pour apprendre un encodage de manière non supervisée, par exemple en utilisant un *Long Short Term Memory* [31]. Cependant, de telles méthodes nécessitent généralement un volume important de données pour optimiser les hyperparamètres du réseau de neurones sous-jacent (le nombre de couches, le nombre de cellules dans chaque couche, la taille de la fenêtre, etc.) et modifieraient les données initiales, risquant ainsi de les biaiser.

2.2.2 Méthodes étudiées

Généralement, la détection des valeurs aberrantes peut être considérée comme un problème de classification. Une approche courante consiste à transformer les séries temporelles en une représentation en nuage de points facilitant la séparation entre les points normaux et anormaux. Pour effectuer cette transformation, l'idée ici est d'attribuer un score à chaque point afin de caractériser son "degré d'anormalité". Les différentes scores utilisées dans notre cas pour détecter ces pics sont décrites et illustrées ci-dessous en se basant sur l'exemple de la Figure 2.10.

Z-Score

Le Z-Score est une mesure statistique couramment utilisée pour détecter les anomalies dans un ensemble de données [32]. Pour une série de données de taille N, $\boldsymbol{X} = [X_1, \cdots, X_N]$ avec une moyenne μ et un écart-type σ, le Z-Score associé à chaque mesure X_i, noté ici $Z(X_i)$, est donné par $Z(X_i) = (X_i - \mu)/\sigma$.

Le Z-Score permet d'évaluer la position d'une valeur par rapport à la moyenne et à l'écart-type d'un ensemble de données. Généralement, on suppose implicitement que les données suivent une distribution gaussienne et on interprète le Z-score comme une score d'anormalité : plus le Z-Score est élevé (en valeur absolue), plus la mesure est éloignée de la moyenne et plus la probabilité que la mesure correspondante soit aberrante est élevée.

En pratique, la détection des valeurs aberrantes s'effectue en utilisant un seuil : si le Z-Score d'une valeur dépasse ce seuil, elle est considérée comme aberrante. Cela entraîne donc l'élimination d'un certain pourcentage de données. Par exemple, lors de l'application du Z-Score à un grand nombre de points issue d'un tirage gaussien, l'utilisation d'un seuil de 2 ou 3 revient, respectivement, à supprimer environ 4.6% et 0.3% des données.

Dans notre cas, en supposant que les données $\boldsymbol{X}$ suivent une distribution gaussienne et qu'environ 1% des données collectées correspondent à des anomalies (pics), un seuil d'environ 2.5 est recommandé. Dans le cas étudié, les données étant non stationnaire (c'est-à-dire avec une moyenne μ et un écart-type σ variables dans le temps), la méthode du Z-Score doit être appliquée sur une fenêtre glissante d'une longueur appropriée, en fonction de la dynamique de l'accumulation des déchets. À titre d'exemple, nous avons appliqué le Z-Score sur les données de la Figure 2.10 en utilisant un seuil de 2.5, d'abord sur l'ensemble complet des données (voir la Figure 2.11), puis sur une fenêtre glissante d'une longueur de 5 jours (voir la Figure 2.12). Il est important de noter que réduire ce seuil revient à rejeter un plus grand nombre de points, ce qui améliore la détection des pics mais augmente aussi le nombre d'anomalies faussement identifiés. Inversement, augmenter le seuil réduit les fausses détections au détriment des performances de détection des pics.

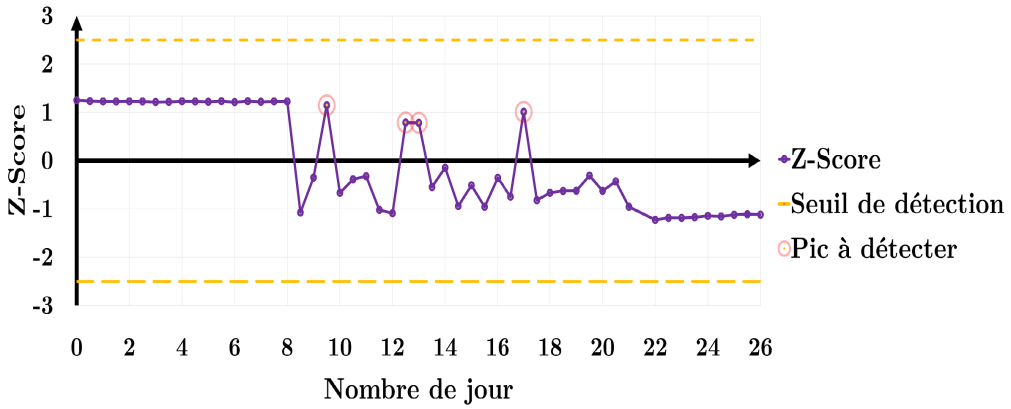

FIGURE 2.11 – Exemple d'application du Z-Score sur les données présentées dans la Figure 2.10. Dans ce cas, le Z-Score est appliqué sur l'ensemble des données et ne détecte aucune anomalie.

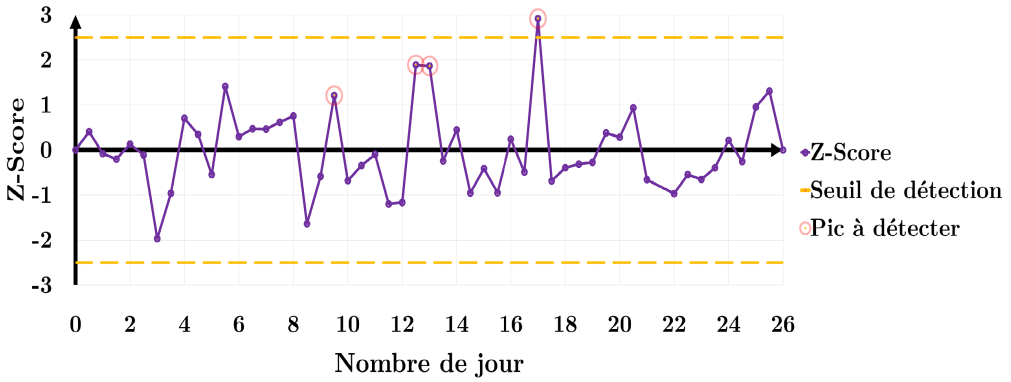

FIGURE 2.12 – Exemple d'application du Z-Score sur les données présentées dans la Figure 2.10. Dans ce cas, le Z-Score est appliqué sur une fenêtre glissante de longueur de 5 jours et identifie un des pics au jour ~ 17.

Notons que le choix de la taille de la fenêtre glissante choisie (5 jours) est basé sur le retour d'expérience des opérateurs terrain : on estime à première vue que l'on peut associer aux avaloirs dynamiques (avec des variations régulières du niveau d'encrassement) un temps caractéristique de remplissage allant de quelques jours à une semaine. Ainsi l'hypothèse d'une distribution gaussienne pour une fenêtre de taille 5 jours nous paraît raisonable.

Le choix de la taille de la fenêtre est donc basé sur un ordre de grandeur (qui peut ne pas être adapté à toutes les dynamiques). Ainsi, pour une performance optimale, à taille de fenêtre fixe, le seuil de détection considéré devrait idéalement être défini de manière "personnalisée" pour chaque avaloir et être adaptatif, pour correspondre au mieux, en temps réel, à la dynamique de remplissage observée. La mise en

œuvre d'un tel seuillage (personnalisé et adaptatif) est évidemment complexe et nécessiterait, par exemple, de passer par la détection de chaque évènement de remplissage, de lessivage ou de curage, entraînant ainsi une complexité considérable du traitement des données.

Détection des variations opposées

À la recherche d'une meilleure détection des valeurs aberrantes, nous avons d'abord développé une méthode simple et intuitive consistant à identifier ces pics comme correspondant à un changement "inhabituel" de la distance mesurée, sous forme d'une variation soudaine importante (négative ou positive) de D. Nous appelons cette méthode *Opposite Variation Detection* (OVD). Considérons la dérivation temporelle de D à l'instant i comme suit :

$$\delta_i = \frac{D_{i+1} - D_i}{\Delta T_i}, \tag{2.2.1}$$

avec ΔT_i l'intervalle de temps entre les mesures D_i et D_{i+1}. Pour rappel, en fonctionnement nominal, on a $\forall i, \Delta T_i = 12$ heures.

Afin d'identifier des changements de variations de D, on se concentre sur le changement de signe de δ_i en définissant le paramètre :

$$\kappa_i = \frac{\text{sign}(\delta_{i-1}) - \text{sign}(\delta_i)}{2}, \tag{2.2.2}$$

avec sign$(\cdot)$ la fonction signe. Par construction, κ_i est égale à zero si les signes de δ_i et de δ_{i+1} sont identiques, et à ± 1 dans le cas contraire. Un pic, par définition, se traduit par une variation positive puis négative (ou inversement négative puis positive) de la mesure. L'objectif de κ_i est donc de repérer ce motif ; plus précisément, X_i est un pic potentiel si $\kappa_i \neq 0$. À partir de κ_i, on définit un score :

$$\theta_i = \kappa_i \cdot \min(|\delta_i|, |\delta_{i-1}|) \tag{2.2.3}$$

où $|\cdot|$ est la fonction valeur absolue. L'idée derrière l'équation (2.2.3) est d'associer à chaque point un score d'anomalie, qui est obtenu à partir de la valeur absolue des pentes locales du tracé de D (par rapport aux points de données précédents et suivants). Cette dernière est le minimum des deux pentes calculées afin de tenir compte des discontinuités du signal. Ainsi, les pics potentiels correspondraient à des valeurs relativement élevées de θ_i. Afin d'expliquer plus amplement cette idée, deux exemples d'application de ce score sont montrés sur les Figures 2.13 et 2.14. La Figure 2.13 montre un point P_1 normal tandis que la Figure 2.14 montre un pic à détecter P_2. Ici, les deux points peuvent être considérés comme des pics potentiels

car, d'après l'équation (2.2.2), les κ_i correspondants sont différents de zéro [1]. Le score θ_i est calculé de manière à distinguer ces deux cas.

Désignons les scores correspondants par θ_{P_1} et θ_{P_2}. Les pentes de la mesure D au point P_1, c'est-à-dire δ_{P_1}, sont désignées par $SP_{1,1}$ et $SP_{1,2}$; de même, les pentes au point P_2 sont désignées par $SP_{2,1}$ et $SP_{2,2}$. Ici, pour le point P_1, la plus petite pente est $SP_{1,2}$, donc $\theta_{P_1} = |SP_{1,2}|$. De même, pour le point P_2, la plus petite pente est $SP_{2,1}$, donc $\theta_{P_2} = |SP_{2,1}|$. Nous remarquons que $\theta_{P_1} < \theta_{P_2}$, en d'autres termes, P_2 est plus susceptible d'être un pic que P_1, ce qui est effectivement le cas ici.

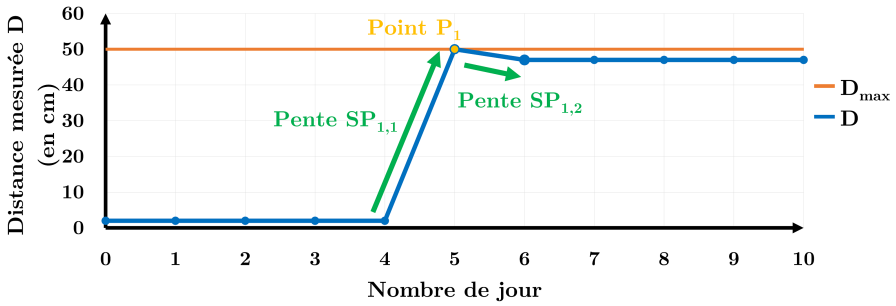

FIGURE 2.13 – Illustration de l'idée derrière le score θ : le point P_1 est un point normal, marquant une discontinuité dans le signal. Dans ce cas, θ sera "faible" en valeur absolue car il est associé (par définition) à la plus petites des deux pentes, $SP_{1,2}$.

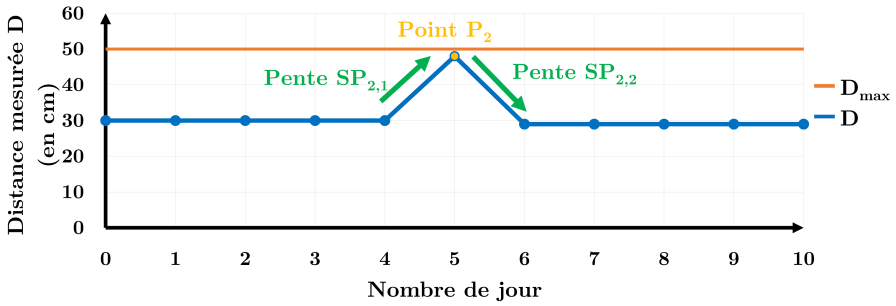

FIGURE 2.14 – Illustration de l'idée derrière le score θ : le point P_2 est un pic à détecter. Dans ce cas, θ sera d'une valeur "plutôt élevée" en comparaison du cas présenté en Figure 2.13 car la plus petites des deux pentes, $SP_{2,1}$, reste importante dans ce cas.

Nous avons également montré sur la Figure 2.15 une illustration de la détection des valeurs aberrantes en calculant le score θ et en appliquant un seuil. En comparant cette figure avec la Figure 2.10, on peut voir que, par exemple, pour un seuil de 2.5, la méthode OVD a détecté deux pics mais a manqué les deux autres. En réalité,

1. La raison est que les variations sont de signes opposées : on a une augmentation puis une diminution de la mesure dans les deux cas.

le score OVD est proche de zéro dans le cas de pics composées de deux points successifs, ce qui est bien sûr inapproprié.

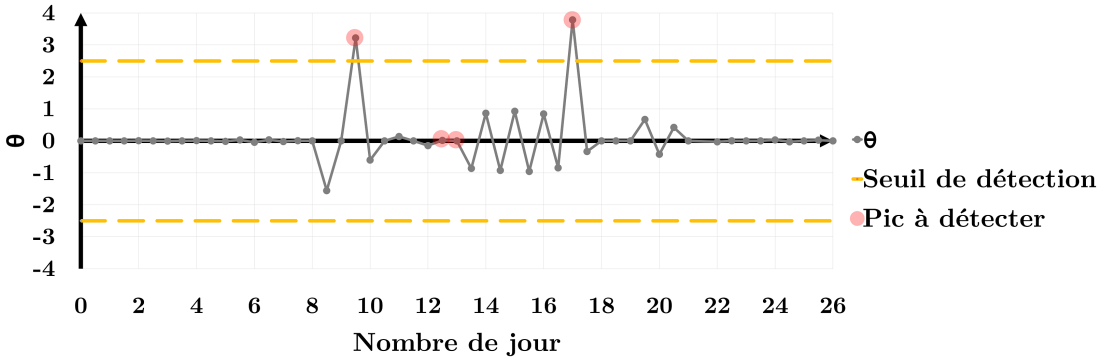

FIGURE 2.15 – Détection des pics en utilisant un seuil de 2.5 en valeur absolue à le score θ. Dans ce cas, seul deux des quatres points anormaux sont détectés.

Solution proposée

Comme décrit précédemment et étant donnée la particularité des données collectées, les méthodes basées sur le Z-Score avec une fenêtre glissante et la méthode OVD montrent des performances limitées en pratique pour détecter les valeurs aberrantes dans les données. Il est important de noter que ces méthodes sont basées sur des scores qui sont proches de 0 lorsqu'elles sont appliquées à des points normaux. Ici, nous proposons une solution plus efficace, que nous appelons le *Peak Pattern based Z-score* (PPZ), combinant les deux scores précédents sous la forme d'une analyse bidimensionnelle (2D) des données, comme décrit ci-dessous. L'idée derrière cette approche est de compléter les informations fournies par la méthode Z-Score avec celles obtenues à partir de l'OVD.

La méthode proposée consiste à calculer pour chaque point X_i de la série temporelle, les scores associés Z_i et θ_i, ce qui définit les vecteurs $\boldsymbol{v}_i = (Z_i, \theta_i)$, $i = 1, ..., N$. Notons les vecteurs correspondants $\boldsymbol{Z} = (Z_1, Z_2, ..., Z_N)$ et $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_N)$; on calcule ensuite la norme de chaque vecteur $\boldsymbol{v}_i$ pour la métrique de Mahalanobis [33] :

$$D_M(\boldsymbol{v}_i) = (\boldsymbol{v}_i - \boldsymbol{\mu})^T \boldsymbol{K}_{\boldsymbol{Z},\boldsymbol{\theta}}^{-1} (\boldsymbol{v}_i - \boldsymbol{\mu}) \tag{2.2.4}$$

où μ_Z et μ_θ, les moyennes respectives de $\boldsymbol{Z}$ et $\boldsymbol{\theta}$, définissent le vecteur moyenne $\boldsymbol{\mu} = (\mu_Z, \mu_\theta)$ et où $\boldsymbol{K}_{\boldsymbol{Z},\boldsymbol{\theta}}$ est la matrice de covariance de $\boldsymbol{Z}$ et $\boldsymbol{\theta}$, de taille (2×2).

Une mesure i est considérée comme anormale si cette distance, donnée par l'équation (2.2.4), est supérieure à un seuil ξ. Du point de vue géométrique, cela revient à représenter $\boldsymbol{Z}$ et $\boldsymbol{\theta}$ sous forme de nuage de points. Comme chacun

de ces scores donne une valeur proche de zéro pour les données “normales”, le nuage de points obtenu sera aproximativement centré autour de 0. Les points correspondant aux valeurs aberrantes seront, par construction, éloignés du centre. Ainsi, en appliquant un seuil 2D, sous la forme d'une ellipse, on peut distinguer les données normales des pics à supprimer. Le réglage du seuil pour la détection des valeurs aberrantes est basé sur la décomposition en valeurs propres de la matrice de covariance $\boldsymbol{K}_{\boldsymbol{Z},\boldsymbol{\theta}}$, appelée *Standard Deviation Ellipse* (SDE) [34]. Les vecteurs propres résultants déterminent l'angle de l'ellipse (seuil de détection ici), tandis que la longueur des axes de cette dernière est définie en multipliant la racine carrée des valeurs propres correspondantes par une constante ξ.

La Figure 2.16 illustre la détection des pics à l'aide de la méthode PPZ proposée, appliquée à l'ensemble des données collectées de la Figure 2.10. Ici, la constante ξ est fixée à 2.5, comme précédemment pour les méthodes de Z-Score et OVD. On observe que chaque point dans le graphique 2D correspond aux Z-Score et θ calculés pour chaque mesure (comme montré précédemment dans les Figures 2.12 et 2.15). La méthode permet la détection de tous les pics, bien qu'elle identifie incorrectement le point aux alentours du jour 3 comme un pic [2] (cf. Figure 2.17).

2. La variabilité de la mesure entre les jours 0 et 8 étant minime, le Z-Score associé à ce point est d'environ ~ -2, ce qui le place en dehors de l'ellipse, malgré que l'écart entre les mesures soit très petit (moins de 1 cm) et que le score OVD soit nul.

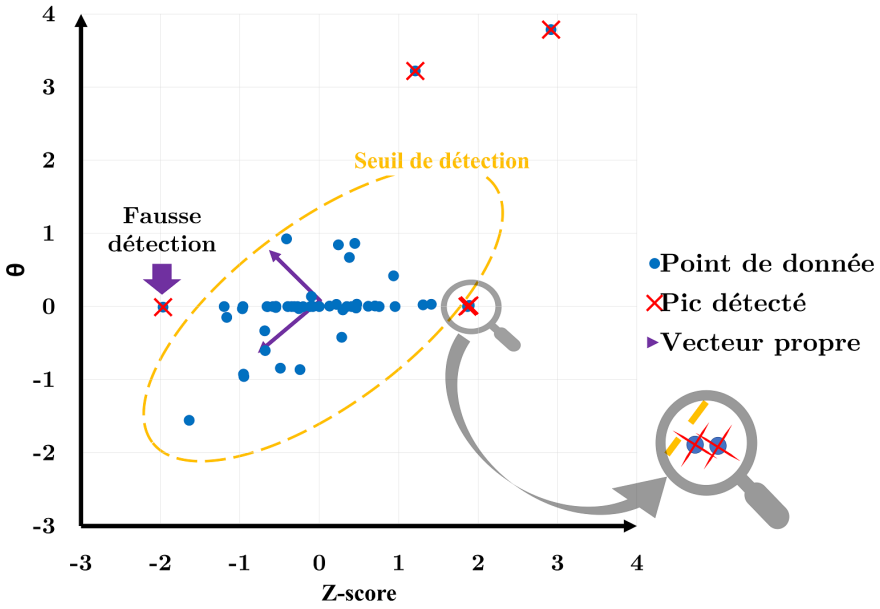

FIGURE 2.16 – Application de la méthode PPZ aux données présentées en Figure 2.10. La détection des anomalies est présentée sous la forme d'une réprésentation 2D des scores utilisées ici avec l'application d'un seuil de détection elliptique ξ fixée à 2.5. Dans ce cas, 5 points sont détectés comme anormaux.

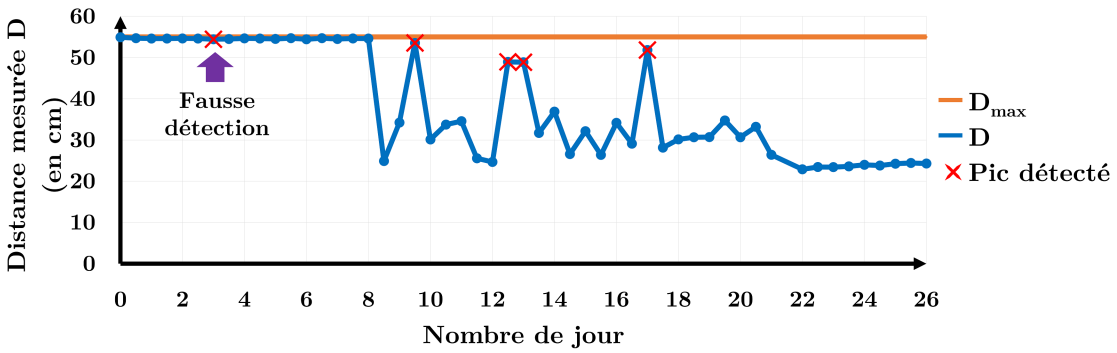

FIGURE 2.17 – Illustration des points de la série temporelle de la Figure 2.10 détectées comme anormaux après l'application de la méthode PPZ pour un seuil ξ de 2.5. Dans ce cas, toutes les anomalies sont détectées, avec en plus une fausse détection.

2.2.3 Évaluation des performances

L'approche courante pour comparer les performances de différentes méthodes de détection de valeurs aberrantes consiste à comparer les courbes *Receiver Operating*

Characteristic (ROC). Ces courbes sont obtenues en calculant la matrice de confusion, dont les éléments comprennent le nombre de vrais positifs N_{TP} (les valeurs aberrantes détectées correctement), les vrais négatifs N_{TN} (les points normaux correctement non détectés), les faux positifs N_{FP} (les points normaux identifiés incorrectement comme des valeurs aberrantes) et les faux négatifs N_{FN} (les valeurs aberrantes non détectées). Les courbes ROC sont obtenues en traçant le *True Positive Rate* (TPR) défini comme $N_{\mathrm{TP}}/(N_{\mathrm{TP}}+N_{\mathrm{FN}})$, en fonction du *False Positive Rate* (FPR) défini comme $N_{\mathrm{FP}}/(N_{\mathrm{FP}}+N_{\mathrm{TN}})$.

L'évaluation des performances peut être réalisée à partir de données étiquetées, simulées ou réelles. Or, à ce stade des travaux, chaque avaloir semble avoir un comportement distinct et les principaux types de dynamique et les proportions qu'ils représentent sont inconnus. La mise en place d'une approche simulée semble difficile car elle n'est pertinente que sous condition que les données simulées reflètent fidèlement les données réelles.

La solution que nous avons choisie consiste à d'évaluer les performances des méthodes présentées précédemment sur un ensemble de donnée réelles, sélectionnées aléatoirement et étiquetées manuellement. L'étiquetage manuel des données s'est effectué de la manière suivante : pour chaque capteur, la série temporelle a été visualisée et chaque mesure a été étiquetée comme étant un pic ou non (tâche fastidieuse). Ainsi, un ensemble de deux mois de données provenant d'environ 300 capteurs a été étiqueté, représentant environ 33 600 points, parmi lesquels plus de 460 points (soit 1.37% des données) ont été étiquetés comme pics.

Nous avons comparé les courbes ROC correspondantes de la méthode PPZ proposée avec celles du Z-Score et de l'OVD dans la Figure 2.18, où l'on peut remarquer la supériorité de la première. Pour l'algorithme proposé, le paramètre de seuil ξ à choisir peut être ajusté pour assurer un TPR minimum ou un FPR maximum requis. Dans notre cas, nous avons opté pour un seuil de 2.5, donnant $\mathrm{TPR} = 0.85$ et $\mathrm{FPR} = 2.5 \times 10^{-2}$. Notez que les performances des algorithmes peuvent également être quantifiées en calculant l'aire sous chaque courbe ROC. Dans le cas de la Figure 2.18, les aires sont de 0.92, 0.95 et 0.98 pour le Z-Score, l'OVD et la méthode proposée, respectivement.

Enfin, en ce qui concerne la complexité calculatoire de notre méthode proposée, pour un capteur avec N mesures, le calcul de chaque score $\boldsymbol{Z}$ et $\boldsymbol{\theta}$ a une complexité $\mathcal{O}(N)$. Ensuite, l'algorithme calcule l'ellipse de seuil, qui est basée sur la décomposition en valeurs propres de la matrice de covariance de dimension (2×2), comme décrit dans la Section 2.2.2, avec une complexité relativement faible. Pendant ce temps, l'estimation de chacune des quatre entrées de la matrice de covariance entraîne une complexité de l'ordre de $\mathcal{O}(N)$. Ainsi, la méthode proposée a une complexité calculatoire de l'ordre de $\mathcal{O}(N)$.

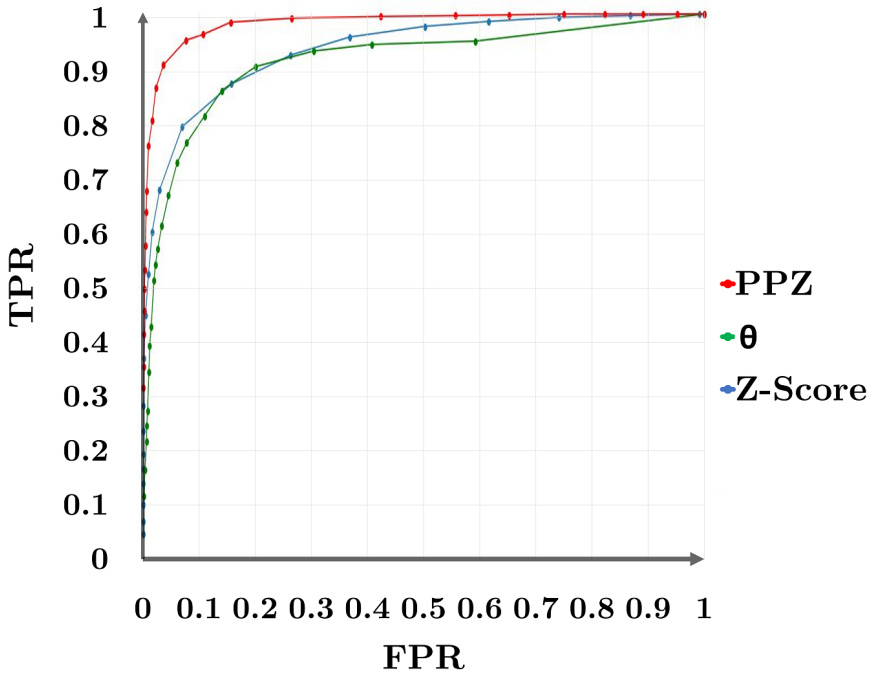

FIGURE 2.18 – Comparaison des performances des trois méthodes de détection d'anomalies étudiées ici (Z-Score, OVD et PPZ) appliquées sur les données réelles étiquetées.

2.3 Synthèse du chapitre

Ce chapitre explique les différents types de problème dans les données issues des capteurs et décrit les méthodes utilisées pour les traiter. En particulier, un nouvel algorithme de détection d'anomalies de type pic a été appliqué aux données collectées. Cette méthode, appelée PPZ, offre des performances supérieures par rapport à la méthode classique du Z-score, avec une complexité calculatoire relativement faible. L'idée principale derrière cet algorithme a été d'améliorer les performances du Z-score en l'associant à un score de détection de motifs, appelé OVD, conçu spécifiquement pour détecter les anomalies de type pic. En d'autres termes, le Z-score offre une détection de valeurs aberrantes à large spectre. Cependant, pour obtenir des performances optimales, il est nécessaire d'utiliser une fenêtre glissante d'une longueur appropriée ainsi qu'un seuil adapté à chaque avaloir, évoluants en fonction de la dynamique observée. La distribution arbitraire du signal et sa non-stationnarité rendent ces exigences difficiles à satisfaire en général. En revanche, à travers la méthode OVD, le score θ a été spécifiquement conçu pour détecter les pics indépendamment de la taille de la fenêtre.

À ce stade d'avancement des travaux, la qualité des données issues du prétraitement nous semble suffisamment correcte pour nous permettre de nous concentrer sur la problématique principale de la thèse, c'est-à-dire l'étude de la dynamique d'encrassement des avaloirs. Le chapitre suivant porte sur cette étude et explique comment nous avons identifié les dynamiques principales et les proportions qu'elles représentent. Plus précisément, nous détaillons comment nous avons regroupé les avaloirs en fonction de leur dynamique en utilisant des algorithmes de classification non supervisée.

Chapitre 3

Étude de la dynamique d'encrassement

Sommaire

3.1 Aperçu général

En observant l'historique des niveaux d'encrassement des avaloirs, on constate divers comportements tels que des remplissages progressifs, des remplissages soudains, des lessivages (pertes de déchets), etc. Ce chapitre est consacré à l'analyse de cette dynamique d'encrassement afin d'identifier les principaux comportements et les proportions qu'ils représentent. Plus précisément, nous avons regroupé les avaloirs selon leur dynamique en utilisant des algorithmes de classification non supervisée (*clustering*). La démarche adoptée pour ces travaux est basée sur des attributs (ce qu'on appelle *feature-based approach*). Cette étude nous a permis de tester de multiples combinaisons d'attributs, d'algorithmes et d'hyperparamètres, tout en nous assurant que les résultats obtenus soient interprétables et en nombre réduits afin de minimiser le temps nécessaire d'analyse (visualisation, interprétation, etc.). Les travaux présentés dans ce chapitre sont basés sur un lot d'environ 2200 avaloirs ayant au moins 1 an d'historique de données.

Ce chapitre est structuré de la manière suivante : nous présentons d'abord les différents attributs que nous avons considérés pour effectuer le clustering, qu'ils soient empiriques, inférentiels ou basés sur des excursions [1]. Nous présentons ensuite comment nous avons sélectionné les attributs, les algorithmes à utiliser et les hyperparamètres associés. Par la suite, nous abordons la sélection de résultats effectuée afin de minimiser le temps nécessaire à l'analyse et à l'interprétation des catégories (*clusters*) obtenus. Finalement, nous présentons les résultats retenus et les interprétations associées en termes de dynamique d'encrassement.

3.2 Construction d'attributs

La construction d'attributs est une phase fondamentale en analyse de données et, plus largement, en science des données. Elle sert à extraire des informations pertinentes des données brutes, facilitant la découverte de motifs, de relations, et tendances cachés qui pourraient autrement rester inaperçus. L'utilisation de connaissances préalables sur les données, lorsqu'elles sont disponibles, peut faciliter le processus de construction d'attributs et enrichir les modèles. En l'absence de telles connaissances, des attributs plus génériques peuvent être employés pour caractériser les données.

Par exemple en traitement d'images, les attributs couramment utilisés comprennent l'intensité des pixels, l'histogramme des couleurs et les attributs de texture (comme les motifs binaires locaux, les filtres de Gabor, et les matrices de cooccurrence) [35, 36, 37]. Ces derniers sont essentiels aux tâches comme la classification

1. Dans cette étude, les *excursions* sont définies comme des périodes durant lesquelles le niveau d'encrassement des avaloirs excède un seuil prédéterminé.

ou la segmentation de textures. Les attributs de détection de bord, comme le détecteur de bords de Canny [38], soulignent les frontières entre différentes régions d'une image. D'autres descripteurs, tels que les moments de Hu [39], capturent les propriétés géométriques des objets dans une image, essentiels pour la reconnaissance et la classification d'objets.

Pour ce qui est de l'étude de séries temporelles qui nous concerne, des attributs statistiques tel que la moyenne, la médiane et l'écart-type, calculables sur l'ensemble de la série ou sur une fenêtre glissante. Les techniques dérivées de l'analyse de Fourier ou de l'analyse en ondelettes, peuvent permettre d'obtenir des informations sur la périodicité de la série étudiée. D'autres méthodes plus classiques comme l'autocorrélation mesurent la relation d'une série avec ses valeurs antérieures, tandis que la corrélation croisée compare deux séries différentes. Les paramètres issus de modèles classiques de séries temporelles, tels que les modèles AR, MA ou ARIMA, peuvent également servir d'attributs [40, 41].

3.3 Attributs empiriques

Dans un premier temps, nous nous sommes concentrés sur la quantification de certains aspects importants de la dynamique d'encrassement du point de vue opérationnel, afin de répondre à certaines questions liées à la maintenance du réseau d'avaloirs. Nous avons créé un attribut opérationnel permettant de caractériser chacun de ces aspects. Ces attributs sont basés sur l'opposé de la variation des mesures :

$$\gamma_i = -\delta_i = -\frac{D_{i+1} - D_i}{\Delta T_i} \tag{3.3.1}$$

où ΔT_i l'intervalle de temps entre les mesures D_i et D_{i+1}. Nous avons choisi de prendre comme référence l'opposé de la variation afin de simplifier l'interprétation : si γ_i est positif, alors l'encrassement augmente ; sinon, il diminue. Par la suite, et sauf mention contraire, nous utilisons γ_i comme référence lorsque nous évoquerons des "variations positives" ou des "variations négatives". Deux exemples d'attributs sont présentés ci-dessous.

Fréquence de remplissage : Pour évaluer la fréquence à laquelle un avaloir est susceptible de recevoir des déchets, nous avons créé un attribut nommé *Fréquence de Remplissage*, noté $\overline{F_\zeta}$. Celui-ci correspond à la fréquence des variations supérieures à un certain seuil ζ. Pour le calculer, on définit d'abord :

$$N_{\gamma_i > \zeta} = \sum_i I(\gamma_i > \zeta)\,, \tag{3.3.2}$$

où I est la fonction indicatrice définie par :

$$I(\text{condition}) = \begin{cases} 1 & \text{si condition est vraie,} \\ 0 & \text{si condition est fausse.} \end{cases} \tag{3.3.3}$$

$N_{\gamma_i>\zeta}$ représente le nombre de variations supérieures à ζ. À partir de cela, on définit :

$$\overline{F_\zeta} = \frac{N_{\gamma_i>\zeta}}{N} \tag{3.3.4}$$

avec N le nombre total de variations.

Amplitude de remplissage : Afin d'obtenir une estimation de la taille des déchets entrant dans un avaloir donné, nous avons également défini un attribut correspondant à la moyenne empirique des variations supérieures à un seuil ζ :

$$\overline{A_\zeta} = \begin{cases} \dfrac{\sum_i \gamma_i \cdot I(\gamma_i > \zeta)}{N_{\gamma_i>\zeta}} & \text{si } N_{\gamma_i>\zeta} > 0 \\ \zeta & \text{sinon} \end{cases} \tag{3.3.5}$$

Afin d'illustrer ces deux attributs nous donnons en Figure 3.1 un exemple de données artificielles (pour 60 mesures prises à une fréquence de 2 mesures par jour). La Figure 3.2 représente les variations γ_i calculés pour cet exemple et la Figure 3.3 représente la distribution de ces variations. Dans cet exemple, le niveau d'encrassement augmenterait respectivement de ~ 10, ~ 6, ~ 10 et ~ 14 cm aux alentours des jours 4, 8, 17.5 et 25 (indiqué par les flèches vertes sur la Figure 3.1). Comme la période entre deux mesures est de 12 heures, les variations γ_i associées sont respectivement de ~ 20, ~ 12, ~ 20 et ~ 28 cm par jour (indiqué par les flèches vertes sur la Figure 3.2).

Dans cet exemple, en choisissant un seuil $\zeta = 10$ cm/jour, la fréquence de remplissage correspond au rapport entre le nombre de variations supérieures au seuil choisi (indiqué comme étant à droite des pointillés rouges sur la Figure 3.3), soit 4 variations dans cet exemple et le nombre de variations $(60 - 1)$, soit : $\overline{F_{10}} = 4/(60-1) \simeq 6.8\%$.

L'amplitude de remplissage correspond à la somme des variations supérieures à ζ divisée par le nombre de variations supérieures à ζ soit : $\overline{A_{10}} \simeq (20{+}13{+}20{+}28)/4 \simeq 20.3$ cm/jour.

À partir de ces deux attributs, on peut interpréter que cet avaloir (simulé) se remplit $\sim 6.8\%$ du temps de déchets supérieurs à 10 cm en hauteur, causant en moyenne, une variation de 20.3 cm/jour des mesures au moment du remplissage.

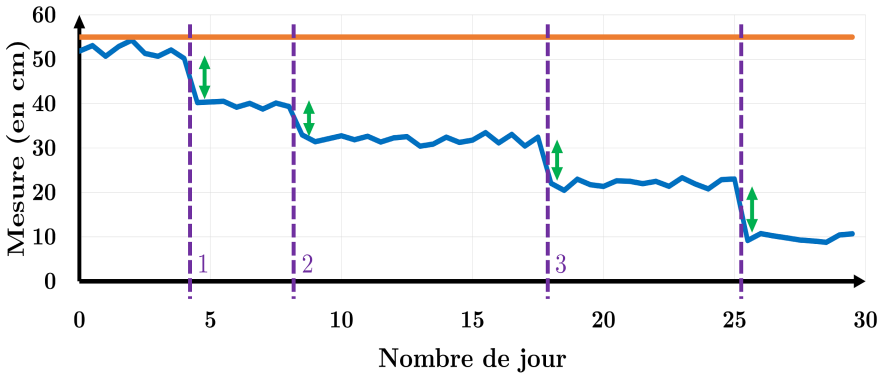

FIGURE 3.1 – *fréquence de remplissage* et *amplitude de remplissage* sur pour courbe simulée.

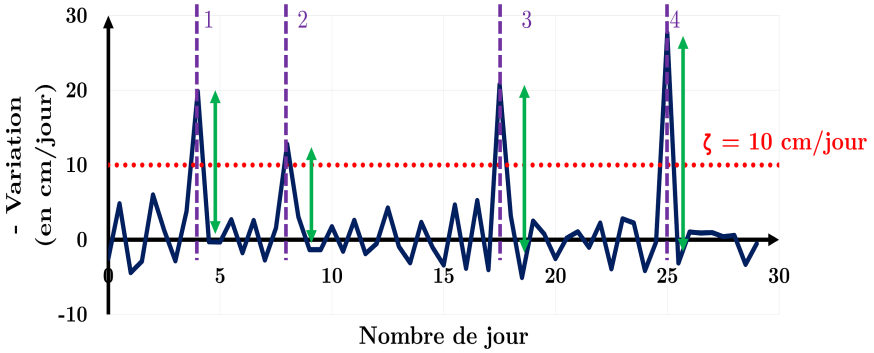

FIGURE 3.2 – *fréquence de remplissage* et *amplitude de remplissage* pour la courbe de la Figure 3.1.

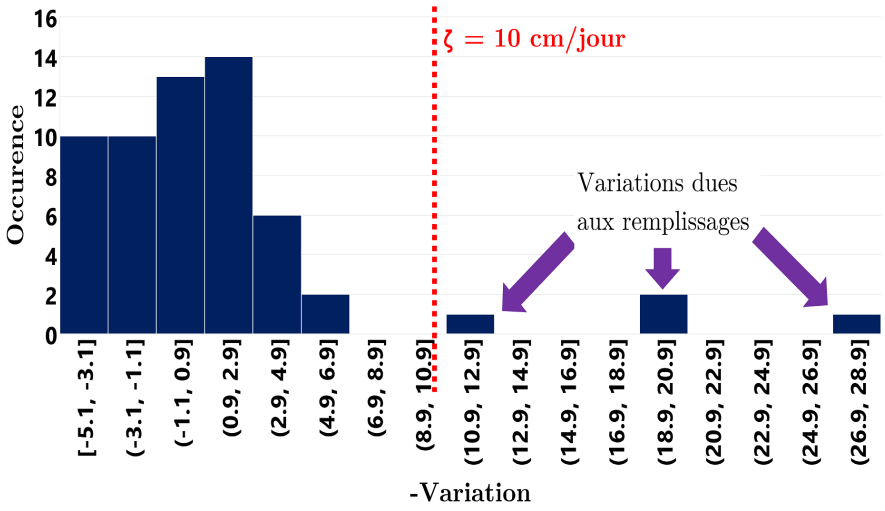

FIGURE 3.3 – Distribution associée aux variations de l'exemple de la Figure 3.2. Les variations qui nous intéressent pour le calcul des attributs *fréquence de remplissage* et *amplitude de remplissage* sont à droite du seuil ζ.

Synthèse des attributs empiriques Les attributs élaborés, ainsi que les aspects qu'ils mesurent, sont présentés ci-dessous. Une description plus détaillée de la construction de ces attributs est fournie en Annexe C.1.

Comme mentionné précédemment, nous avons développé des attributs pour mesurer la fréquence et l'amplitude de remplissage d'un l'avaloir (afin d'évaluer la taille moyenne des déchets y entrant). Ces attributs sont calculés pour différents seuils, à savoir respectivement $\zeta = 2.5\,\text{cm}, 5\,\text{cm}, \text{et } 10\,\text{cm/jour}$.

Un autre attribut intéressant est la vitesse moyenne de remplissage, qui correspond à la moyenne des variations positives de l'encrassement ($\gamma_i > 0$) sur l'ensemble des mesures. On notera cependant que cet attribut est sensible à les faibles variations issues de la variabilité intrinsèque des mesures. En d'autres termes, avec cet attribut, une partie des petites fluctuations des mesures sont considérées comme de légères augmentations du niveau d'encrassement. Idéalement, ces variations intrinsèques ne devraient pas être prises en compte dans son calcul. Cependant, comme mentionné dans la Section 2.1.6, la variabilité des mesures est influencée par la forme de la surface mesurée par le capteur, qui peut elle-même varier dans le temps. C'est pourquoi éliminer cette variabilité s'avère particulièrement complexe. Nous avons donc opté pour la sommation de toutes les variations positives du niveau d'encrassement.

D'autre part, le retour d'expérience terrain suggère que certains avaloirs sont encrassés périodiquement. Par exemple, les avaloirs situés à proximité des plages et du littoral connaissent un encombrement accru pendant la période estivale.

Un autre cas est celui des jours de la semaine, liées aux marchés hebdomadaires, qui peuvent générer des objets ou déchets abandonnés en fin de journée. Nous avons examiné trois facteurs : l'influence de la saison, le jour de la semaine, et l'impact des jours ouvrés par rapport aux week-ends. Nous avons associé à chaque variation enregistrée la période étudiée et réalisé une analyse de variance ou ANOVA (*Analysis Of Variance*).

L'ANOVA est un test statistique utilisé pour déterminer si les moyennes de différents groupes sont significativement différentes. Ce test génère une *p-value*, qui peut être interprétée comme la probabilité d'obtenir un certain résultat sous l'hypothèse nulle. En théorie, l'ANOVA repose sur plusieurs hypothèses : les observations au sein de chaque groupe doivent être indépendantes, les données de chaque groupe doivent suivre une distribution gaussienne, et les variances entre les différents groupes doivent être égales. En pratique, ces conditions ne sont pas toujours strictement respectées. Cependant, lorsque la taille de l'échantillon dans chaque groupe est suffisamment importante (souvent considérée comme supérieure à 30), il est généralement admis que ces hypothèses sont respectées grâce au Théorème Central Limite (TCL). Ce théorème indique que la moyenne d'un échantillon tend à suivre une distribution gaussienne à mesure que la taille de l'échantillon augmente, même si les données d'origine ne le sont pas. Autrement dit, ces hypothèses sont asymptotiquement vérifiées lorsque la quantité de données augmente.

Par exemple, pour évaluer l'impact de la saison, nous supposons que celle-ci n'influence pas les variations mesurées. La *p-value* permet de quantifier la vraisemblance des variations observées. Si la *p-value* est faible, on considère que l'hypothèse nulle est rejetée. En pratique, nous avons utilisé les *p-values* obtenues comme attributs pour évaluer l'impact des trois facteurs mentionnés précédemment : l'influence de la saison, du jour de la semaine, et de la distinction entre jours ouvrés et week-ends.

Nous avons également élaboré des attributs décrivant la distribution en calculant les différents quantiles des variations positives en déchets ($\gamma_i > 0$). Les quantiles choisis sont les 25^{e}, 50^{e}, 75^{e} et 90^{e} percentiles.

Finalement, à ce stade des travaux, les 13 attributs suivants ont été construits :

- fréquence de remplissage :$\overline{F_\zeta}$ avec $\zeta = 2.5\,\text{cm}, 5\,\text{cm}, 10\,\text{cm/jour}$;
- amplitude de remplissage : $\overline{A_\zeta}$ avec $\zeta = 2.5\,\text{cm}, 5\,\text{cm}, 10\,\text{cm/jour}$;
- *P-values* pour étudier la saison, le jour de la semaine, la différence entre jour ouvré et week-end ;
- Percentiles des variations positives ($\gamma_i > 0$) : 25^{e}, 50^{e}, 75^{e}, et 90^{e}.

3.4 Correction et exploration des attributs empiriques

3.4.1 Sous-échantillonage des données

Comme illustré précédemment, les attributs que nous avons construits sont basés sur la distribution de la variation des mesures. Cependant, ces variations dépendent de l'intervalle de temps ΔT_i entre deux mesures, cf. l'équation 3.3.2. Il convient de rappeler que, même si la période nominale entre deux mesures est de $\Delta T_i = 12$ heures, celle-ci peut être modifiée par l'opérateur. En pratique, l'historique des mesures d'un capteur peut contenir des mesures prises à des intervalles de temps différents. Par conséquent, la distribution des variations des mesures, et donc les attributs construits, seront affectés par ces changements de fréquence de mesure imposés par l'opérateur.

Pour mieux comprendre ce problème, nous présentons en Figure 3.4 un exemple basé sur des données simulées qui correspond à des mesures prises à une fréquence de une mesure par heure, avec un remplissage d'environ 20 cm le jour 15. La Figure 3.5 montre la distribution associée à cet exemple, où on peut observer cette variation due au remplissage. Plus précisément, la variation associée correspond à un remplissage de 20 cm observé entre deux mesures avec une période de mesure de une heure, soit une variation de 20 cm/heure. Ensuite, en reprenant ce même exemple mais en supposant cette fois que la fréquence de mesure a été reconfigurée par l'opérateur à une mesure toutes les 12 heures à partir du jour 12, on constate, sur les Figures 3.6 et 3.7, que le remplissage n'apparaît plus dans la distribution des variations de mesures. En effet, un remplissage de 20 cm sur une période de 12 heures se traduit par une augmentation moyenne d'environ 1.7 cm par heure, ce qui se confond avec les variations des mesures observées entre les jours 0 et 12, période pendant laquelle la fréquence était d'une mesure par heure. Autrement dit, le changement de fréquence influence la forme de la distribution des variations mesurées.

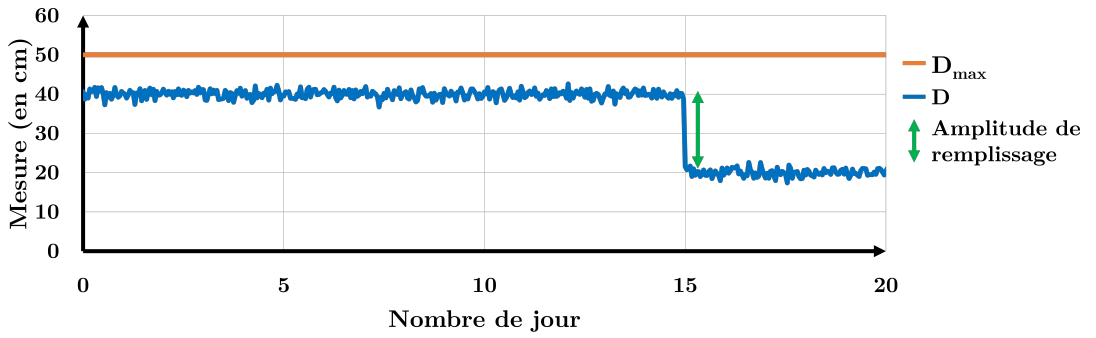

FIGURE 3.4 – Simulation de mesures à fréquence unique (une mesure par heure) avec un remplissage d'environ 20 cm aux alentours du jour 15.

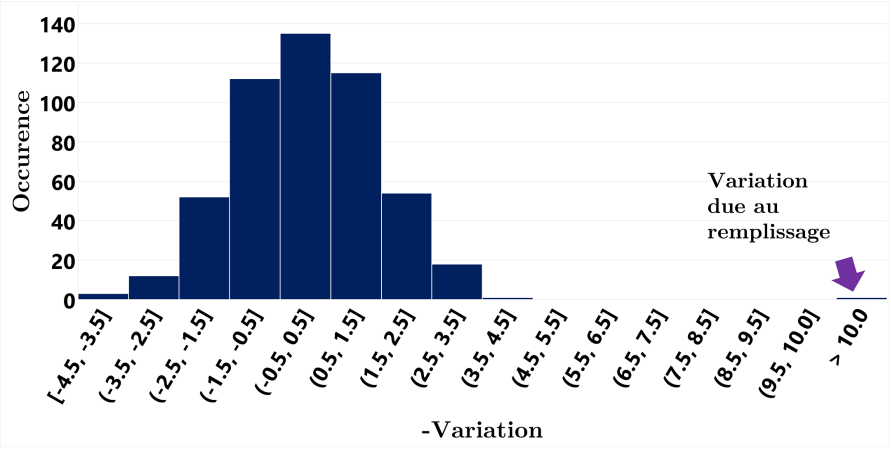

FIGURE 3.5 – Distribution des variations des mesures associée à l'exemple de la Figure 3.4. La variation liée au remplissage du jour 15 apparaît bien sur la distribution.

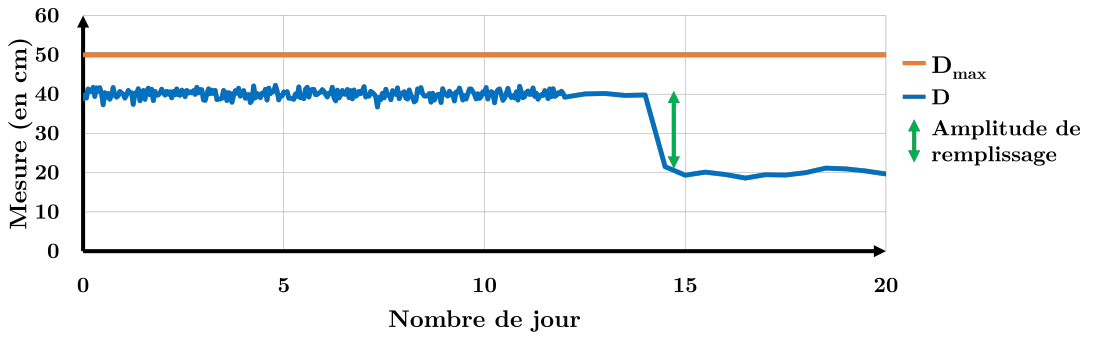

FIGURE 3.6 – Simulation de mesures à fréquence variable pour l'exemple de la Figure 3.4 : la fréquence de mesure est de une mesure par heure entre les jours 0 et 12 puis de une mesure toutes les 12 heures. On retrouve bien le remplissage d'environ 20 cm aux alentour du jour 15 sur le graphique.

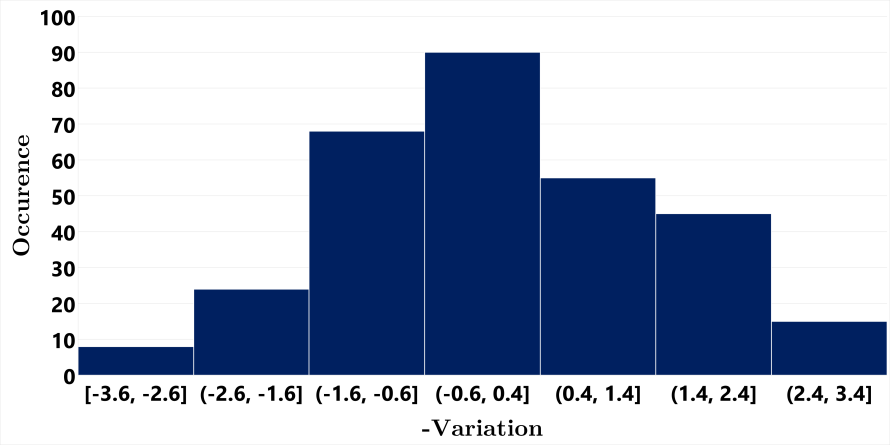

FIGURE 3.7 – Distribution des variations de mesures associée à l'exemple de la Figure 3.6. On constate que la variation liée au remplissage du jour 15 n'apparaît plus sur la distribution.

Pour éviter ce problème, nous avons opté pour un sous-échantillonnage des mesures à une fréquence d'une mesure par jour afin d'uniformiser les distributions des variations de mesures. De cette manière, toutes les variations étudiées seront des variations journalières et pourront être comparées entre elles. Nous avons opté pour une fréquence d'une mesure par jour car on considère que l'on dispose pour chaque capteur, d'au moins une mesure fiable par jour en pratique.

Plus précisément, nous avons choisi de retenir la mesure de la distance D journalière la plus élevée, correspondant au niveau d'encrassement minimal. Ce critère a été retenu car, du point de vue opérationnel, on considère le niveau minimal

d'encrassement comme référence, ce qui permet en particulier d'éviter les fausses alertes liées aux déchets ayant une forme élancée. On peut imaginer le cas d'une bouteille qui serait placée verticalement dans un avaloir. Dans ce cas, le capteur mesure parfois l'écho correspondant au sommet de cette bouteille, rajoutant donc quelques dizaines de centimètres au niveau d'encrassement (cf. Figure 2.9). En sélectionnant le niveau d'encrassement minimal, on s'assure que les interventions de maintenance sont véritablement nécessaires. L'ensemble des travaux qui suivent est basé sur ces données sous-échantillonnées.

3.4.2 Correction des tests

Lors de notre analyse, les tests ANOVA effectués pour certains avaloirs rendent des p-values inférieures au seuil choisi (5%). Ces résultats suggèrent donc que la dynamique d'encrassement de ces avaloirs est influencée par la saison ou par les jours de la semaine. Cependant, pour un critère donné (l'influence de la saison, par exemple), comme nous effectuons un test pour chaque avaloir, cela correspond à plus de 2000 tests réalisés. Or, lorsque l'on effectue des tests statistiques multiples, comme c'est le cas ici, une correction des p-values est nécessaire pour minimiser les erreurs de type I (rejeter à tort l'hypothèse nulle), car le risque d'obtenir une p-value faible par hasard augmente avec le nombre de tests réalisés. Quelques méthodes courantes de correction sont présentées ci-dessous.

- La correction de Bonferroni est l'une des méthodes les plus simples pour gérer le problème des comparaisons multiples [42]. Elle consiste à multiplier la p-value obtenue par le nombre total de tests réalisés. Si le produit dépasse 1, la p-value corrigée est fixée à 1. Notez qu'en pratique cela est équivalent à diviser le seuil choisi (généralement 5%) par le nombre de tests. Cette méthode est utile lorsque le nombre de tests à réaliser est relativement faible, que l'erreur de type I doit être strictement contrôlée et est plutôt adaptée lorsque les tests sont indépendants.
- La correction de Holm (ou procédure de Holm-Bonferroni) est une amélioration de la correction de Bonferroni [43]. Elle trie les p-values en ordre croissant et les compare à $\alpha/(m-i+1)$ où i est le rang de la p-value (en commençant par $i = 1$ pour la plus petite p-value), m est le nombre total de tests, et α est le risque d'erreur (généralement 5%). Cette méthode est intéressante lorsque l'on a besoin d'une approche moins conservatrice que la correction de Bonferroni mais toujours basée sur le contrôle strict de l'erreur de type I. Elle est adaptée aux situations où les tests sont indépendants ou légèrement dépendants.
- La correction de Benjamini-Hochberg se concentre sur le contrôle du taux de fausses découvertes, c'est-à-dire le pourcentage attendu de rejets incorrects

de l'hypothèse nulle [44]. Après avoir trié les p-values, elle identifie la plus grande p-value qui est inférieure ou égale à $\frac{i}{m} \cdot \alpha$, où i est le rang de la p-value, m est le nombre total de tests, et α est le risque d'erreur (généralement 5%). Toutes les p-values inférieures ou égales à cette valeur sont considérées comme significatives. Cette méthode est souvent moins conservatrice que les deux précédentes. Elle est utile dans des contextes où des milliers de tests sont réalisé et où il est acceptable d'avoir quelques faux positifs en échange d'une plus grande puissance pour détecter les vrais effets.

Dans notre cas, en supposant que les avaloirs sont indépendants et n'ayant pas besoin d'un contrôle stricte de l'erreur de type I et étant dans une démarche d'exploration, nous avons opté pour la correction de Benjamini-Hochberg.

Après correction des p-values pour les tests ANOVA multiples, en analysant l'impact de la saisonnalité, nous avons constaté que seuls les avaloirs présentant une activité faible affichaient des p-values significatives. Pour ces avaloirs, nous avons observé un ou deux cas d'encrassement au cours d'une saison donnée, sans autre activité notable le reste de l'année. Comme nous le pressentions, la base de données historiques s'avère trop réduite pour conduire une analyse robuste de l'impact de la saison car la majorité des capteurs disposent d'un historique de mesure à peine supérieure à une année, chaque saison n'est donc représentée qu'une seule fois (un seul hiver, un seul été, etc.). Par conséquent, nous avons choisi d'écarter les attributs concernant l'impact de la saison pour la suite.

En ce qui concerne les deux autres facteurs (l'influence du jour de la semaine ou de la comparaison jour de la semaine par rapport aux week-ends), aucune des p-values corrigées n'est inférieure au seuil fixé. Nous avons tout de même examiné les données des avaloirs les plus significatifs (ceux qui avaient montré des p-values inférieures au seuil d'erreur de 5% avant la correction). Cependant, il est visuellement difficile de conclure à une influence claire de ces éléments.

3.4.3 Nettoyage supplémentaires des données

Dans notre étude des attributs calculés, en particulier ceux relatifs à la fréquence de remplissage, à l'amplitude de remplissage et aux percentiles fondés sur les fluctuations des mesures, nous avons détecté certaines valeurs anormalement élevées. Par exemple, nous avons noté des avaloirs avec une $\overline{F_\zeta}$ et une $\overline{A_\zeta}$ indiquant des augmentations de l'encrassement de plusieurs dizaines de centimètres presque quotidiennement. Bien que possible, ce phénomène est peu commun et ne serait plausible que dans le cas où l'avaloir est soumis à un curage très fréquent ou possède une profondeur exceptionnelle. Sans cela, l'avaloir serait rapidement saturé, ce qui stabiliserait les mesures du capteur à niveau constant indiquant cette saturation.

En examinant de plus près les données des capteurs concernés, nous avons

identifié un ensemble de capteurs dont les mesures semblaient peu fiables en raison d'une grande variabilité des mesures ou d'éventuels problèmes d'installation, comme expliqué précédemment en Section 2.1.5. Environ 160 de ces capteurs ont été exclus de la suite de notre étude. Ainsi, ces attributs ont également été utiles pour détecter les anomalies dans les données des capteurs.

Finalement, conscients que regrouper les avaloirs en fonction de leur dynamique d'encrassement exclusivement à partir de ces attributs empiriques pourrait négliger certains aspects, nous avons décidé de créer d'autres attributs afin de compléter les informations extraites. Ces nouveaux attributs sont soit issus de la modélisation de la distribution des γ_i, soit directement basés sur les mesures elles-mêmes.

3.5 Attributs inférentiels et d'excursions

3.5.1 Attributs inférentiels

Pour compléter notre analyse de la dynamique, nous avons ajouté à nos attributs empiriques de nouveaux attributs qui se concentrent sur la distribution des γ_i, cherchant ainsi à capturer des comportements plus globaux. Plus précisément, nous avons utilisé un modèle statistique qui postule la présence de trois lois différentes (voir la Figure 3.8) : une loi gaussienne $\mathcal{G}_V$ reflétant la variabilité intrinsèque du signal ; une deuxième loi gaussienne $\mathcal{G}_D$ pour les variations liées à la dynamique de l'avaloir, comme le remplissage ou le lessivage ; et enfin, une loi uniforme $\mathcal{U}$ pour les variations exceptionnelles ou extrême, par exemple lorsque l'avaloir est subitement complètement rempli ou vidé. Ce choix des distributions associées à chacun de ces phénomènes est basé sur les observations qui ont pu être faîtes sur les distributions des variations et des connaissances a priori dont on dispose à propos des variations de mesures du point de vue opérationnel. Dans ce modèle, chaque variation observée est supposée suivre l'une de ces trois lois avec une certaine probabilité. Ces probabilités, ainsi que les paramètres spécifiques à chaque loi, sont estimés en utilisant l'algorithme *Expectation-Maximization* (EM) [45]. Ces probabilités et paramètres sont ensuite considérés comme des attributs supplémentaires pour décrire la dynamique en question.

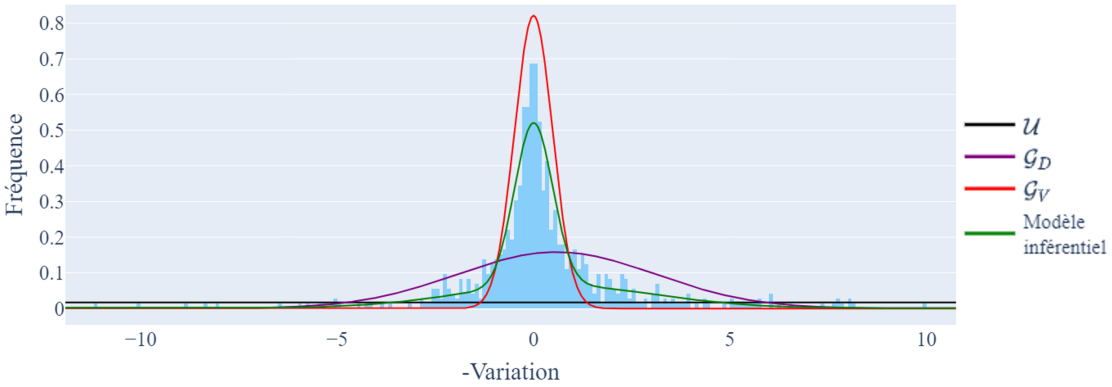

FIGURE 3.8 – Exemple d'estimation des caractéristiques du modèle avec l'algorithme EM sur des données réelles. Les trois lois de notre modèle $\mathcal{U}$, $\mathcal{G}_D$ et $\mathcal{G}_V$ sont respectivement représentées en noir, en violet et en rouge. La courbe verte représente le modèle global ajusté à la distribution de la variation des mesures du capteur étudié.

L'algorithme Expectation-Maximization (EM) est une méthode itérative d'optimisation utilisée pour estimer les paramètres d'un modèle de mélange probabiliste lorsque les données sont incomplètes. Cet algorithme fonctionne en deux étapes : la phase "Expectation" (E) et la phase "Maximization" (M). Dans l'étape E, on utilise les données disponibles et notre estimation actuelle des paramètres pour estimer les données manquantes. Dans l'étape M, on utilise les estimations précédentes afin de mettre à jour les paramètres du modèle. Ces deux étapes sont ensuite répétées jusqu'à ce que le modèle converge. Plus précisément, nous appliquons l'algorithme de la manière suivante.

- **Initialisation des paramètres des lois** :
  Les paramètres de chacune des trois lois de notre modèle $\mathcal{G}_V$, $\mathcal{G}_D$, $\mathcal{U}$, sont respectivement notés $(\mu_{\mathcal{G}_V}, \sigma_{\mathcal{G}_V})$, $(\mu_{\mathcal{G}_D}, \sigma_{\mathcal{G}_D})$, et $(a_{\mathcal{U}}, b_{\mathcal{U}})$. Nous imposons à l'algorithme EM, les contraintes suivantes vis-à-vis des paramètres :
  - $\mu_{\mathcal{G}_V}$ est fixé à 0 car il représente la variabilité moyenne intrinsèque à notre signal (ou bruit de mesure)
  - On fixe $a_{\mathcal{U}} = -D_{\max}$ et $b_{\mathcal{U}} = +D_{\max}$ car les variations extrêmes ne peuvent pas dépasser la profondeur de l'avaloir. Autrement dit, dans les cas les plus extrêmes, l'avaloir peut soudainement se remplir ou se vider complètement.

  Les autres paramètres sont initialisés aléatoirement, de sorte à respecter les ordres de grandeur empirique y étant associés. Par exemple, $\mu_{\mathcal{G}_D}$ représente le remplissage journalier moyen d'un avaloir. Empiriquement, on s'attend à ce que cette valeur soit de l'ordre de quelques centimètres. En pratique, nous avons initialisé cette valeur en effectuant un tirage aléatoire d'une loi

gaussienne de moyenne 4 et d'écart-type 1, de sorte que la valeur tirée se situe généralement entre 1 et 7. L'Annexe C.1 présente plus en détails les choix faits pour chaque initialisation.

- **Étape E** :
  Nous estimons, pour chacune des variations de la distribution, la probabilité d'appartenance aux trois lois respectives étudiées, en prenant en compte les paramètres initialisés précédemment.
- **Étape M** :
  On utilise ensuite les probabilités estimées pour mettre à jour les paramètres des différentes lois.
- **Déroulement de l'algorithme** :
  On alterne E et M à chaque étape en injectant les quantités mises à jour. L'algorithme se termine lorsque les paramètres des lois convergent numériquement, c'est-à-dire que la variation des paramètres d'une itération à l'autre est inférieure à un seuil suffisamment petit, fixé ici à 10^{-4} ici.

3.5.2 Attributs basés sur les excursions

Pour garantir que tous les comportements potentiels sont pris en compte, nous avons étendu notre analyse afin d'inclure non seulement des attributs basés sur la variabilité des mesures, mais aussi deux attributs supplémentaires axés sur le suivi des excursions, c'est-à-dire des périodes pendant lesquelles les mesures dépassent un seuil prédéfini, ce qui permet de conserver la dimension temporelle des données.

Le premier attribut, noté ε_N compte le nombre de fois où le niveau minimal d'encrassement de l'avaloir sur une journée dépasse un certain seuil. Ce nombre est ensuite rapporté à la quantité de données disponibles, c'est-à-dire au nombre total de jours. La Figure 3.9 illustre le calcul de cet attribut pour un seuil correspondant à 30% de remplissage de l'avaloir.

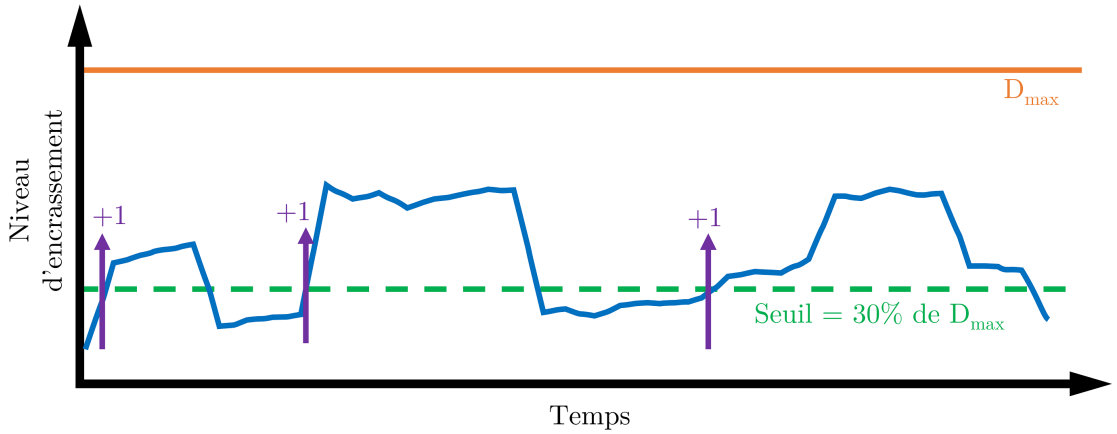

FIGURE 3.9 – Illustration du calcul de l'attribut ε_N qui compte les excursions (nombre de fois où le niveau d'encrassement dépasse un seuil fixé ici à 30% de $D_{\max}$).

Le second attribut noté ε_T évalue la durée maximale en jours consécutifs pendant laquelle le niveau d'encrassement est resté supérieur à un seuil. Cette durée est également rapportée à la quantité de données historiques disponibles. La Figure 3.10 illustre le calcul de cet attribut pour un seuil de dépassement de 30% de la distance de fond.

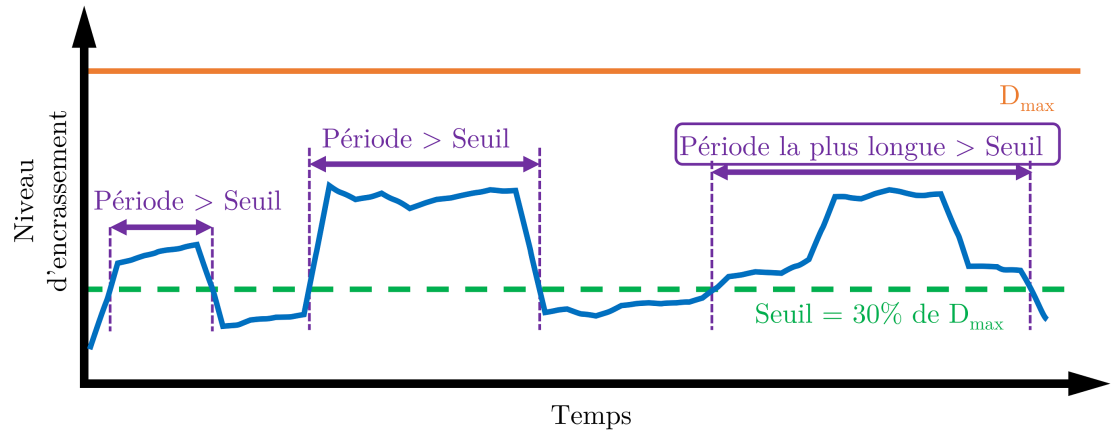

FIGURE 3.10 – Calcul de l'attribut ε_T qui mesure l'excursion la plus longue (période la plus longue pendant laquelle le niveau d'encrassement est supérieur à un seuil fixé ici à 30% de $D_{\max}$).

Dans notre cas, un seuil de 30% de la distance de fond $D_{\max}$ a été retenu pour les excursions. Ce seuil a été empiriquement fixé pour éviter à la fois que des avaloirs actifs ne soient constamment signalés et afin d'éviter d'ignorer des situations d'encrassement significatives, en particulier, pour les avaloirs les plus

profonds. Une valeur de 30% est suffisante à signaler une accumulation notable sans être déclenché par de faibles niveaux d'encrassement, permettant ainsi de distinguer les avaloirs peu actifs des autres.

Enfin, il est important de noter que ces attributs basés sur les excursions ont un inconvénient : ils dépendent de la maintenance des avaloirs. Par exemple, un avaloir particulièrement sensible à l'encrassement pourrait présenter un ε_N faible et un ε_T long s'il n'est pas curé régulièrement car l'encrassement va s'accumuler (et donc rester systématiquement au dessus du seuil choisi). À l'inverse, un avaloir pourrait avoir un ε_N élevé et un ε_T court si celui-ci est nettoyé fréquemment. Pour ces raisons, nous vérifierons l'impact des curages sur les résultats obtenus en Section 3.11.

Finalement, à ce stade des travaux, nous avons créé un large éventail d'attributs afin de ne pas négliger des détails importants dans les données disponibles. Cependant, cette approche peut mener à des redondances et au risque de surapprentissage (*overfitting*), car certaines caractéristiques auront un poids plus important que d'autres, ce qui peut diminuer l'efficacité du modèle. Par conséquent, il est bénéfique de réduire le nombre d'attributs en ne gardant que ceux qui sont réellement pertinents.

3.6 Sélection et préparation des attributs

3.6.1 Sélection d'attributs

Dans le contexte de l'apprentissage automatique, des techniques de réduction de dimension sont parfois utilisées pour réduire le nombre d'attributs d'un ensemble de données, tout en conservant les informations essentielles à analyser. Il existe deux catégories principales de techniques de réduction de dimension : la sélection d'attributs et l'extraction d'attributs (*feature selection* et *feature extraction*). La sélection d'attributs consiste à choisir un sous-ensemble des attributs originaux qui représente au mieux l'ensemble de données, tandis que l'extraction d'attributs consiste à créer de nouveaux attributs qui sont des combinaisons linéaires ou non linéaires des attributs originaux.

Cependant, l'extraction d'attributs peut poser des problèmes d'interprétabilité car les attributs d'origine sont modifiés ou combinés pour créer de nouveaux attributs, ce qui peut rendre difficile leur interprétation. Dans notre cas, notre objectif a été de simplifier autant que possible l'interprétabilité des classes obtenues à terme. C'est pourquoi la méthodologie présentée ici se concentre sur la sélection d'attributs. Il est évident que faire les bons choix dans la sélection d'attributs est essentiel pour obtenir des résultats pertinents. Dans le but de fournir un rapide

aperçu de ces techniques, on présente ici quelques méthodes de sélection d'attributs [46, 47], classés en différentes catégories :

- Les méthodes de filtrage qui attribuent un score à chaque attribut, soit en se basant sur une mesure statistique ou soit un critère heuristique liés à sa pertinence par rapport à une variable cible, soit en comparant les attributs entre eux.
- Les méthodes intégrées (*embedded methods*) qui intègrent le processus de sélection dans l'apprentissage du modèle lui-même. Ces méthodes optimisent à la fois les performances du modèle et la sélection des attributs, en les intégrant dans l'algorithme d'apprentissage. Autrement dit, elles évaluent le poids des attributs pendant le processus d'apprentissage, permettant d'identifier efficacement les attributs les plus pertinents pour la tâche cible.
- Les méthodes enveloppantes (*wrapper methods*) qui sélectionnent des sous-ensembles d'attributs en les évaluant conjointement avec un algorithme d'apprentissage ou un modèle spécifique. Contrairement aux méthodes de filtrage qui évaluent les attributs de manière indépendante, les méthodes enveloppantes utilisent les performances de l'algorithme d'apprentissage comme critère d'évaluation pour la sélection des attributs. Elles considèrent la sélection d'attributs comme un problème de recherche, explorant différentes combinaisons pour trouver le sous-ensemble optimal qui maximise les performances du modèle. Ces méthodes visent à tester un grand nombre de combinaisons possibles d'attributs. Elles nécessitent souvent un temps de calcul plus long et peuvent conduire à un surapprentissage [2], surtout si le nombre de combinaisons d'attributs à évaluer est très élevé ou si la quantité de données disponibles est limitée, ce qui ne permet pas de généraliser les résultats à de nouvelles données.

Dans notre cas, afin de réduire l'effort humain et le temps de calcul de la sélection d'attributs, nous nous sommes concentrerons sur les méthodes de filtrage. Parmi les métriques classiques utilisées dans le cadre des méthodes de filtrage appliquées aux problèmes de clustering (où les étiquettes ne sont pas connues), on trouve en particulier le coefficient de corrélation (qui permet d'évaluer la similarité statistique entre deux attributs), et l'information mutuelle (qui permet d'estimer la quantité d'information partagée entre deux attributs). En utilisant ces métriques pour regrouper les attributs similaires et en choisissant de la sorte quelques attributs représentatifs pour décrire l'ensemble du groupe, il est possible de réduire la redondance des informations.

2. On appelle surapprentissage le cas où le modèle d'apprentissage automatique s'adapte trop précisément aux données d'entraînement, perdant ainsi sa capacité à généraliser et à effectuer correctement des prédictions sur de nouvelles données non vues auparavant.

3.6.2 Application de la sélection d'attributs

Une sélection des attributs a été effectuée en fonction de leur corrélation : les attributs fortement corrélés ont été regroupés au sein de la même famille. Seul un ou deux attributs représentatifs de chaque famille sont conservés par la suite. Dans cette étude, deux attributs sont considérés comme corrélés si la valeur absolue du coefficient de corrélation est supérieure à 0.85. Ce seuil représente un bon compromis car il permet de regrouper les attributs qui sont significativement liés, tout en offrant la flexibilité nécessaire pour identifier des familles d'attributs. Un seuil plus bas risquerait de regrouper des attributs présentant une corrélation marginale, qui pourraient porter des informations distinctes. À l'inverse, un seuil trop élevé empêcherait le regroupement efficace des attributs. La Figure 3.11 est la matrice de corrélation qui décrit les attributs liés. Pour des raisons de clarté, seul un extrait de cette matrice est présenté (matrice complète de taille $P \times P$). De plus, les attributs ont été renommés en f_i, avec $i \in 1, 2, ..., P$.

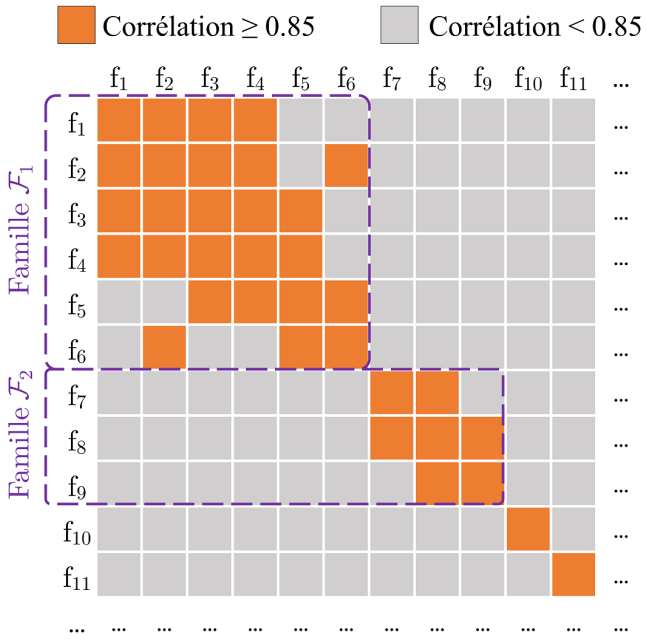

FIGURE 3.11 – Matrice de corrélation des attributs qui met en évidence deux lots d'attributs.

Comme on peut le constater sur la Figure 3.11, deux principales familles d'attributs peuvent être extraites : une famille regroupant les attributs f_1 à f_6 (nommée $\mathcal{F}_1$) et une autre pour les attributs f_7 à f_9 (appelée $\mathcal{F}_2$).

La famille $\mathcal{F}_1$ regroupe les attributs de fréquence de remplissage $\overline{F_\zeta}$, les attributs correspondants aux percentiles, ainsi que la vitesse de remplissage. La famille $\mathcal{F}_2$

regroupe les trois attributs liés à l'amplitude de remplissage $\overline{A_\zeta}$.

Le reste des attributs est considéré comme unique. Les relations entre les attributs de chaque famille sont synthétisées dans la Figure 3.12, où les flèches représentent une corrélation supérieure à 0.85.

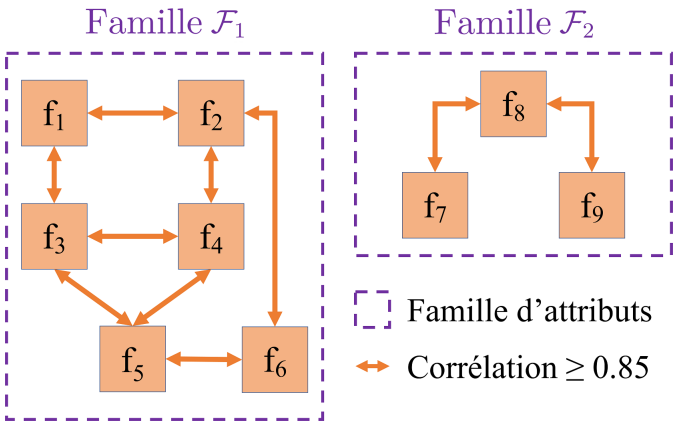

FIGURE 3.12 – Synthèse des familles d'attributs $\mathcal{F}_1$ et $\mathcal{F}_2$

Pour chaque famille, un ou deux attributs peuvent être choisis pour la représenter. Par la suite, le choix d'un ou de deux attributs pour représenter chaque famille sera considéré comme un paramètre.

Nous avons choisi de retenir un représentant d'une famille, l'attribut le plus corrélé avec les autres dans la situation la moins favorable. Formellement, nous choisissons l'attribut f' comme représentant de sa famille k si cet attribut maximise $\min_{f\in\mathcal{F}_k} \mathrm{cor}(f, f')$, avec cor la fonction de corrélation. Par exemple, dans le cas de $\mathcal{F}_1$, f_5 est l'attribut le plus corrélé avec les autres en ce sens, car $\forall i, j \in \{1, 2, ..., 6\}, \min(\mathrm{cor}(f_5, f_j)) \simeq 0.82 \geq \min(\mathrm{cor}(f_i, f_j))$. De même, f_8 peut être choisi comme représentant de $\mathcal{F}_2$.

Lorsque nous souhaitons choisir deux représentants par famille, nous sélectionnons les deux attributs les moins corrélés de chaque famille : pour $\mathcal{F}_1$, ces attributs seraient f_1 et f_6 car $(\forall i, j \in 1, 2, ..., 6, \min(\mathrm{cor}(f_i, f_j)) \geq \mathrm{cor}(f_1, f_6) \simeq 0.7)$. De manière similaire f_7 et f_9 représentent $\mathcal{F}_2$. L'idée derrière ce choix est de conserver les deux attributs ayant le moins d'information redondante possible.

3.6.3 Préparation des attributs

En résumant les données en attributs, il est courant de rencontrer des problèmes liés à la gestion de différentes échelles d'attributs, à l'impact des valeurs extrêmes sur les modèles. Des méthodes de transformation des attributs, telles que celle de *Yeo-Johnson* [48] (décrite plus en détail en Annexe C.2), sont fréquemment utilisées pour

résoudre ces problèmes. Ces transformations adaptent automatiquement l'échelle des caractéristiques et minimisent l'influence des valeurs extrêmes. Dans notre étude, après avoir visualisé les distributions de chaque attribut construit, nous avons appliqué cette transformation aux attributs présentant des problèmes d'échelle ou de valeurs extrêmes, tels que les percentiles. La Figure 3.13 illustre l'application de cette transformation sur l'attribut "Percentile-50", on peut y voir que la distribution originale de l'attribut, représentée en bleu, montre une asymétrie notable. Après l'application de la transformation de Yeo-Johnson, la distribution, maintenant en vert, démontre une symétrie plus grande et une réduction significative des valeurs extrêmes, se rapprochant plus d'une gaussienne.

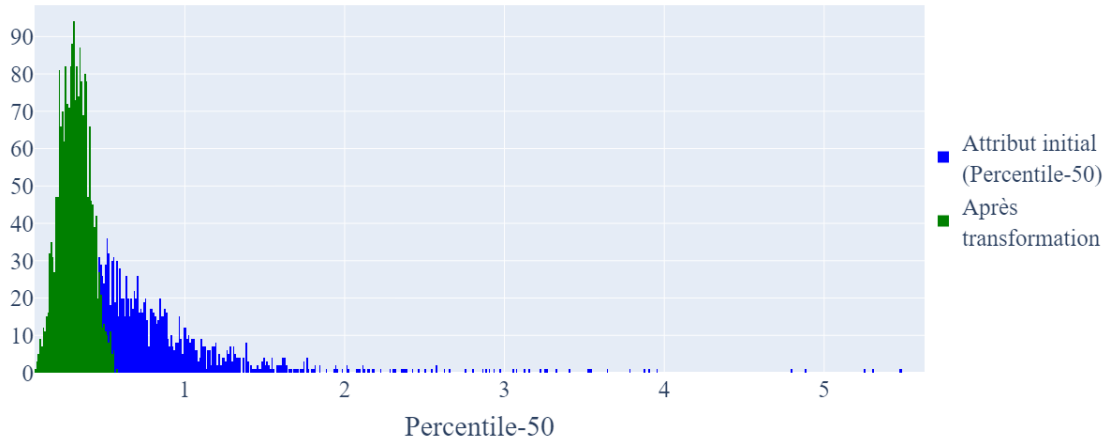

FIGURE 3.13 – Comparaison des distributions de l'attribut "Percentile-50" avant et après la transformation de Yeo-Johnson.

À partir des attributs restants, nous avons créé différents sous-ensemble d'attributs que l'on souhaite essayer pour regrouper les dynamiques à partir d'algorithmes de clustering qui restent à choisir. Ces sous-ensembles sont créées en choisissant 1 ou 2 représentants pour les familles $\mathcal{F}_1$ ou $\mathcal{F}_2$ et différents mélanges des attributs restants en fonction de leur nature (attributs empiriques, statistiques, de saisonalité, etc.). Enfin, notez que certains attributs comme $\mu_{\mathcal{G}_V}$ et $\sigma_{\mathcal{G}_V}$, décrivant la variabilité intrinsèque des mesures, sont écartés car on ne souhaite pas regrouper les avaloirs à partir de cette variabilité qui dépend du capteur et de la surface mesurée (dont entre autres la forme du fond) et non de la dynamique d'encrassement. La liste des sous-ensemble utilisées est disponible en Annexe C.3.

3.7 Choix des algorithmes de clustering

3.7.1 Sélection des algorithmes de manière générale

Les algorithmes de clustering sont utilisés pour regrouper les points de données en clusters en fonction de leur similarité. À titre d'exemple, nous présentons ci-dessous quelques algorithmes de clustering (avec leur complexité calculatoire pour n points de données) [49, 50] :

- L'algorithme *K-means* attribue chaque point de données au centre du cluster le plus proche en utilisant la distance euclidienne, le nombre de clusters étant un hyperparamètre à définir. Sa complexité calculatoire est $O(n \times K \times i \times d)$, où K le nombre de clusters, i le nombre d'itérations, et d le nombre de dimensions.
- L'algorithme *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) regroupe les points de données en fonction de leur densité dans l'espace des attributs. Cet algortihme a deux paramètres qui sont respectivement le rayon ϵ dans lequel les points voisins sont considérés comme faisant partie du même cluster, et le nombre minimum de points requis dans le disque de rayon ϵ autour d'un point pour que celui-ci soit considéré comme un point central. La complexité calculatoire est généralement $O(n \log n)$, mais peut atteindre $O(n^2)$ dans le pire des cas.
- Le *Spectral Clustering* projette les points de données dans un espace de dimension réduite en utilisant la matrice de Laplacien du graphe correspondant, puis les regroupe à l'aide du K-means. Le nombre de clusters, le choix de la mesure de similarité et la dimension de l'espace transformé sont des hyperparamètres à définir. Sa complexité dépend de la décomposition en valeurs propres, souvent $O(n^3)$ pour de petites dimensions.
- Le *Gaussian Mixture Model* (GMM) consiste à modéliser les points de données comme un mélange de plusieurs distributions gaussiennes. Chaque distribution gaussienne représente un "cluster" potentiel dans les données. La complexité de ce modèle, principalement due à l'estimation des paramètres de ces distributions, est $O(n \times k \times d^3)$, où k est le nombre de clusters (nombre de distributions gaussiennes) dans le modèle et d désigne la dimension de l'espace des caractéristiques, c'est-à-dire le nombre de variables ou d'attributs pour chaque point de données.
- Le regroupement hiérarchique construit une hiérarchie de clusters basée sur deux hyperparamètres : le choix des critères de liaison et le nombre de clusters. Sa complexité peut varier de $O(n^2 \log n)$ à $O(n^3)$ en fonction de l'implémentation.

On peut trouver dans la littérature de nombreuses autres méthodes, y compris des variantes de celles mentionnées ci-dessus et des algorithmes plus récents que l' [51, 52, 53].

Le choix de l'algorithme employé est généralement effectué en fonction d'une combinaison de facteurs, incluant la nature des données, les objectifs de l'analyse et la complexité calculatoire. Par exemple, la méthode K-means utilisée avec la distance euclidienne, peut être très performante si les clusters sont sphériques, tandis que le regroupement hiérarchique peut être plus flexible pour identifier des clusters de formes et de tailles variées, mais peut ne pas bien se prêter à des ensembles de données plus volumineux.

Dans les cas où des connaissances préalables ou une compréhension de la structure des données existent, il est possible d'utiliser et de comparer les algorithmes appropriés entre eux. Sinon, il convient de considérer plusieurs algorithmes, y compris des algorithmes classiques, qui reposent sur des principes différents.

3.7.2 Algorithmes retenus

Ici, sans connaissances particulière sur la structure des données, nous avons choisi d'utiliser des algorithmes de clustering classiques : K-means, spectral clustering et DBSCAN, mentionnés précédemment. Ces algorithmes sont basés sur des principes différents, chacun avec ses forces et ses faiblesses. En employant une diversité d'algorithmes, notre objectif est d'acquérir une compréhension plus complète de la structure des données et de réaliser un meilleur clustering.

3.8 Calibration des hyperparamètres

Une fois les attributs et les algorithmes à utiliser établis, la problématique principale est celle de la sélection des hyperparamètres. Dans le domaine du clustering, cette sélection dépend des caractéristiques des données et des objectifs de l'analyse. Une certaine quantité d'expériences et d'essais de configuration sont souvent nécessaire pour trouver les hyperparamètres qui fonctionnent le mieux pour un regroupement spécifique. Dans notre cas, afin d'éviter de devoir visualiser et interpréter les résultats pour chaque ensemble d'hyperparamètres (et ainsi minimiser le temps et les efforts humains), les différents résultats sont évalués à l'aide de scores de clustering. Parmi les scores les plus classiques [53, 54], on retrouve :

- le Silhouette Score qui mesure à quel point chaque échantillon d'un cluster appartient à son propre cluster par rapport aux autres clusters. Il varie de -1 à 1, avec des valeurs plus élevées indiquant des clusters mieux définis.
- l'indice de Davies-Bouldin qui mesure la similarité moyenne entre les clusters et la distance inter-cluster. Une valeur plus basse indique un meilleur

regroupement, avec 0 étant le meilleur score.

- l'indice de Calinski-Harabasz, également connu sous le nom de critère du rapport de variance, qui mesure le rapport entre la dispersion inter-clusters et intra-clusters. Des valeurs plus élevées suggèrent des clusters mieux définis.
- le *Within-Cluster Sum of Squares* (WCSS) qui calcule la somme des distances au carré entre chaque point d'un cluster et son centroïde. Des valeurs basses suggèrent que les clusters sont compacts et bien séparés.

En pratique, pour un jeu d'attributs et un algorithme fixés, cela permet à l'utilisateur de tester différents hyperparamètres et ne sélectionner que les résultats donnant le meilleur score.

Dans notre étude, les algorithmes de clustering sélectionnés sont appliqués sur différents sous-ensembles des attributs, l'objectif étant d'explorer différentes configurations intéressantes. Ces configurations incluent par exemple des ensembles composés uniquement d'attributs empiriques, d'attributs déduits, ou encore des assortiments variés incluants différents types d'attributs. Concrètement, nous expérimentons avec environ une douzaine de configurations d'attributs différentes pour chacun des trois algorithmes de clustering sélectionné. Pour chaque sous-ensemble d'attributs et chaque algorithme, les hyperparamètres sont choisis en fonction de scores de clustering tels que le Silhouette score ou l'indice de Calinski-Harabasz. À titre d'exemple, la Figure 3.14 présente le calibrage de K, le nombre de clusters pour K-means appliqué à un ensemble donné d'attributs. Dans cet exemple, les trois meilleurs Silhouette score étant obtenus pour $K = 2$, $K = 4$, et $K = 6$. L'Annexe C.3 décrit plus en détails les sous-ensembles retenus.

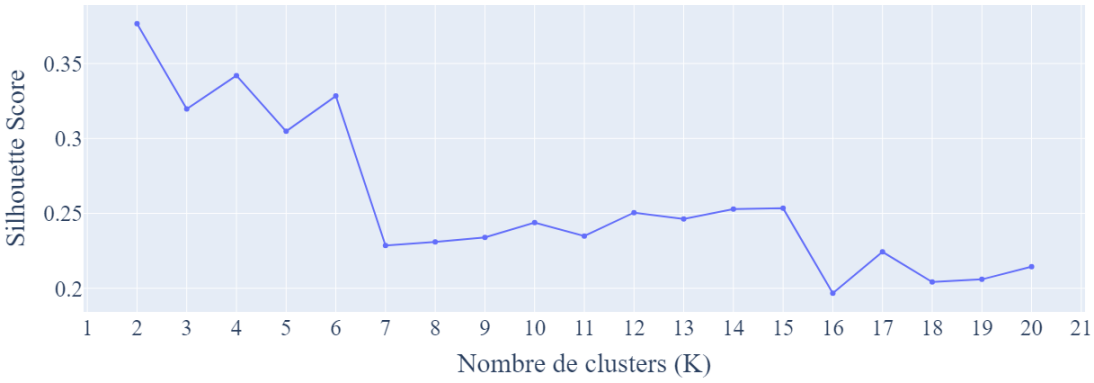

FIGURE 3.14 – Évolution du Silhouette Score en fonction du nombre de clusters (K). Dans cet exemple, les trois meilleurs scores sont obtenus pour $K = 2$, $K = 4$ et $K = 6$.

3.9 Sélection des résultats

Pour chaque jeu d'attributs, on applique chacun des trois algorithmes retenus (cf. Section 3.7.2) en utilisant les hyperparamètres comme décrit précédemment avec l'aide du Silhouette score et/ou de l'indice de Calinski-Harabasz. On appelle par la suite *expérience*, une tentative de clustering utilisant un sous-ensemble d'attributs, un algorithme et un jeu d'hyperparamètres fixés. Chaque expérience résulte donc en un clustering possible des avaloirs. En appliquant cette méthode pour une diversité de configurations, le nombre d'experiences et donc le nombre de résultats peuvent être assez élevés : si on a N_C sous-ensemble d'attributs à tester, N_A algorithmes et N_H ensembles d'hyperparamètres à tester pour chaque algorithme, le nombre total de résultats est $N_C \cdot N_A \cdot N_H$.

Afin de réduire le nombre de résultats à interpréter, une possibilité est de regrouper les résultats similaires. Par exemple, il arrive que deux expériences résultent en deux résultats quasiment identiques ; autrement dit, les regroupements sont effectués de la même manière à quelques détails près. On peut alors considérer que ces deux résultats sont suffisamment similaires pour ne visualiser et interpréter qu'un seul. Regrouper les résultats de manière pertinente nécessite donc de pouvoir mesurer leur similarité. Il existe différents scores que l'on peut utiliser pour comparer les résultats de clustering [55, 56]. Nous en présentons quelques-uns ci-dessous.

- le *Cohen's Kappa*, $\kappa_{\mathcal{C}}$ qui mesure la concordance entre deux *évaluateurs*[3] indépendants qui classent des éléments en catégories, dans le cas où chaque élément ne peut appartenir qu'à une seule catégorie à la fois (sans chevauchement possible entre les catégories). Le score $\kappa_{\mathcal{C}}$ tient compte de la concordance qui pourrait survenir par hasard. Un $\kappa_{\mathcal{C}} = 1$ indique une concordance parfaite, tandis qu'un $\kappa_{\mathcal{C}} = 0$ suggère qu'il n'y a pas plus de concordance que ce qui serait attendu par hasard. Des valeurs négatives de Kappa indiquent que les résultats des évaluateurs sont significativement différents [57].
- l'*Adjusted Rand Index* (ARI) qui quantifie la similarité entre deux ensembles de clusters tout en prenant en compte la possibilité que la concordance entre ces deux ensembles soit due au hasard. Il varie de -1 à 1, avec des valeurs proches de 1 indiquant une forte concordance, 0 indiquant une concordance non significative et des valeurs proches de -1 indiquant des ensembles de clusters significativement différents. Contrairement au $\kappa_{\mathcal{C}}$, l'ARI permet de traiter les situations où les clusters ne sont pas mutuellement exclusifs [58].
- le *Fowlkes-Mallows Index* qui combine la précision et le rappel des résultats de clustering pour mesurer la similarité entre deux ensembles de clusters. [59]

3. Dans notre cas, les évaluateurs correspondent aux différentes experiences réalisées, chaque expérience proposant un résultat de clustering.

Dans notre cas, tout en reconnaissant l'existence d'autres scores pertinents, nous avons opté pour $\kappa_{\mathcal{C}}$ comme score de similarité pour rassembler les résultats similaires. Cependant, ce score ne peut s'appliquer que lorsque le nombre de clusters est identique parmi les deux résultats de clustering. Ainsi, nous utiliserons ce score pour comparer les résultats de clustering ayant le même nombre de clusters dans le but de regrouper les résultats similaires. L'idée étant de ne visualiser et interpréter qu'un seul résultat par famille. Par exemple, la Figure 3.15 montre la comparaison de résultats de clustering à 4 classes, nommés E_i, avec $i \in 1, 2, ..., 13$, sous la forme d'une matrice symétrique [4] où chaque case représente la comparaison d'une paire de résultats. En pratique, on considère que deux résultats sont similaires si $\kappa_{\mathcal{C}} > 0.6$. Le seuil de 0.6 a été choisi car il est généralement considéré comme le point de départ d'une corrélation substantielle selon l'échelle de Landis et Koch [60], qui est fréquemment utilisée pour interpréter les valeurs de $\kappa_{\mathcal{C}}$ [5]. On constate sur cette figure que les résultats de clustering forment 5 "blocs" : $\{E_1$ à E_7 ; E_8 et E_9 ; E_{10} et E_{11} ; E_{12} ; $E_{13}\}$.

Nous avons choisi de ne conserver qu'un seul résultat pour chaque regroupement de résultats afin de minimiser les efforts de visualisation et d'interprétation. Dans l'exemple présenté dans la Figure 3.15, nous avons choisi E_4 comme représentatif du groupe E_1 à E_7, car il est le résultat le plus proche des autres en termes de $\kappa_{\mathcal{C}}$ minimales : $\forall i, j \in \{1, 2, ..., 13\}, \min_j(\kappa_{\mathcal{C}}(E_i, E_j))$ est maximisé pour $i = 4$.

Nous avons ainsi réduit le nombre de résultats de clustering à visualiser et à interpréter.

4. La matrice obtenue est symétrique car $\kappa_{\mathcal{C}}$ est symétrique : $\kappa_{\mathcal{C}}(E_i, E_j) = \kappa_{\mathcal{C}}(E_j, E_i)$, où E_i et E_j sont des résultats de clustering.

5. Cette échelle qualifie les valeurs de $\kappa_{\mathcal{C}}$ comme indiquant une "bonne" concordance si $\kappa_{\mathcal{C}}$ est entre 0.61 et 0.80, et une "très bonne" concordance au delà.

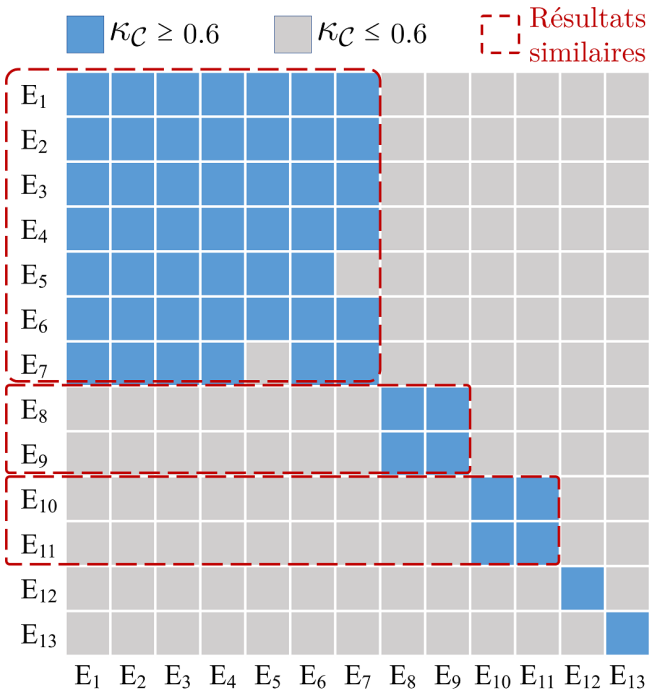

FIGURE 3.15 – Regroupement des résultats similaires en utilisant le Kappa de Cohen κ_C avec un seuil de 0.6.

3.10 Synthèse de la méthodologie de clustering

Avant de présenter en détail les résultats de clustering obtenus et les interprétations correspondant, nous présentons ici une synthèse de la méthodologie adoptée. Pour rappel, cette approche vise à explorer les différentes possibilités de choix d'algorithmes, d'attributs et d'hyperparamètres. À chaque étape, nous avons cherché à minimiser le nombre de combinaisons possibles, dans le but de réduire la quantité de résultats à analyser et de faciliter au maximum l'interprétabilité de ces résultats. La Figure 3.16 expose les différentes étapes de la méthodologie de clustering ainsi que les choix effectués à chaque étape, décrites ci-dessous.

La première étape est celle de la **la création d'attributs**, qui consiste à utiliser des attributs génériques ou à en créer de nouveaux basés sur l'intuition ou des connaissances préalables. Nous avons ainsi créé trois types d'attributs : empiriques, construits pour répondre à des questions opérationnelles ; et inférentiels, qui sont déduits à partir de la distribution de la variation des mesures ; et ceux basés sur des excursions, afin de décrire la dynamique d'encrassement tout en conservant la dimension temporelle des données.

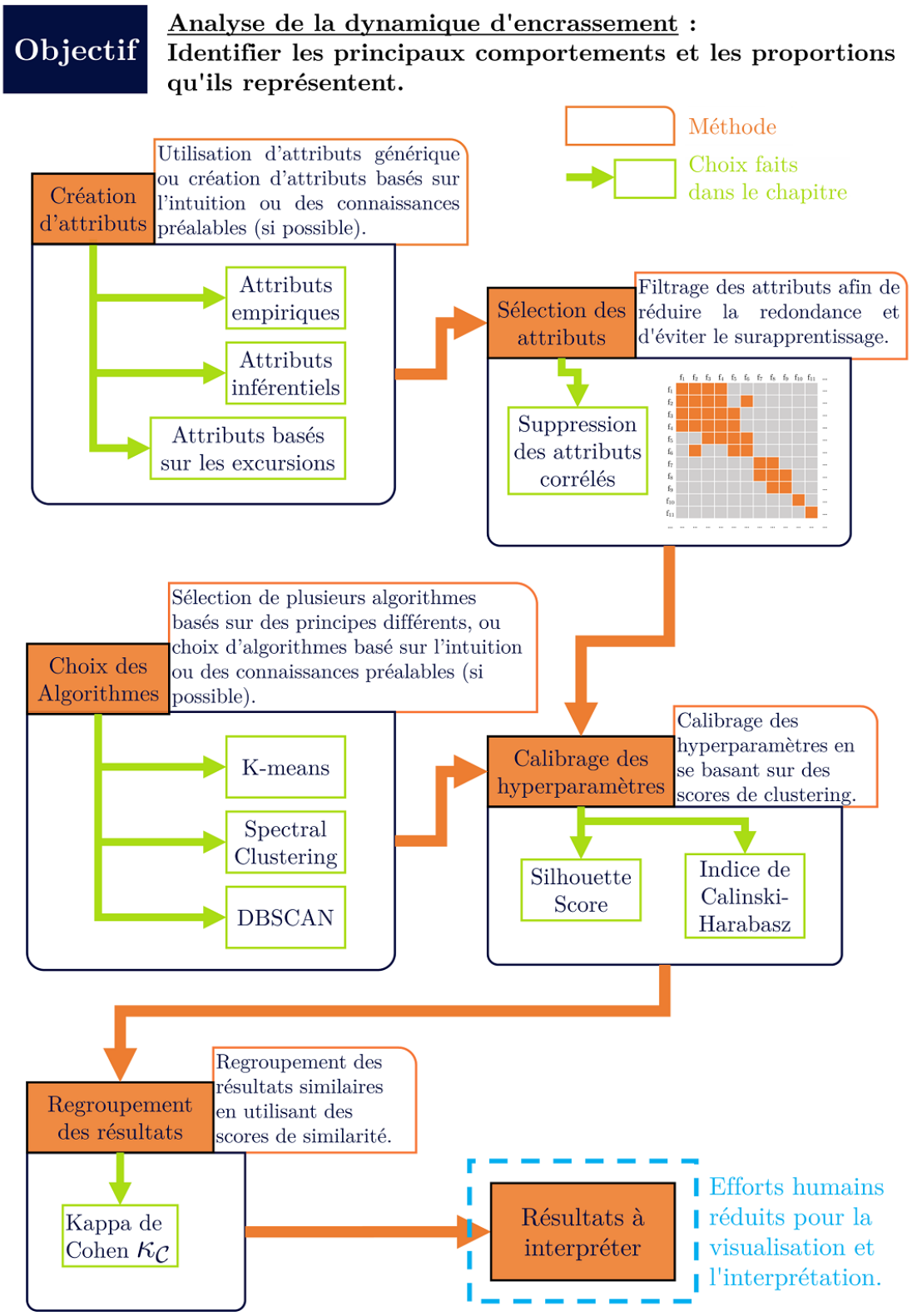

FIGURE 3.16 – Synthèse de la méthodologie de clustering utilisée.

L'étape suivante est la **sélection des attributs**, qui permet de ne conserver que les plus pertinents et ainsi minimiser la redondance. Dans notre cas, nous avons évalué la corrélation entre chaque couple d'attributs et les avons regroupés en familles en fonction de leur corrélation. Après avoir formé ces familles d'attributs, nous sélectionnons un ou deux attributs par famille, ceux qui représentent le mieux l'information. Finalement, à partir de ces attributs sélectionnés, nous formons différents sous-ensembles que nous utiliserons par la suite pour le clustering.

Le **choix des algorithmes** de clustering peut se fonder sur une compréhension préalable de la structure des données, ou sur le choix de divers algorithmes basés sur des principes variés. La seconde approche, que nous avons privilégiée dans notre étude, consiste à sélectionner différents algorithmes, chacun fondé sur des principes distincts. Cette méthode permet d'aborder le problème sous différents angles, augmentant ainsi les chances de découvrir des clusters pertinents. Dans notre étude, nous avons employé les trois algorithmes : K-means, Spectral Clustering et DBSCAN.

Pour chaque combinaison d'attributs et d'algorithmes, le **calibrage des hyperparamètres** est effectué en utilisant des scores de clustering. Nous avons utilisé le Silhouette Score et l'indice de Calinski-Harabasz, qui mesurent la qualité des clusters en matière de cohésion interne et de séparation entre les clusters. Ainsi, pour chaque ensemble d'attributs et d'algorithmes, nous sélectionnons les combinaisons d'hyperparamètres qui génèrent les scores les plus élevés.

Enfin, nous évaluons plusieurs sous-ensemble d'attributs, d'algorithmes de clustering et d'hyperparamètres. Pour optimiser ce processus et éviter l'analyse individuelle de chaque résultat de clustering, nous adoptons une stratégie de **regroupement des résultats similaires**. Cette méthode repose sur l'utilisation de scores de similarité pour comparer les couples de résultats. Les résultats jugés similaires sont ensuite regroupés, nous permettant de ne visualiser et d'interpréter qu'un seul résultat représentatif par groupe. Nous avons choisi d'utiliser le score Kappa de Cohen κ_C pour évaluer la similarité entre les résultats de clustering.

Les résultats obtenus en appliquant cette méthodologie sont présentés et discutés par la suite.

3.11 Résultats et interprétations

Après avoir appliqué le processus de clustering décrit ci-dessus, les résultats restants sont représentés visuellement et interprétés. L'objectif est de comprendre

ce que les différents clusters révèlent en termes de dynamique d'encrassement du point de vue opérationnel. Un exemple de visualisation des données sous la forme d'une *matrix plot* est présentée en Figure 3.17. Comme le montre cet exemple, un matrix plot est un type de visualisation qui permet de représenter les différents attributs par couple dans un jeu de données. Les graphiques sur la diagonale représentent la densité de la distribution de chaque attribut individuellement. En dehors de la diagonale, les graphiques affichent des nuages de points qui permettent d'observer directement la répartition des valeurs des attributs au sein de chaque cluster, chacun étant représenté par une couleur différente.

La Figure 3.17 présente un résultat à 3 clusters, obtenu en utilisant 5 attributs en entrée et l'algorithme Spectral Clustering. Pour des raisons de clarté, nous ne présentons que les deux premières lignes de cette matrice de graphiques. On peut observer sur cette figure que l'attribut $\overline{A_5}$ permet de distinguer le cluster 3 (en vert) car, si l'on observe la deuxième ligne de cette matrice, le nuage de points vert est éloigné des autres points, tout comme le centre de ce nuage appelé *centroïde*, représenté par une croix verte, qui est éloigné des autres centroïdes (représentés par des croix de couleurs différentes). Autrement dit, sur la deuxième ligne de graphiques, quel que soit l'attribut en abscisse, on constate que le nuage vert représente des avaloirs ayant une valeur de $\overline{A_5}$ plus élevée que dans les autres clusters. Inversement, on peut constater que cet attribut ne permet pas de distinguer les clusters 2 (en bleu) et 1 (en orange) car les nuages bleus et oranges sont toujours confondus sur la deuxième ligne. On peut également le constater sur les distributions associées à $\overline{A_5}$ situées sur la deuxième colonne de la deuxième ligne que les densités bleu et orange se superpose. Un exemple d'interpréation qu'il est possible de tirer des résultats présentés en Figure 3.17 est que, puisque le cluster vert correspond à des avaloirs ayant une $\overline{A_5}$ forte mais une $\overline{F_{2.5}}$ faible, c'est-à-dire aux avaloirs se remplissant peu souvent (fréquence de remplissage faible) mais ayant de forte variations (amplitude de remplissage élevé). Autrement dit, il pourrait s'agir des avaloirs qui se remplissent peu souvent, mais de déchets volumineux (ou bien d'agglomérats de petits déchets).

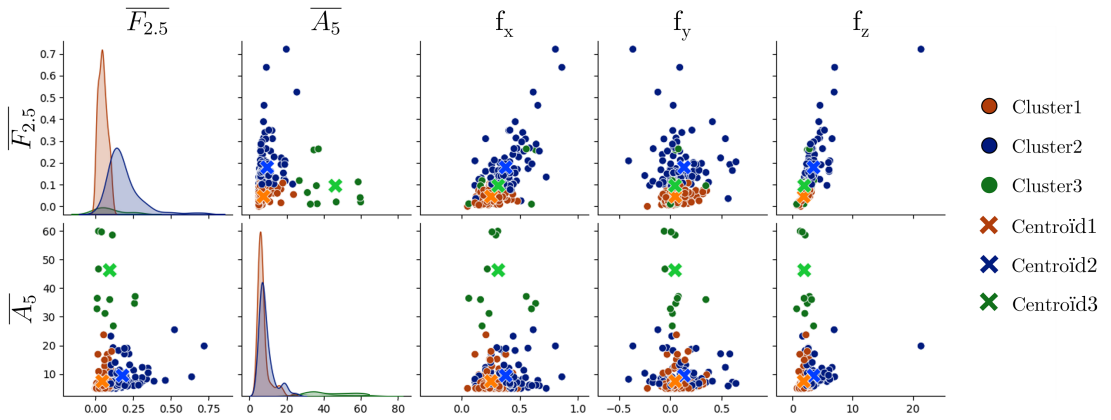

FIGURE 3.17 – Deux premières ligne d'une matrice de graphiques représentant la répartition des différents clusters en fonction des attributs utilisés. Dans cette figure, les avaloirs sont représentés par des points et les couleurs représentent les clusters associés. Les croix représentent les centres respectifs des différents nuages, en respectant le code couleur associé.

Parmi les résultats de clustering, obtenus sur un échantillon d'environ 2 000 avaloirs connectés ayant au moins un an d'historique de données, nous avons trouvé deux découpages particulièrement intéressants (ayant respectivement 3 et 4 clusters), comme décrit dans la Figure 3.18. Par la suite, nous appellons ces résultats $\mathcal{C}^3$ et $\mathcal{C}^4$, et nommons chacun des clusters de ces deux résultats, $\mathcal{C}_i^3$ et $\mathcal{C}_j^4$ respectivement, avec $i \in \{1, 2, 3\}$, et $j \in \{1, 2, 3, 4\}$.

Dans les deux découpages, une proportion significative ($> 50\%$) d'avaloirs à faible niveau d'activité est observée, correspondant aux classes $\mathcal{C}_1^3$ et $\mathcal{C}_1^4$. Bien que les deux clusters correspondent tous deux à une faible dynamique, on observe que leur délimitation n'est pas identique, l'un étant presque inclus dans l'autre : $\sim 96\%$ des avaloirs se trouvant dans $\mathcal{C}_1^3$ se trouvent également dans $\mathcal{C}_1^4$. Cela souligne la difficulté de définir ce qu'est un avaloir dynamique.

On observe également des clusters d'avaloirs présentant de grandes variations positives du niveau d'encrassement (7%) correspondant aux clusters $\mathcal{C}_3^3$ et $\mathcal{C}_4^4$. Ces clusters sont également assez similaires (plus de 80% des avaloirs d'un cluster se trouvent dans l'autre et inversement). En visualisant plus en détail les courbes associées à ces avaloirs, nous avons pu identifier une soixantaine de capteurs renvoyant des mesures anormales. On peut interpréter ces clusters $\mathcal{C}_3^3$ et $\mathcal{C}_4^4$ comme regroupant des avaloirs sensibles à des déchets volumineux (ou bien à des agglomérats de petits déchets).

Parmi les clusters restant, $\mathcal{C}_2^3$ contient les avaloirs se remplissant de manière progressive (fréquence élevée de petites variations positives), correspondant éventuellement aux avaloirs sensibles aux déchets et à d'autres débris de petit volume,

comme des mégots ou des feuilles d'arbres. Une attention particulière peut être accordée au cluster $\mathcal{C}_2^4$, qui peut être interprété comme les avaloirs avec *remplissages et pertes réguliers* en termes d'encrassement. Autrement dit, ce sont les avaloirs pour lesquels on mesure des diminutions du niveau d'encrassement. Cependant, certains attributs définissant ce cluster, tels que ceux caractérisant les excursions, peuvent dépendre des curages, il est important de vérifier si ces diminutions correspondent à des curages. En pratique, nous avons pu constater que parmi les $\sim$ 430 avaloirs de ce cluster, une majorité n'ont pas été curés pendant la période étudiée (environ 320 avaloirs) ou ont été curés une seule fois (environ 80 avaloirs). Ainsi, ce cluster regroupe des avaloirs risquant potentiellement de rejeter des déchets dans le réseau ou dans l'environnement. Finalement, contrairement à $\mathcal{C}_2^4$, le cluster $\mathcal{C}_3^4$ regroupe des avaloirs que l'on peut identifier comme ceux ayant tendance à accumuler l'encrassement sans le perdre. Ces avaloirs sont donc également intéressants du point de vue opérationnel, car ce sont potentiellement des avaloirs qui peuvent déborder fréquemment ou attirer d'autres nuisances (telles que des odeurs, des rats, etc.) si les déchets qui s'y trouvent stagnent.

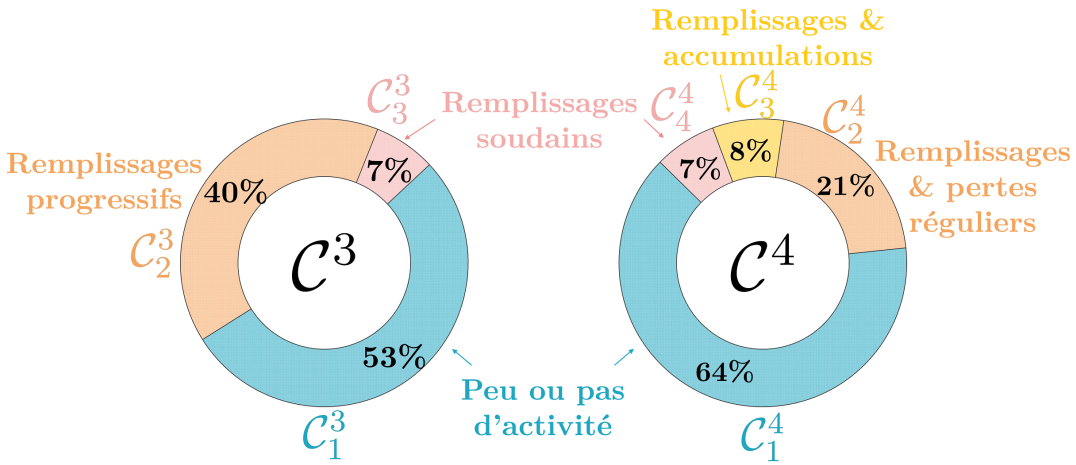

FIGURE 3.18 – Résultats de clustering $\mathcal{C}^3$ et $\mathcal{C}^4$ et interprétation de chaque cluster.

3.12 Synthèse du chapitre

En examinant les évolutions de l'encrassement des avaloirs équipés de capteurs, on observe différents comportements, tels qu'une accumulation lente de l'encrassement, des variations soudaines, etc. L'objectif de ce chapitre était d'examiner ces phénomènes afin d'identifier les comportements prédominants et les proportions qu'ils représentent. La méthodologie adoptée permet une exploration de différentes possibilités de choix d'algorithmes, d'attributs et d'hyperparamètres, tout

en facilitant autant que possible l'interprétabilité des résultats à chaque étape. La Figure 3.16 synthétise les différentes étapes de l'approche adoptée. Ces travaux ont permis d'obtenir des résultats stables et interprétables et d'identifier les principaux comportements des avaloirs et les proportions qu'ils représentent. Du point de vue opérationnel, on observe aujourd'hui qu'un certain nombre d'éléments contextuels influencent la dynamique d'encrassement. Nous étudions l'impact de l'environneent des avaloirs sur leur dynamique dans le dernier chapitre de cette thèse.

Chapitre 4

Corrélations contextuelles et impact sur la dynamique

Sommaire

Dans le chapitre précédent, nous avons dégagé plusieurs clusters/groupes de capteurs correspondants à diverses dynamiques d'encrassement. Ces groupes nous ont permis de mettre en évidence certains comportements prédominants et de quantifier leur importance. On constate aujourd'hui que certains éléments contextuels peuvent expliquer les types de déchets et leurs quantités retrouvés dans les avaloirs. Cette observation nous amène à nous interroger sur la nature de ces éléments contextuels influant et sur leur impact sur la dynamique. Cette problématique est traitée dans ce chapitre, où nous expliquons d'abord les types de contextes que nous avons identifiés avec l'aide d'experts opérationnels, avant de décrire les travaux de collecte et de préparation des données associées. Nous étudions ensuite comment ces données contextuelles sont associées avec les différentes catégories d'avaloirs identifiées au chapitre précédent. Cette analyse est effectuée d'abord de manière bivariée, en recherchant ces associations pour chaque facteur pris individuellement, puis de manière multivariée, où nous essayons de prédire le cluster associé à un avaloir en considérant tous les facteurs contextuels. Cette dernière étude est basée sur deux algorithmes : la régression logistique et le random forest.

4.1 Éléments contextuels identifiés

Cette section décrit les principaux types de contextes ayant potentiellement une influence sur la dynamique d'encrassement des avaloirs que nous nous avons identifés avec des experts opérationnels. Pour chaque type d'éléments, nous donnons quelques exemples. Enfin, notez que dans ce chapitre, nous appelons *éléments contextuels* les différents éléments identifiés. Une liste plus détaillée de ces éléments est donnée en Annexe D.

4.1.1 Contexte structurel

Dans un premier temps, nous considérons que la structure de l'avaloir influence la dynamique d'encrassement. Quelques exemples de ces éléments structurels sont décrits ci-dessous.

Équipements d'avaloirs : Les avaloirs peuvent être équipés de divers éléments supplémentaires susceptibles d'influencer leur dynamique d'encrassement. L'exemple le plus notable est le barreaudage, présenté en Section 1.4.2, qui a pour objectif de protéger l'avaloir en empêchant les déchets les plus volumineux d'y pénétrer, tels que les emballages de fast-food, les canettes, etc. Il convient de noter qu'il existe d'autres équipements ayant un impact similaire sur le blocage des déchets, comme les bavettes ou les clapets. Ces derniers sont des dispositifs anti-retour conçus

pour bloquer les mauvaises odeurs du réseau tout en assurant un bon écoulement hydraulique.

Taille et forme de l'avaloir : La taille et la forme de l'avaloir peuvent avoir un impact sur la manière dont les macro-déchets se positionnent à l'intérieur. Intuitivement, on peut penser que l'encrassement peut être plus étalé dans un avaloir large, engendrant donc des variations de niveau de déchet moins importantes. Inversement, on peut s'attendre à observer des variations de hauteur plus importantes dans une fosse étroite.

Taille de l'exutoire[1] **:** Intuitivement, la taille de l'exutoire devrait jouer un rôle important sur la dynamique d'encrassement, notamment pour les lessivages (pertes de déchets). Il est raisonnable d'imaginer qu'un avaloir doté d'une évacuation de grand diamètre laissera plus fréquemment les déchets passer à l'intérieur du réseau. Ainsi, l'avaloir lui-même est moins susceptible d'accumuler les déchets ou d'être obstrué et de déborder. À l'inverse, un avaloir avec une évacuation de petite taille aura moins tendance à évacuer les déchets vers le réseau mais présentera un risque plus élevé d'accumulation de déchets, d'obstruction, voire de débordement.

Inclinaison de la rue : La pente de la rue et le profil de la route où se trouve l'avaloir, ainsi que son emplacement au sein de cette rue, vont également avoir une influence. On observe que des avaloirs voisins peuvent avoir des dynamiques très différentes. Par exemple, un avaloir donné peut absorber tous les déchets d'une rue, "protégeant" ainsi l'avaloir voisin.

4.1.2 Contexte spatial

On observe que certains éléments contextuels proches de l'avaloir peuvent déterminer le type de déchet trouvé à l'intérieur et donc influencer sa dynamique d'encrassement.

Bacs à ordure : À Marseille, on observe que la présence de bacs à ordures peut aggraver l'encrassement des avaloirs à proximité. Ceci peut s'expliquer par le fait que ces bacs peuvent déborder ou que les ordures peuvent être déposées à côté. En conséquence, la quantité de déchets présents dans la rue (et finalement dans les avaloirs) est généralement plus importante.

1. On rappelle que l'exutoire correspond au tuyau d'évacuation de l'avaloir, cf. Figure 1.6

Fast-foods : Parmi les avaloirs à proximité des fast-foods, on peut parfois retrouver des déchets spécifiques à la restauration rapide, tels que des canettes et des emballages de nourriture.

Arbres : Des éléments naturels peuvent aussi encombrer les avaloirs, par exemple les arbres qui perdent leurs feuilles en automne.

4.1.3 Contextes temporel et spatio-temporel

Enfin, des événements peuvent également influencer la dynamique d'encrassement.

Pluie : Les précipitations entraînent les déchets des surfaces urbaines vers les avaloirs, augmentant ainsi la quantité de débris qui s'y accumulent. Dans certains cas, la pluie peut aussi provoquer un phénomène de siphonnage et entraîner l'évacuation des déchets dans le réseau. Enfin, si l'avaloir est obstrué, l'eau de pluie peut s'accumuler et causer un débordement. Dans cette situation, les déchets peuvent être rejetés en surface.

Vent : Le vent peut déplacer ou éjecter certains déchets légers, tels que les emballages, les sacs plastiques ou encore des éléments naturels comme les feuilles d'arbres.

Évènements publics : Certains événements peuvent entraîner une forte concentration de population dans des espaces restreints, ce qui peut augmenter la quantité de déchets générés. On peut citer ici les événements festifs, tels que les festivals ou les concerts, ou encore les rencontres sportives. Ces événements peuvent être ponctuels ou plus réguliers, comme les marchés ou les vide-greniers. À Marseille, on peut parfois constater d'importantes quantités de déchets après les marchés.

Activités de nettoyage : Enfin, d'autres activités doivent être prises en compte ; par exemple les services en charge de la collecte des déchets ménagers ou les services de nettoyage de voirie. Le fonctionnement de ces services (fréquence de ramassage des ordures, méthode de nettoyage des rues, grèves, etc.) va également influencer la dynamique d'encrassement des avaloirs.

4.2 Collecte et préparation des données

4.2.1 Référencement exploitation

Une partie des informations contextuelles est déjà à notre disposition, en particulier, celles relatives à l'infrastructure. Hormis quelques avaloirs pour lesquels les données sont manquantes, on connaît la forme, les dimensions et la présence éventuelle de barreaudages pour les avaloirs. Cependant, certaines données relatives à la structure restent manquantes, comme l'inclinaison de l'exutoire ou la pente et le profil de la rue.

4.2.2 Données cartographiques

La localisation d'un certain nombre d'éléments contextuels identifiés a été récupérée via OpenStreetMap [61], tels que la localisation des bacs à ordures, de bac de tri, des arrêts de transport en commun, des fast-foods, des bars/cafés/pubs, etc.

4.2.3 Préparation des données et découpage spatial

Étudier l'impact de certains éléments contextuels nécessite de définir un rayon d'influence pour chacun de ces facteurs. Par exemple, nous avons besoin de définir la distance à partir de laquelle on considère qu'un arbre influence la dynamique d'encrassement d'un avaloir. En pratique, ne connaissant pas ce rayon d'action, nous avons choisi de construire différents *facteurs contextuels* pour différents rayons d'action. Ces derniers ont été choisis de manière à couvrir différents ordres de grandeur en se basant sur l'expérience opérationnelle. Nous appellerons donc *facteurs contextuels* les caractéristiques construites afin d'évaluer l'impact des *éléments contextuels* sur la dynamique d'encrassement des avaloirs. Ainsi, un élément contextuel peut être associé à plusieurs facteurs. Par exemple, nous évaluerons l'impact de la présence d'arbres (élément contextuel) en définissant les facteurs "présence d'arbres dans un rayon de 3 mètres" et "présence d'arbres dans un rayon de 10 mètres". Dans la suite, nous indiquons ce rayon d'influence (noté $\mathcal{R}$) pour chaque facteur.

Nous avons également souhaité étudier les impacts de certains éléments, moyennés sur une zone d'étude plus étendue que l'entourage immédiat de l'avaloir, afin de mettre en évidence des caractéristiques telles que l'affluence de la population. Par exemple, en prenant le quartier comme référence spatiale et en calculant le nombre et la densité de bars [2] par quartier. En plus des quartiers comme unité de

2. En pratique, nous avons considéré la présence de bars, de cafés et de pubs comme un élément unique bar/café/pub.

découpage spatial, nous avons aussi adopté le découpage en *Îlots Regroupés pour l'Information Statistique* (IRIS) proposé par l'*Institut National de la Statistique et des Études Économiques* (INSEE) [62] décrit en Annexe E. De manière similaire, nous avons construit des facteurs contextuels liés à la quantité ou à la densité de certains éléments, tels que le nombre de restaurants par IRIS, par exemple.

4.2.4 Synthèse des données utilisées

Finalement, nous avons collecté et préparé les données contextuelles sous forme de facteurs dont nous souhaitons étudier l'impact. L'ensemble de ces données peut être synthétisé sous forme d'un tableau regroupant pour chaque avaloir les facteurs contextuels préparés, comme illustré en Tableau 4.1. Les facteurs peuvent être qualitatifs, tels que la présence d'arbres dans un certain rayon (*Vrai* ou *Faux*), ou bien quantitatifs, comme le nombre de fast-foods dans le quartier associé ou la densité de pubs dans la zone IRIS. La liste des facteurs contextuels étudiés et leurs descriptions sont données en Annexe F.

ID Capteur	**Présence Arbre 10m**	**Présence Arbre 30m**	**ID Quartier**	**ID IRIS**	**Nb Fast-Food Quartier**	**...**
1	Vrai	Vrai	Hôtel de Ville	132020303	5	...
2	Faux	Faux	Hôtel de Ville	132020303	5	...
3	Vrai	Faux	La Joliette	132020401	13	...

TABLE 4.1 – Illustration des données contextuelles après mise en forme.

Dans la suite de notre étude, nous nous sommes concentrés sur l'impact des contextes structurel et spatial.

4.3 Description de l'étude

Nous rappelons que les travaux de clustering présentés au chapitre précédent sont basés sur un ensemble d'environ 2000 avaloirs. Dans cet ensemble, certains avaloirs ont un référencement incomplet qui ne permet pas de déterminer leur contexte (en raison de l'absence de coordonnées GPS, par exemple). De plus, une soixantaine d'anomalies supplémentaires ont été identifiées suite à ces travaux de clustering. En excluant les avaloirs dont le contexte est inconnu ou dont les

données ont été identifiées comme anormales, il reste un lot de 1935 avaloirs que nous utilisons pour l'étude qui suit.

À partir des résultats de clustering du chapitre précédent, plus précisément le découpage en 3 clusters $\mathcal{C}^3$ et le découpage en 4 clusters $\mathcal{C}^4$, nous proposons un nouveau découpage $\mathcal{C}^{\text{Dyn}}$ visant à regrouper les avaloirs en fonction de la présence ou de l'absence de dynamique. Plus précisément, un avaloir est considéré comme non dynamique s'il a été catégorisé comme tel par au moins un des deux résultats $\mathcal{C}^3$ ou $\mathcal{C}^4$ (cf. Figure 4.1). Par conséquent, les clusters sont définis comme suit.

- $\mathcal{C}_1^{\text{Dyn}} = \mathcal{C}_1^3 \cup \mathcal{C}_1^4$, le cluster des avaloirs non dynamiques
- $\mathcal{C}_2^{\text{Dyn}} = \overline{\mathcal{C}_1^{\text{Dyn}}} = \overline{\mathcal{C}_1^3} \cap \overline{\mathcal{C}_1^4}$, le cluster des avaloirs dynamiques.

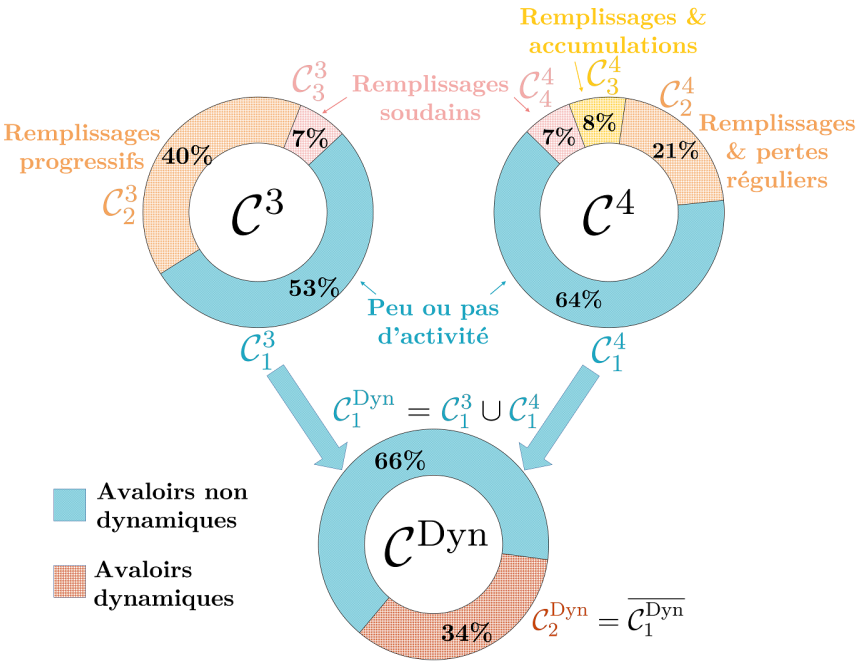

FIGURE 4.1 – Définition du découpage $\mathcal{C}^{\text{Dyn}}$ à partir des découpages $\mathcal{C}^3$ et $\mathcal{C}^4$.

Cette étude se concentre sur l'analyse et la prédiction des trois découpages $\mathcal{C}^3$ et $\mathcal{C}^4$ et $\mathcal{C}^{\text{Dyn}}$ à partir des facteurs contextuels construits.

4.4 Analyse bivariée

Nous avons commencé par analyser la répartition des différents éléments contextuels au sein des clusters d'avaloirs des trois découpages $\mathcal{C}^3$, $\mathcal{C}^4$ et $\mathcal{C}^{\text{Dyn}}$ afin d'identifier les premiers éléments d'influence.

4.4.1 Facteurs qualitatifs

Nous commençons nos analyses en évaluant l'impact des facteurs contextuels qualitatifs sur nos données. Pour ce faire, nous avons établi des matrices de contingence et effectué des tests du χ^2 afin de déterminer l'influence de chaque facteur contextuel sur les clusters. Il est important de souligner que nous avons pris en compte une cinquantaine de facteurs de ce type.

Le test de χ^2 est utilisé pour vérifier l'indépendance entre deux variables qualitatives. L'hypothèse nulle $\mathcal{H}_0$ postule que les deux variables sont indépendantes. La *p-value* obtenue à partir de ce test offre une mesure de l'accord entre les données observées et ce à quoi on s'attendrait si $\mathcal{H}_0$ était vraie. Si la p-value est inférieure à un seuil prédéfini (généralement fixé à 5%), alors on rejette l'hypothèse nulle. Comme mentionné dans le chapitre précédent, une correction est nécessaire lorsqu'on réalise plusieurs tests afin de minimiser le risque d'erreurs de type I notamment. Comme précédemment nous utiliserons la correction de Benjamini-Hochberg car elle offre une plus grande puissance pour détecter les vrais effets (cf Section 3.4.2).

Le Tableau 4.2 donne les facteurs significatifs après correction des tests avec la méthode de Benjamini-Hochberg pour le découpage $\mathcal{C}^3$. Dans ce tableau et par la suite, nous désignons par *shop* tout type de magasin ou boutique (coiffeur, épicerie, bijoutier, boulangerie, magasin de vêtements, etc.).

Facteur contextuel étudié	**Description**	**p-value corrigée**
Arbre-10m	Variable binaire correspondant à la présence d'au moins un arbre dans un rayon de $\mathcal{R} = 10$ m.	1×10^{-3}
Bar-Café-10m	Variable binaire correspondant à la présence d'au moins un bar ou d'un café ou d'un pub dans un rayon de $\mathcal{R} = 10$ m.	5×10^{-2}
Bar-Café-50m	Variable binaire correspondant à la présence d'au moins un bar ou d'un café ou d'un pub dans un rayon de $\mathcal{R} = 50$ m.	3×10^{-2}
Is-Grid	Variable binaire vérifiant si l'avaloir possède au moins une grille.	1×10^{-4}
Is-Shop-Quartier	Variable binaire qui évalue si l'avaloir se trouve dans un quartier faisant partie des 25% des quartiers comptant le plus grand nombre de *shops* parmi les 101 quartiers en tout. Plus précisément, cela est équivalent à examiner les quartiers ayant 33 shops ou plus.	3×10^{-2}

TABLE 4.2 – liste des facteurs contextuels significatif au seuil de 5% après correction des tests du χ^2 appliqué sur le résultat de clustering $\mathcal{C}^3$.

Ces tests statistiques montrent que les présences d'arbres, de bars/cafés/pubs, de grilles, et de shops, sont corrélées à la dynamique d'encrassement. Par exemple, le Tableau 4.3 présente le tableau de contingence entre la présence d'un arbre dans un rayon de 10m et la répartition des avaloirs dans les différents clusters de $\mathcal{C}^3$.

On peut y voir que la présence d'arbre semble plus importante parmi les clusters $\mathcal{C}_2^3$ et $\mathcal{C}_3^3$. Autrement dit, la présence d'arbre semble liée à la présence ou à l'absence de dynamique. Plus précisément, parmi les facteurs significatifs listé en Tableau 4.2, tous les facteurs, excepté de *Is-Shop-Quartier*, semblent liés à la présence ou à l'absence de dynamique et sont étudiés plus en détail par la suite. Le facteur *Is-Shop-Quartier* semble plus particulièrement lié au cluster $\mathcal{C}_2^3$ comme le montre le Tableau 4.4.

	Absence d'arbre pour $\mathcal{R} = 10$ m	**Présence d'arbre pour $\mathcal{R} = 10$ m**	Total
$\mathcal{C}_1^3$ Absence/peu de dynamique	986 (93%)	77 (7%)	1063 (100%)
$\mathcal{C}_2^3$ Remplissages progressifs	683 (87%)	105 (13%)	788 (100%)
$\mathcal{C}_3^3$ Remplissages soudains	74 (88%)	10 (12%)	84 (100%)
Total	1743 (90%)	192 (10%)	1935 (100%)

TABLE 4.3 – Tableau de contingence entre la présence d'arbre dans un rayon de 10m et $\mathcal{C}^3$. Les proportions par ligne sont représentées en vert.

	Is-Shop-Quartier : Non	**Is-Shop-Quartier : Oui**	Total
$\mathcal{C}_1^3$ Absence/peu de dynamique	762 (72%)	301 (28%)	1063 (100%)
$\mathcal{C}_2^3$ Remplissages progressifs	508 (64%)	280 (36%)	788 (100%)
$\mathcal{C}_3^3$ Remplissages soudains	61 (73%)	23 (27%)	84 (100%)
Total	1331 (69%)	604 (31%)	1935 (100%)

TABLE 4.4 – Tableau de contingence entre *Is-Shop-Quartier* et $\mathcal{C}^3$. Les proportions par ligne sont représentées en vert.

Afin de quantifier l'influence de *Is-Shop-Quartier* sur les chances qu'un avaloir se remplisse progressivement (i.e., appartenir au cluster $\mathcal{C}_2^3$), nous avons calculé le *risque relatif* (RR) associé. Le RR est une mesure quantitative permettant de comparer la probabilité qu'un événement se produise dans une population exposée à un facteur par rapport à une population non exposée. Il est défini comme le rapport des risques entre ces deux groupes :

$$RR = \frac{\text{Probabilité de l'événement chez les exposés}}{\text{Probabilité de l'événement chez les non exposés}} \tag{4.4.1}$$

Si $RR = 1$, il n'y a pas de différence de risque entre les deux groupes. Si $RR > 1$, l'exposition augmente le risque de l'événement, et si $RR < 1$, l'exposition réduit le risque.

À partir des données du Tableau 4.4, on peut calculer le RR qu'un avaloir se trouve dans le cluster $\mathcal{C}_2^3$ en fonction de *Is-Shop-Quartier*, qui est égal à $\frac{280}{604}/\frac{508}{1331} \simeq 1.2$.

On peut interpréter que la présence des commerces augmente les chances qu'un avaloir ait un remplissage progressif. Du point de vue opérationnel, on peut supposer que ce comportement est dû, par exemple, à la présence plus importante de déchets tels que des mégots dans des endroits plus fréquentés.

En procédant similairement, on établit le Tableau 4.5 synthétise les facteurs significatifs après correction des tests pour le découpage $\mathcal{C}^4$.

Facteur contextuel étudié	**Description**	**p-value corrigée**
Arbre-3m et *Arbre-10m*	Variable binaire correspondant à la présence d'au moins un arbre pour un rayon de $\mathcal{R} = 3$ m et $\mathcal{R} = 10$ m respectivement.	1×10^{-3} et 3×10^{-2}
Fast-Food-10m et *Fast-Food-50m*	Variable binaire correspondant à la présence d'au moins un fast-food pour un rayon de $\mathcal{R} = 10$ m et $\mathcal{R} = 50$ m respectivement.	3×10^{-2} dans les deux cas
Is-Fast-Food-IRIS et *Is-Fast-Food-Quartier*	Variable binaire qui évalue si l'avaloir fait partie des 25% d'IRIS (respectivement de quartiers) comptant le plus grand nombre de fast-foods. Plus précisément, cela correspond à examiner les IRIS (respectivement quartiers) ayant 2 (respectivement 3) fast-food ou plus.	5×10^{-4} et 3×10^{-2}
Bar-Cafe-10m et *Bar-Cafe-50m*	Variable binaire correspondant à la présence d'au moins un bar / café / ou pub pour un rayon de $\mathcal{R} = 10$ et $\mathcal{R} = 50$ m respectivement.	1×10^{-2} et 4×10^{-2}
Is-Grid	Variable binaire vérifiant si l'avaloir possède au moins une grille.	2×10^{-4}
Is-Shop-IRIS et *Is-Shop-Quartier*	Variable binaire qui évalue si l'avaloir fait partie des 25% d'IRIS (respectivement de quartiers) comptant le plus grand nombre de *shop*. Plus précisément, cela correspond à examiner les IRIS (respectivement quartiers) ayant 12 (respectivement 33) shops ou plus.	5×10^{-4} et 2×10^{-2}
Is-Density-Shop-IRIS	Variable binaire qui évalue si l'avaloir fait partie des 25% respectivement d'IRIS ayant la plus grande densité de *shop*. Plus précisément, cela correspond à examiner les IRIS ayant ~ 48 shops par km^2 ou plus.	2×10^{-2}

TABLE 4.5 – Synthèse des facteurs contextuels significatifs au seuil de 5% après correction des tests du χ^2 appliqué sur le résultat de clustering $\mathcal{C}^4$.

Comme précédemment, en regardant le tableau de contingence associé à chaque

facteur, on constate que tous les facteurs semblent liés à la présence ou à l'absence de dynamique (ce que nous étudions plus bas), à l'exception de *Fast-Food-50m*, *Is-Fast-Food*-IRIS, *Is-Fast-Food-Quartier*, et *Is-Shop*-IRIS qui sont liés à certains clusters spécifiques.

La présence de fast-foods dans un rayon de 50 mètres semble liée aux clusters $\mathcal{C}_3^4$ "remplissage et accumulation" et $\mathcal{C}_4^4$ "grand remplissage". Les risques relatifs associés sont de 1.8 et 1.6, respectivement. Pour expliquer cela, on peut supposer que les déchets pouvant être générés par les fast-foods sont de plus grandes dimensions (emballages cartonnés, canettes, etc.) et causent donc des variations de mesures plus importantes en rentrant dans l'avaloir et sont moins facilement évacués dans le réseau. Cependant, la perte de ces déchets est également liée à d'autres facteurs comme la dimension de l'exutoire.

Les facteurs restants semblent être plutôt liés à $\mathcal{C}_3^4$ "remplissage et accumulation". Le Tableau 4.6 liste ces facteurs et le RR associé à chacun. De manière similaire, on peut interpréter que ces facteurs sont liés à l'émission de déchets étant plus susceptibles d'être retenus dans l'avaloir.

Facteur contextuel	**Risque relatif** ($\in \mathcal{C}_3^4$)
Is-Fast-Food-IRIS	2.1
Is-Fast-Food-Quartier	1.7
Is-Shop-IRIS	2.0

TABLE 4.6 – Synthèse des facteurs particulièrement liés à $\mathcal{C}_3^4$ et le RR associé à chacun.

Afin de quantifier plus généralement l'impact du contexte, nous étudions plus spécifiquement la corrélation entre ces facteurs et la présence de dynamique d'encrassement décrite par $\mathcal{C}^{\text{Dyn}}$. Nous ne détaillons pas chacun des facteurs ressortant de ces tests, car leurs comportements sont très similaires à ceux déjà présentés. Il est cependant intéressant de calculer les RR des différents facteurs sur la présence de dynamique (être dans le cluster $\mathcal{C}_1^{\text{Dyn}}$). Le Tableau 4.7 synthétise ces résultats.

Facteur contextuel	**Risque relatif** ($\in \mathcal{C}_2^{Dyn}$)
Arbre-3m	1.4
Arbre-10m	2.0
Bar-Cafe-10m	1.9
Bar-Cafe-50m	1.4
Fast-food-50m	1.4
Is-Grid	0.7
Is-Shop-Quartier	1.2

TABLE 4.7 – Synthèse de l'évaluation de l'impact des facteurs contextuels significatifs (selon les tests du χ^2 réalisés).

Le dernier facteur qualitatif mis en avant par les tests que nous n'avons pas encore discuté est la présence de grilles dont le RR est inférieur à 1, indiquant que ce facteur semble lié à une dynamique faible ou absente. Sachant que les grilles étaient historiquement installées dans les zones nécessitant une absorption accrue des eaux de pluie, une interprétation possible est que la plus grande quantité d'eau absorbée pourrait favoriser l'évacuation de l'encrassement dans le réseau. De plus, les grilles pourraient permettre à la pluie de s'écouler tout en retenant les déchets qu'elle entraîne. Sans grilles, la pluie s'infiltre uniquement par l'ouverture au niveau du trottoir, emportant avec elle les déchets vers l'avaloir. Cet effet bloquant des grilles sur les macro-déchets pourrait également expliquer la faible dynamique d'encrassement de ces avaloirs.

4.4.2 Facteurs quantitatifs

Similairement, nous avons exploré les corrélations entre les facteurs contextuels quantitatifs et la répartition des avaloirs selon les trois découpages étudiés $\mathcal{C}^3$, $\mathcal{C}^4$ et $\mathcal{C}^{Dyn}$. Pourcela, le test du χ^2 utilisé précédemment est remplacé par une ANOVA (cf. Section 3.3) tout en conservant une démarche globalement similaire. Comme précédemment, nous utiliserons la correction de Benjamini-Hochberg pour les p-values obtenues avec un seuil de 5%.

Il est à noter que certaines données structurelles peuvent manquer. Par exemple, sur un total de 1935 avaloirs retenus dans cette étude, seulement 1835 disposent d'une information sur la largeur d'avaloir et 1900 sur sa profondeur dû à des problèmes de référencement, indépendants des clusters étudiés. Nous effectuons donc les tests ANOVA uniquement sur les avaloirs pour lequels ces données sont disponibles.

Les résultats montrent que, en plus des facteurs déjà évoqués (arbres, shops, etc.), les caractéristiques structurelles des avaloirs se distinguent nettement. Le Tableau 4.8 synthétise ces résultats.

Facteur contextuel étudié	**p-value corrigée pour $\mathcal{C}^3$**	**p-value corrigée pour $\mathcal{C}^4$**	**p-value corrigée pour $\mathcal{C}^{Dyn}$**
Diamètre de l'exutoire	1×10^{-10}	2×10^{-21}	2×10^{-8}
Largeur de l'avaloir	7×10^{-4}	7×10^{-4}	8×10^{-5}
Profondeur de l'avaloir	1×10^{-35}	5×10^{-63}	2×10^{-33}

TABLE 4.8 – Liste des facteurs contextuels quantitatifs, liés à la structure de l'avaloir, et dont l'influence sur la dynamique d'encrassement est statistiquement significative.

Les facteurs significatifs sont le diamètre de l'exutoire, la largeur de l'avaloir, et la profondeur de l'avaloir. En examinant la répartition du diamètre de l'exutoire dans les différents clusters d'avaloirs on constate que les avaloirs du cluster $\mathcal{C}_3^4$ "remplissage et accumulation" ont un diamètre d'exutoire plus petit comparé aux autres clusters. Plus précisément, le diamètre médian des exutoires des avaloirs appartenant au cluster $\mathcal{C}_3^4$ est de 30 cm contre 40 cm pour le reste des clusters. Ce résultat correspond à l'intuition que si l'exutoire est de plus petite taille alors les déchets présents dans l'avaloir auront plus de difficulté à pénétrer dans le réseau, ce qui tend à favoriser l'accumulation de l'encrassement.

Finalement, la largeur et la profondeur de la fosse apparaissent également comme des éléments significatifs. En regardant les distributions intra-classes, on constate que les avaloirs ayant les fosses les plus larges figurent notamment parmi les avaloirs peu dynamiques. On peut supposer que les déchets entrant dans l'avaloir sont plus étalés, diminuant ainsi les variations de hauteur d'encrassement mesurées par le capteur. Enfin, l'influence de la profondeur de l'avaloir peut sembler étonnante car on ne s'attend pas à ce que la quantité d'éléments (déchets, feuilles d'arbres, etc.) entrant dans l'avaloir dépend de sa profondeur. On peut donc imaginer qu'elle influence la manière dont les déchets tombent dans l'avaloir (par exemple, les déchets allongés comme les bouteilles pourraient tomber plutôt à la verticale ou à l'horizontale), ou bien que la mesure elle-même (variabilité/fiabilité) dépend de cette profondeur.

Afin d'explorer cette deuxième idée, nous avons regardé plus spécifiquement l'influence de la profondeur sur les capteurs anormaux que nous avions jusqu'à présent écarté de l'étude. Nous avons écarté environ 180 capteurs au fur et à

mesure de notre étude en raison de la fiabilité insatisfaisante de la mesure (liée par exemple à un problème d'installation). En reprenant notre jeu de donnée initial et en comparant les distributions de la profondeur en fonction de si le capteur a été identifié comme anormal, on constate que la profondeur semble fortement corrélée à ces problèmes(cf. Figure 4.2). La p-value obtenue par ANOVA dans ce cas nous donne une valeur de l'ordre de 2×10^{-47}, confirmant cette correlation.

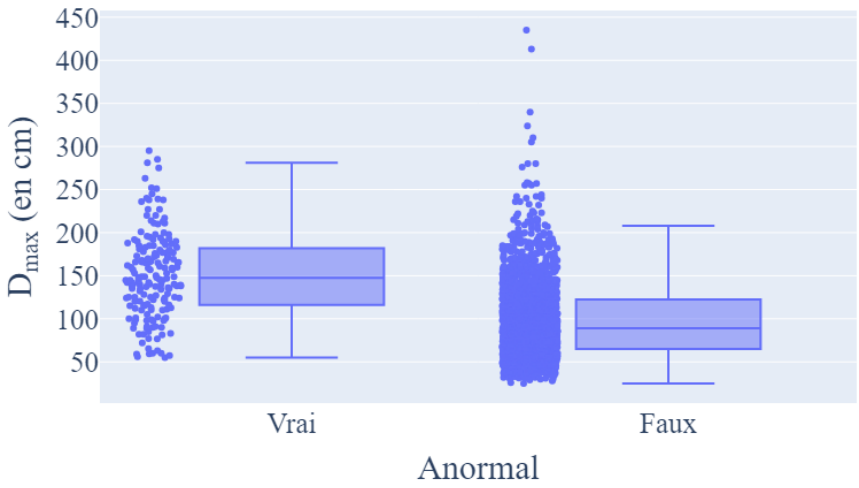

FIGURE 4.2 – Comparaison entre la distribution de la profondeur en fonction de si le capteur a été identifié comme anormal.

Une explication potentielle de ce problème peut venir du fonctionnement du dispositif en lui-même. Le capteur calcule le niveau d'encrassement de l'avaloir en mesurant le temps entre l'émission d'une impulsion ultrasonore et la réception de son écho. Selon le modèle (très) aproximatif du constructeur, cette impulsion se propage de manière conique. Ainsi, la taille de la surface mesurée par le capteur (sur laquelle se fait la réflexion de l'onde) depend de la distance mesurée, comme illustré en Figure 4.3. Le constructeur nous fournit la largeur de ce cône que l'on notera l_c en fonction de la distance mesurée, disponibles en Figure 4.4.

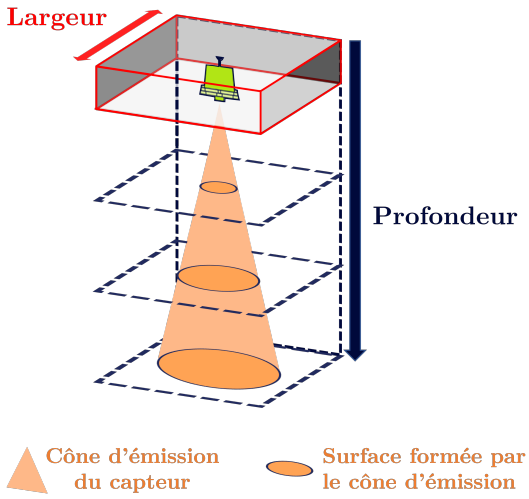

FIGURE 4.3 – Modèle (très) aproximatif du constructeur : le capteur émet une impulsion ultrasonore avec un profil de faisseau d'une forme cônique.

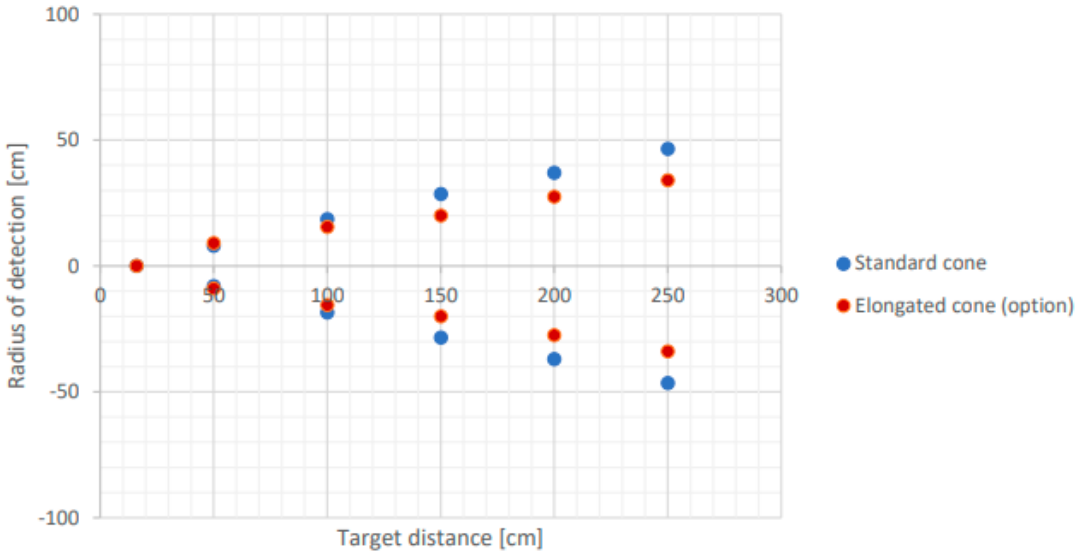

FIGURE 4.4 – Evolution du rayon de mesure du capteur selon le fabriquant. Les capteurs utilisés dans notre étude ont un profil "Standard cone" [63].

À partir des données constructeur disponibles en Figure 4.4 (sachant que la technologie utilisée dans notre réseau est nommée *Standard cone*), nous avons approximé l'évolution du diamètre de ce cône en fonction de la distance mesurée

avec une regression linéaire [3]. La droite affine obtenue correspond à $l_c \simeq 0.4 \cdot \mathrm{D} - 4.2$.

En Figure 4.5, nous comparons la largeur des avaloirs à la largeur du cône dans le cas où l'avaloir est vide $l_c(D_{\max})$. On observe qu'un certain nombre d'avaloirs ont une largeur inférieure à celle de ce cône (points bleus en dessous de la droite rouge). Les étoiles oranges représentent les capteurs identifiés comme anormaux. Pour comparer plus précisément les deux critères "largeur insuffisante" (correspondant aux avaloirs dont la largeure est inférieur au cône de mesure lorsque l'avaloir est vide) et "identifié comme anormal", nous avons établi la matrice de contingence associée (cf. le Tableau 4.9). La p-value résultante associée à un test de χ^2 appliqué à ce tableau est de $\sim 10^{-29}$, affirmant donc le rejet de l'hypothèse $\mathcal{H}_0$, i.e. les deux critères ne sont pas indépendantes. Notez que dans ce cas, seuls les capteurs que nous avons visuellement invalidés sont considérés comme anormaux ; le reste des capteurs étant considérés par défaut comme normaux [4]. Ces résultats montrent une perspective intéressante à explorer afin d'améliorer la fiabilité du réseau d'avaloirs connectés. Cela necessiterait d'étudier plus en détails l'évolution de la mesure du capteur lorsque l'onde émise se réfléchie partiellement sur les parois ou sur une surface de réflexion quelconque. De plus, d'autres paramètres comme la forme de l'avaloir (notamment celle du fond), la position exacte du capteur par rapport à l'avaloir (on peut supposer que l'installation du capteur n'est généralement pas parfaitement centrée), etc., devraient être pris en compte afin de conclure sur les conditions de fonctionnement optimales du dispositif (cf. Section 1.5). Ce sujet ne faisant pas partie des objectifs de cette thèse, nous ne le discutons pas davantage.

3. Les données utilisées sont $X = \{0, 50, 100, 150, 200, 250\}$ et $Y \simeq \{0, 16, 38, 56, 74, 94\}$.
4. Ceux-ci inclu donc potentiellement des anomalies encore non détectées.

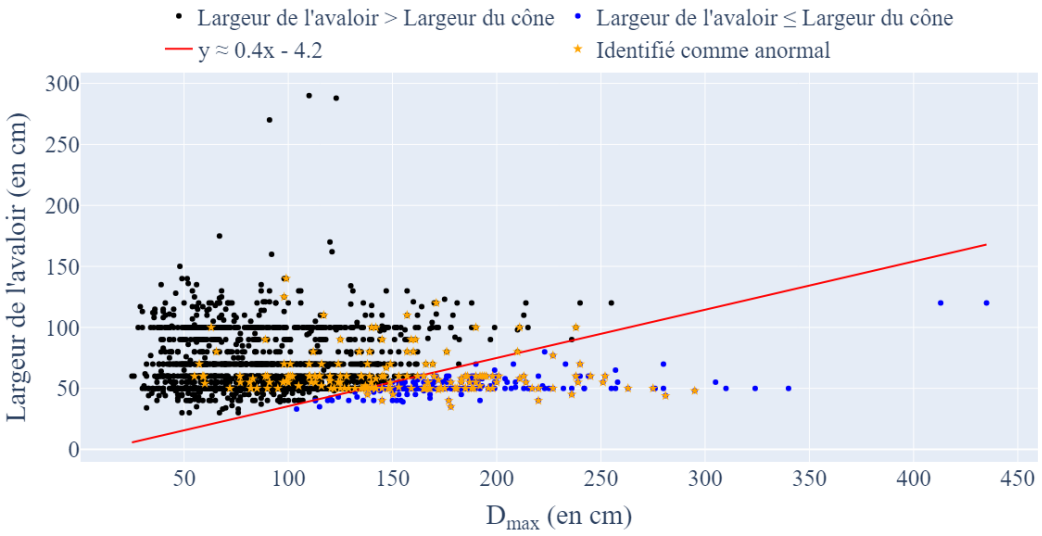

FIGURE 4.5 – Comparaison entre la largeur des avaloir étudiés et celle du cône de mesure dans le cas où l'avaloir est vide (trait rouge) en fonction de la profondeur de l'avaloir $D_{\max}$. Un point noir (respectivement bleu) signifie que la largeur de l'avaloir est supérieure (inférieure) à celle du cône de mesure. Les étoiles oranges représentent les capteurs identifiés comme anormaux.

		Largeur insuffisante		Total
		Non	Oui	
Identifié comme anormal	Non	1629 (86%)	271 (14%)	1900 (100%)
	Oui	94 (52%)	86 (48%)	180 (100%)

TABLE 4.9 – Tableau de contingence comparant les deux critères "largeur insuffisante" (autrement dit, la largeur de l'avaloir est inférieur à celle du cône) et "identifié comme anormal".

Lors de notre première analyse, nous avons identifié individuellement les facteurs contextuels qui semblent être corrélés aux résultats de clustering du chapitre précédent. Cependant, il est possible que certains facteurs aient peu d'influence individuellement, mais que leur combinaison ait un impact sur la dynamique d'encrassement. Ainsi, la section suivante traite d'une approche multivariée en tentant de prédire les résultats $\mathcal{C}^3$, $\mathcal{C}^4$ et $\mathcal{C}^{Dyn}$ pour identifier de nouveaux facteurs d'influence et les combinaisons possiblement associées. Enfin, du point de vue opérationnel, il serait bénéfique de déterminer si nous sommes en mesure de prédire la catégorie d'un avaloir en nous basant sur ces facteurs contextuels préalablement définis. Cela indiquerait que nous pouvons anticiper le comportement typique d'un

avaloir selon son environnement, ce qui serait intéressant pour la maintenance des avaloirs non équipés de capteurs.

4.5 Analyse multivariée

4.5.1 Données en entrée

Comme mentionné précédemment, il se peut que certaines données contextuelles soient incomplètes, en particulier celles concernant la structure de l'avaloir. Étant donné que les modèles que nous utilisons requièrent des données complètes et sans lacunes, nous avons décidé d'utiliser une base de données de 1794 avaloirs après avoir éliminé ceux avec des informations manquantes.

Il est à noter qu'il existe plusieurs méthodes pour gérer les valeurs manquantes dans les données. Parmi celles-ci, l'imputation multiple, qui remplace les valeurs manquantes par un ensemble de substitutions plausibles et l'imputation par la moyenne, la médiane ou le mode pour les variables quantitatives, sont des techniques couramment utilisées. Pour les variables qualitatives, l'adoption d'une catégorie "manquant" est une approche fréquente [64]. Cependant, dans notre cas, nous avons choisi d'éliminer les lignes présentant des valeurs manquantes. Cette décision a été guidée par le fait que seulement environ 5% des données étaient affectées, ce qui résulte en une perte de données jugée admissible. De plus, cette approche est fondée sur l'hypothèse que l'absence de ces données est indépendante des autres valeurs, réduisant ainsi le risque de biais dans les études ultérieures.

Enfin, les données utilisées sont ensuite divisées en un jeu d'entraînement (composé d'environ 80% des données) et un jeu de test (environ 20% des données restantes). Afin que les jeux de données d'entraînement et de test soient bien représentatifs, il est nécessaire d'effectuer cette division de manière stratifiée, autrement dit, en respectant dans chaque jeu de données, les proportions d'avaloirs au sein des différents clusters. Cette stratification est nécessaire lorsque les différents clusters sont disproportionnés car les clusters les plus petits pourraient ne pas être fidèlement représentés dans les jeux de données construits. Cependant, la stratification nécessite de prendre un découpage comme référence parmi $\mathcal{C}^3$, $\mathcal{C}^4$, et $\mathcal{C}^{Dyn}$ nous étudions ici. Les clusters les plus petits dans notre étude sont $\mathcal{C}_3^3$, $\mathcal{C}_3^4$ et $\mathcal{C}_4^4$. Or, comme les clusters $\mathcal{C}_3^3$ et $\mathcal{C}_4^4$ sont des ensembles très semblables, nous avons choisi de prendre le découpage $\mathcal{C}^4$ comme référence pour notre stratification. Ainsi, les jeux d'entraînement et de test respectent les proportions de $\mathcal{C}^4$ et restent très proches des proportions de $\mathcal{C}^3$ et de $\mathcal{C}^{Dyn}$.

4.5.2 Score de classification

Dans la littérature relative à la classification, plusieurs scores sont couramment utilisés pour évaluer la performance d'un modèle [49]. Parmi les plus populaires, nous retrouvons l'*accuracy* (ou exactitude), la *precision* (précision) et le *recall* (rappel). Par exemple, l'accuracy mesure le rapport entre les prédictions correctes et le nombre total de prédictions, comprise entre 0 et 1.

Cependant, ces scores ont certaines limitations. L'une des plus notables est leur sensibilité à la proportion des classes au sein du jeu de données. Par exemple, si dans un ensemble de données, 95% des échantillons appartiennent à la classe 0 et seulement 5% à la classe 1, un modèle prédisant systématiquement la classe 0 obtiendrait un score d'accuracy de 95%. Dans ce cas, cette performance masque une classification totalement erronée de la classe 1.

Face à ce type de déséquilibre, il est courant en pratique, et surtout dans les contextes multiclasses, de recourir à des scores moyennés. Cela signifie que des scores telles que la précision, le recall ou le *score F1* (défini comme moyenne harmonique de la précision et du recall) sont calculés individuellement pour chaque classe, puis moyennées afin de fournir une estimation globale.

Toutefois, ces scores moyennés peuvent ne pas suffire, car ils ne tiennent pas compte des performances que l'on pourrait obtenir par une simple classification aléatoire. Par exemple, le score *Balanced Accuracy* (BA) [49], compris entre 0 et 1, associé à un modèle aléatoire prédisant la classe 0 une fois sur deux serait de 0.5 [5], ce qui peut laisser penser à un modèle dont les performances seraient éventuellement moyennes.

De manière plus générale, le score BA d'un modèle de classification indépendant des données que l'on souhaite classifier, est $\mathrm{BA_r} = \frac{1}{\text{Nombre de classe}}$, comme montré en Annexe G. Dans notre contexte, où le nombre de catégories $= 2, 3, 4$, le score BA d'un estimateur indépendant des données est non-négligeable. Afin de corriger ce problème d'échelle, des scores tels que l'*Adjusted Balanced Accuracy* (ABA) ont été proposés [49].

Le score ABA évalue la performance en comparant le score BA d'un modèle au score que l'on obtiendrait avec une classification aléatoire [49]. Formellement, si BA est l'accuracy équilibrée du modèle et $\mathrm{BA_r}$ celle d'une classification aléatoire, le score ABA est donnée par

$$\mathrm{ABA} = \frac{\mathrm{BA} - \mathrm{BA_r}}{1 - \mathrm{BA_r}} .$$

5. Car, pour chaque échantillon de la classe 0 (respectivement 1), le modèle a une chance sur deux de prédire la classe 0 (respectivement 1). Ainsi le taux de prediction correct au sein de chaque classe serait de 50%. Le score BA, correspondant à la moyenne de ces taux de prédiction correct, est donc également de 50%.

Ainsi, le score BA est ajusté de manière à ce qu'une performance aléatoire reçoive un score de 0, tandis qu'une performance parfaite obtienne un score de 1. Le score ABA est particulièrement utile, notamment lorsque les classes sont déséquilibrées, car elle offre une perspective plus nuancée des performances réelles du modèle par rapport à un classificateur aléatoire. Pour ces raisons, nous avons décidé par la suite d'évaluer les résultats de classification en utilisant l'ABA.

4.5.3 Régression logistique

Description

Lorsque l'on étudie comment plusieurs facteurs combinés peuvent influencer ou expliquer un résultat catégorique, différentes méthodes peuvent être envisagées. Le choix de la méthode dépend généralement de la nature des données, du nombre de catégories à prédire et des hypothèses sous-jacentes que l'on est disposé à accepter. La première méthode que nous avons utilisée est celle de la *régression logistique*, décrite ci-dessous.

La régression logistique est une technique de modélisation qui vise à estimer la probabilité qu'une observation appartienne à une catégorie donné en fonction de plusieurs facteurs explicatifs en entrée [32]. Dans le cas d'une classification binaire, elle estime la probabilité qu'une observation appartienne à la classe 1, par opposition à la classe 0. Cette probabilité est donnée par la fonction sigmoïde d'une combinaison linéaire des variables d'entrée, où :

$$\text{sigmoid}(z) = \frac{1}{1 + e^{-z}}.$$

Ainsi, l'équation de la régression logistique est

$$p(Y = 1|\boldsymbol{x}) = \hat{y}_i = \text{sigmoid}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)\,,$$

En pratique, on utilise un jeu de données d'entraînement afin d'estimer ces coefficients en maximisant la vraisemblance, c'est à dire en minimisant la fonction de coût logistique. La fonction de coût logistique (ou log loss) pour la régression logistique binaire est définie comme suit :

$$\text{Log Loss}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right],$$

avec y_i comme la valeur réelle de la i-ème observation (0 ou 1 dans le cas binaire) et $\hat{y}_i$ comme la valeur prédite correspondante.

Pour les cas avec plus de deux catégories, on utilise la régression logistique multinomiale. L'approche la plus courante est la méthode *one-versus-rest* où on estime un modèle pour chaque catégorie. L'idée est de revenir à un cas binaire

en estimant la probabilité que chaque observation appartienne à une catégorie particulière par rapport à toutes les autres.

Pour éviter le surapprentissage du modèle, en particulier lorsque l'on utilise un grand nombre de facteurs en entrée, différentes méthodes de pénalisation peuvent être employées. Parmi celles-ci, on retrouve la *Regression Isotropic Distributed Gaussian Estimate* (RIDGE) [32], le *Least Absolute Shrinkage and Selection Operator* (LASSO) [32] et *ElasticNet* [32].

Le LASSO peut réduire les coefficients du modèle à zéro, favorisant ainsi une meilleure sélection de variables en éliminant celles qui ont peu d'impact sur la prédiction. La régularisation RIDGE est particulièrement efficace pour gérer les variables corrélées. Enfin, ElasticNet est une combinaison pondérée des régularisations LASSO et RIDGE.

La fonction de coût pour un modèle de régression avec pénalisation s'écrit :

$$J(\boldsymbol{\beta}) = \text{Log Loss}(\boldsymbol{\beta}) + \alpha \cdot \text{Pénalisation}(\boldsymbol{\beta})$$

où $J(\boldsymbol{\beta})$ est la fonction à minimiser, Log Loss$(\boldsymbol{\beta})$ est la fonction de coût logistique pour les coefficients de régression $\boldsymbol{\beta}$, et α est un paramètre de régularisation qui contrôle la force de la fonction Pénalisation choisie (RIDGE, LASSO, etc.). Une valeur de α plus élevée donnera lieu à une pénalisation plus forte des coefficients. Enfin, on peut noter que la complexité calculatoire de l'algorithme de régression logistique est typiquement de l'ordre de $O(n \times p \times i)$, où n est le nombre d'observations, p le nombre de facteurs explicatifs et i le nombre d'itérations nécessaires pour la convergence.

Application de la régression logistique

Pour établir un lien entre les facteurs contextuels construits et la dynamique d'encrassement, nous avons employé la régression logistique pénalisée. Certains de nos facteurs sont corrélés, comme le montre la Figure 4.6, représentant un extrait de la matrice de corrélation entre les différents facteurs contextuels. Afin d'aborder à la fois la sélection de variables (comme le fait la méthode LASSO) et la gestion de la multicollinéarité (comme le propose la méthode RIDGE), ElasticNet nous est apparu comme un compromis judicieux. Nous avons opté pour un ratio de 0.5 entre LASSO et RIDGE, car ce choix nous semblait équilibré en l'absence d'une raison de privilégier l'une ou l'autre méthode. Par la suite, lors des entraînements du modèle, on compense le déséquilibre entre les différents effectifs des classes (quelque soit le découpage étudié) en attribuant un poids à chaque classe, correspondant à l'inverse de la fréquence d'appartenance de chaque classe.

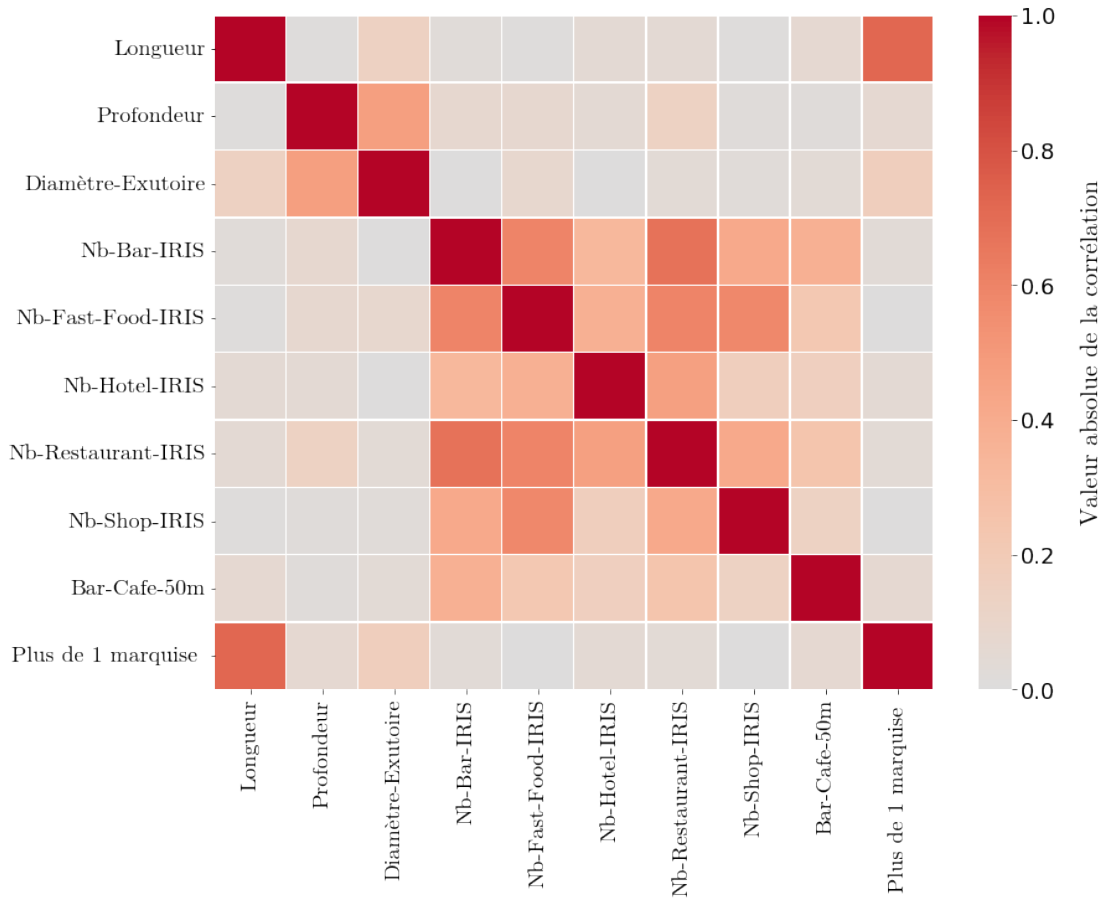

FIGURE 4.6 – Extrait de la matrice de corrélation entre les facteurs contextuels. On peut constater que certains facteurs sont plus ou moins correlés entre eux.

On rappelle que l'entraînement de notre modèle est basé sur une répartition des données, allouant 80% à l'entraînement et 20% pour la validation finale. L'un des principaux hyperparamètres à déterminer est le niveau de pénalisation à appliquer à la régression logistique. En effet, une pénalisation forte tend à réduire de nombreux coefficients vers zéro, favorisant une sélection stricte des variables. À l'opposé, une faible pénalisation pourrait conserver presque tous les facteurs, ne réalisant pas la sélection désirée et présentant ainsi un risque de surajustement. Pour identifier le niveau de pénalisation optimal, nous avons utilisé la *validation croisée* sur le jeu d'entraînement [65]. Nous avons divisé celui-ci en 5 plis ; pour chaque niveau de pénalisation considéré, où 4 plis sont employés pour l'entraînement et le cinquième pour la validation. Cette procédure est répétée cinq fois, de sorte que chaque pli serve une fois pour la validation. Les scores ABA pour chaque pli sont ensuite moyennés afin de produire un score consolidé, offrant une évaluation robuste des performances de prédiction du modèle tout en minimisant les variations liées aux

spécificités d'un sous-ensemble donné. Cette méthode de validation croisée assure que le modèle est bien adapté pour pouvoir ensuite généraliser au-delà des données d'entraînement.

Dans le cas binaire, où l'on souhaite prédire la présence de dynamique (correspondant au découpage $\mathcal{C}^{Dyn}$), on peut tracer les *coefficients paths*, représentés sur la Figure 4.7, qui illustre comment les coefficients β du modèle évoluent en fonction du niveau de pénalisation appliqué. Sur cette figure, chacune des courbes est associée à un facteur contextuel ; pour des raisons de clarté, nous avons seulement identifié trois facteurs (sur les ~ 70 facteurs considérés). Lorsque la pénalisation augmente, certains de ces coefficients sont réduits à zéro, ce qui équivaut à écarter le facteur associé. De manière complémentaire, la Figure 4.8 montre l'évolution du score ABA lors de la validation croisée en fonction du niveau de pénalisation (toujours pour la prédiction de $\mathcal{C}^{Dyn}$).

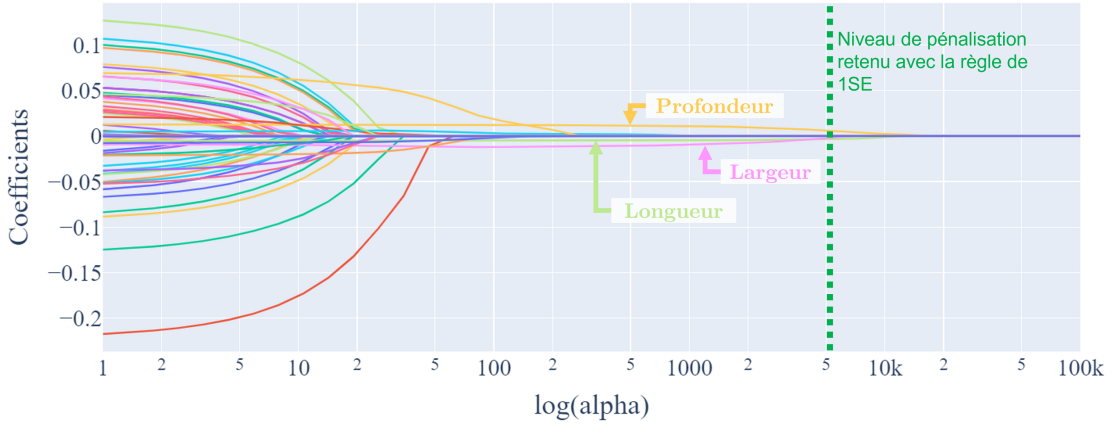

FIGURE 4.7 – Coefficients paths montrant l'évolution des coefficients associés aux différents facteurs contextuels pour la prédiction de $\mathcal{C}^{Dyn}$. Pour des raisons de clarté, le facteur est indiqué seulement sur 3 courbes parmi 70 : ces courbes correspondent aux facteurs sélectionnés après l'application de la méthode 1SE pour la calibration du niveau de pénalisation.

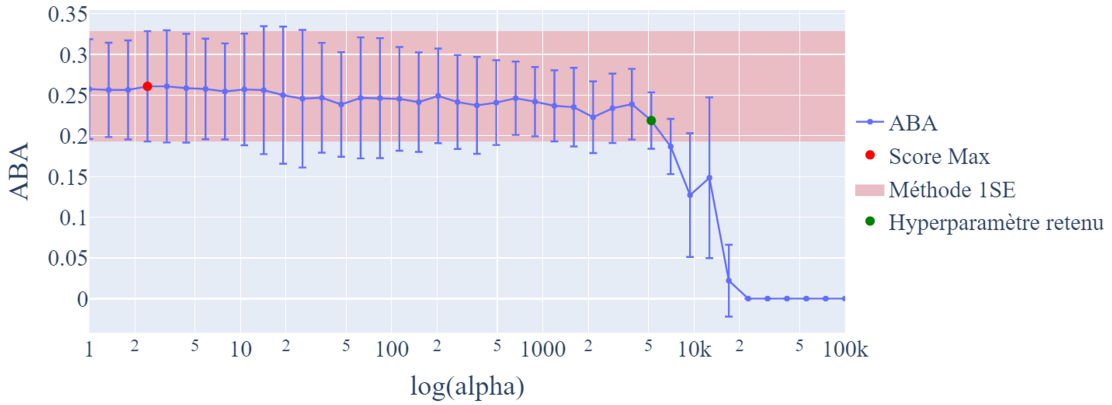

FIGURE 4.8 – Évolution du score ABA de validation croisée en fonction du niveau de pénalisation. La bande rouge illustre l'application de la règle du 1SE pour la sélection du niveau de pénalisation optimal (indiqué par un point vert) pour la prédiction de $\mathcal{C}^{Dyn}$.

En pratique, à partir de ces deux résultats, on choisit les hyperparamètres de sorte à avoir le modèle le plus simple offrant les meilleures performances possibles. Pour ce faire, on peut se baser sur la règle du *One Standard Error* (1SE). Cette règle suggère de choisir le modèle le plus simple dont le score de validation croisée est à moins "d'une erreur standard" du score maximal observé. L'erreur standard est obtenue en calculant l'écart-type empirique des scores de validation croisée pour les différents niveaux de pénalisation évalués. En appliquant la règle du 1SE, on favorise un modèle plus parcimonieux et potentiellement plus généralisable, tout en garantissant des performances proches de celles du modèle optimal.

Dans notre cas, en adoptant la règle 1SE, le niveau de pénalisation retenu est d'environ 5200 (cf. Figure 4.8). À ce niveau de pénalisation, on constate sur la Figure 4.7 que seuls certains facteurs ont un coefficient non nul. Les facteurs sélectionnés par le modèle sont la longueur, la largeur et la profondeur de l'avaloir. En réentraînant le modèle uniquement avec les facteurs sélectionnés, et sans pénalisation cette fois, on peut évaluer les performances de prédiction du modèle sur le jeu de données test à ABA $\simeq 0.28$.

Les Figures 4.9 et 4.10 représentent respectivement l'évolution du score ABA de la validation croisée en fonction du niveau de pénalisation pour les prédictions de $\mathcal{C}^3$ et $\mathcal{C}^4$. Pour des raisons de brièveté, nous n'exposons pas les trajectoires des coefficients associées à ces deux cas. On rappelle que dans le contexte d'un problème de prédiction multiclasses, il faut mettre en place un modèle de régression logistique pour chaque classe. Ainsi, chaque modèle calcule la probabilité qu'une observation appartienne à sa classe spécifique [6]. Comme chaque modèle possède son

6. On rappelle que dans ce cas, la classe attribuée à une observation donnée est celle pour laquelle le modèle a calculé la probabilité la plus élevée.

propre ensemble de coefficients, et que ceux-ci varient avec le niveau de pénalisation, cela se traduirait par une série de 3 graphiques pour $\mathcal{C}^3$ et une autre série de 4 graphiques pour $\mathcal{C}^4$.

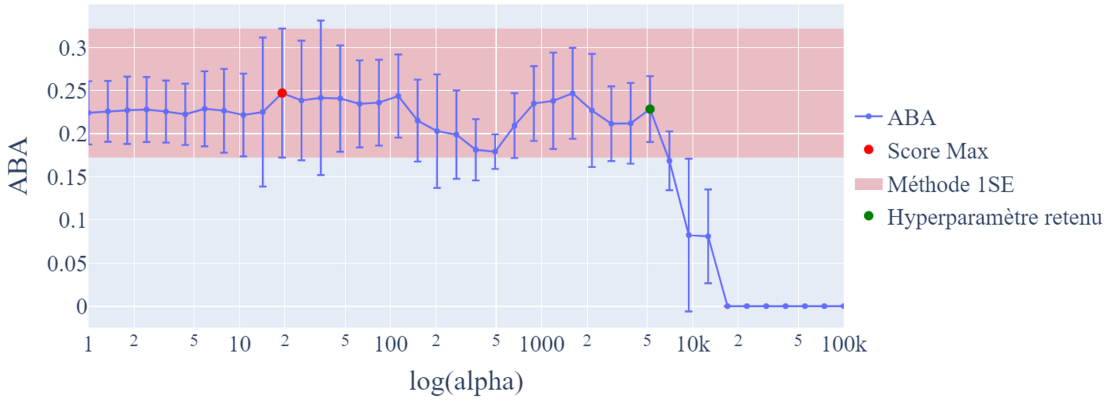

FIGURE 4.9 – Évolution du score ABA de validation croisée en fonction du niveau de pénalisation. La bande rouge illustre l'application de la règle du 1SE pour la sélection du niveau de pénalisation optimal (point vert) dans la prédiction de $\mathcal{C}^3$.

FIGURE 4.10 – Évolution du score ABA de validation croisée en fonction du niveau de pénalisation. La bande rouge illustre l'application de la règle du 1SE pour la sélection du niveau de pénalisation optimal (point vert) dans la prédiction de $\mathcal{C}^4$.

Toujours en appliquant la méthode du 1SE, on constate sur les Figures 4.9 et 4.10 que les niveaux de pénalisation retenus sont respectivement de $\sim 5\,200$, et de ~ 7.9 pour $\mathcal{C}^3$ et $\mathcal{C}^4$. Les scores ABA obtenus sur le jeu de donnée test sont respectivement de ~ 0.24 et ~ 0.14.

Enfin, il est également possible de choisir d'autres niveaux de pénalisation pour chacun des cas étudiés, comme par exemple le niveau lié à la valeur maximale du score ABA obtenu pendant la validation croisée. Il est également possible de choisir le niveau de pénalisation en se basant uniquement sur les coefficients paths. Par exemple, sur la Figure 4.7, on constate qu'à partir d'un niveau de pénalisation ~ 25, la plupart des facteurs sont éliminés par l'ElasticNet, ne laissant donc qu'un nombre réduit de facteurs. Ce niveau de pénalisation peut donc sembler particulièrement intéressant. Il convient de noter que lorsque l'on choisit ce niveau de pénalisation et que l'on réentraîne le modèle uniquement avec les facteurs sélectionnés, le score ABA sur le jeu de données test est de ~ 0.21.

Nous avons exploré plusieurs niveaux de pénalisation possibles pour chacun des trois découpages $\mathcal{C}^3$, $\mathcal{C}^4$ et $\mathcal{C}^{Dyn}$. Nous ne détaillerons pas chacun de ces résultats, cependant, nous pouvons faire les constats notables suivants.

- Globalement, en observant l'évolution des coefficients β de la régression en fonction du niveau de pénalisation, on constate que le premier lot important de facteurs à être éliminé est composé des facteurs liés au contexte spatial proche de l'avaloir (présence d'arbres à 10 mètres, présence de de bar/café/pub à 50 mètres etc.). Ensuite, sont éliminés les facteurs concernants plutôt les zones associés à l'avaloir (nombre de restaurants dans le quartier, nombre de shop dans l'IRIS, etc.), ainsi qu'éventuellement certains éléments structuels comme la présence de grilles ou le diamètre de l'exutoire. Enfin, la largeur, la longueur et la profondeur de l'avaloir font systématiquement parties des facteurs qui subsistent même avec une forte pénalisation.
- Les meilleurs scores ABA obtenus sur le jeu de données test sont ceux obtenus lorsque l'on utilise uniquement les facteurs liés à la structure de l'avaloir [7].

Finalement, les scores ABA supérieurs à zéro confirment que nos modèles surpassent une attribution aléatoire des classes aux avaloirs. Cela montre que les facteurs contextuels, surtout ceux relatifs à la structure des avaloirs, détiennent effectivement des informations explicatives sur leur distribution dans les catégories étudiées. Toutefois, avec des scores ABA demeurant en deçà de 0.3, la capacité de prédire la classe d'un avaloir strictement à partir de ces facteurs par la régression logistique s'avère modeste. Nous nous sommes donc tournés vers des techniques plus avancées comme le Random Forest, dans l'espoir de saisir des interactions plus complexes entre les facteurs et ainsi améliorer la précision de nos prédictions.

7. présence de grilles, diamètre de l'exutoire, largeur, longueur, et profondeur

4.5.4 Random Forest

Description

La méthode des forêts aléatoires ou *Random Forest* est une approche issue de l'apprentissage automatique qui combine plusieurs arbres de décision afin de produire un modèle global plus performant [32, 66]. Un *arbre de décision* est un modèle d'apprentissage automatique qui fonctionne en divisant un ensemble de données en sous-ensembles plus petits et plus homogènes. Cette division se fait en fonction des caractéristiques (également appelées features, ou attributs) des données. À chaque étape, l'algorithme choisit la meilleure caractéristique pour diviser les données, en se basant sur des critères tels que l'indice de Gini ou l'entropie, qui mesurent la pureté des sous-ensembles [66]. La pureté ici signifie que les éléments de chaque sous-ensemble sont aussi similaires que possible en termes de la variable cible. Le processus se poursuit jusqu'à ce que des critères d'arrêt soient atteints, tels que le sous-ensemble ne pouvant plus être divisé ou atteignant une taille minimale. À ce moment, chaque feuille de l'arbre symbolise un résultat prédictif fondé sur les caractéristiques des données dans ce sous-groupe spécifique.

Le *Random Forest* étend ce concept en utilisant plusieurs arbres de décision. Chaque arbre est construit sur un échantillon différent de l'ensemble des données, ce qui est réalisé en utilisant une technique appelée *échantillonnage bootstrap*, correspondant à un tirage avec remise des données. De plus, lors de la construction de chaque arbre, une sous-sélection aléatoire des caractéristiques est effectuée pour chaque division. Cette méthode crée une diversité parmi les arbres et réduit le risque de surapprentissage. La prédiction finale du modèle Random Forest est déterminée en combinant les prédictions de tous les arbres, par exemple en prenant la moyenne des prédictions pour une tâche de régression ou le vote majoritaire pour une tâche de classification. En regroupant les résultats de plusieurs arbres, le Random Forest réduit grandement la variance d'un arbre unique. La complexité computationnelle de cet algorithme est généralement de l'ordre de $O(E \times K \times N \times \log(N))$, où E est le nombre d'arbres dans la forêt, K le nombre de caractéristiques examinées pour chaque division dans un arbre, et N le nombre d'échantillons dans le jeu de données. Le terme $\log(N)$ représente la complexité de la construction d'un arbre de décision individuel, en supposant un équilibrage parfait de l'arbre. Autrement dit, que chaque division de l'arbre divise les données de manière égale, permettant à l'arbre de croître de manière logarithmique par rapport à la taille des données, plutôt que de manière linéaire.

Application du Random Forest

Afin de calibrer le modèle Random Forest, nous avons utilisé la méthode *Grid Search* qui est une technique de recherche exhaustive pour optimiser les

hyperparamètres. Celle-ci consiste à tester différentes combinaisons de valeurs pour les hyperparamètres essentiels du Random Forest, tels que le nombre d'estimateurs, la profondeur maximale des arbres et le nombre minimal d'échantillons requis pour diviser un nœud interne.

Comme précédemment, l'entraînement et la validation du modèle sont effectués sur les jeux de données construits préalablement (pour rappel, nous avons divisé de manière stratifiée les données en deux parties : 80% pour l'entraînement et 20% pour la validation (cf. Section 4.5.1)). Pour juger de l'efficacité de chaque combinaison d'hyperparamètres, nous avons utilisé à une validation croisée en 5 plis (cf. Section 4.5.3).

Suite à cette calibration, nous avons identifié la combinaison d'hyperparamètres la plus performante selon le score ABA. Le modèle optimal a été réentrainé avec cette configuration pour évaluer sa performance finale. Cependant, malgré la variété des combinaisons d'hyperparamètres testées, les performances de classification du modèle Random Forest sont restées très limitées, avec un score ABA $\simeq 0.1$ dans le meilleur cas. Ces performances, parfois comparable à une classification aléatoire (ABA $\simeq 0$) ou à peine meilleure indique une difficulté significative du modèle à discerner les motifs pertinents dans les données pour réaliser des prédictions fiables.

Par contraste, bien que la régression logistique fournisse des performances modestes avec un score ABA de ~ 0.25, ce résultat est supérieur à celui du Random Forest. Cela indique que malgré sa simplicité, la régression logistique parvient à mieux saisir la structure linéaire potentielle des données.

Quand un modèle plus simple tel que la régression logistique surpasse un modèle ensembliste comme le Random Forest, cela peut refléter divers scénarios : soit les données ne présentent pas de relations complexes que le Random Forest serait en mesure d'exploiter, soit le bruit présent dans les données rend les modèles plus sophistiqués moins performants car ils se concentrent sur celui-ci.

4.6 Synthèse du chapitre

Les observations métier montrent que certains éléments contextuels peuvent influencer l'encrassement des avaloirs. Ce chapitre porte sur l'étude de l'impact de ce contexte sur la dynamique d'encrassement des avaloirs et, plus précisément, sur la répartition des avaloirs dans les différentes catégories de comportement établies dans le chapitre précédent. Les différentes étapes de cette étude sont synthétisées sur la Figure 4.11.

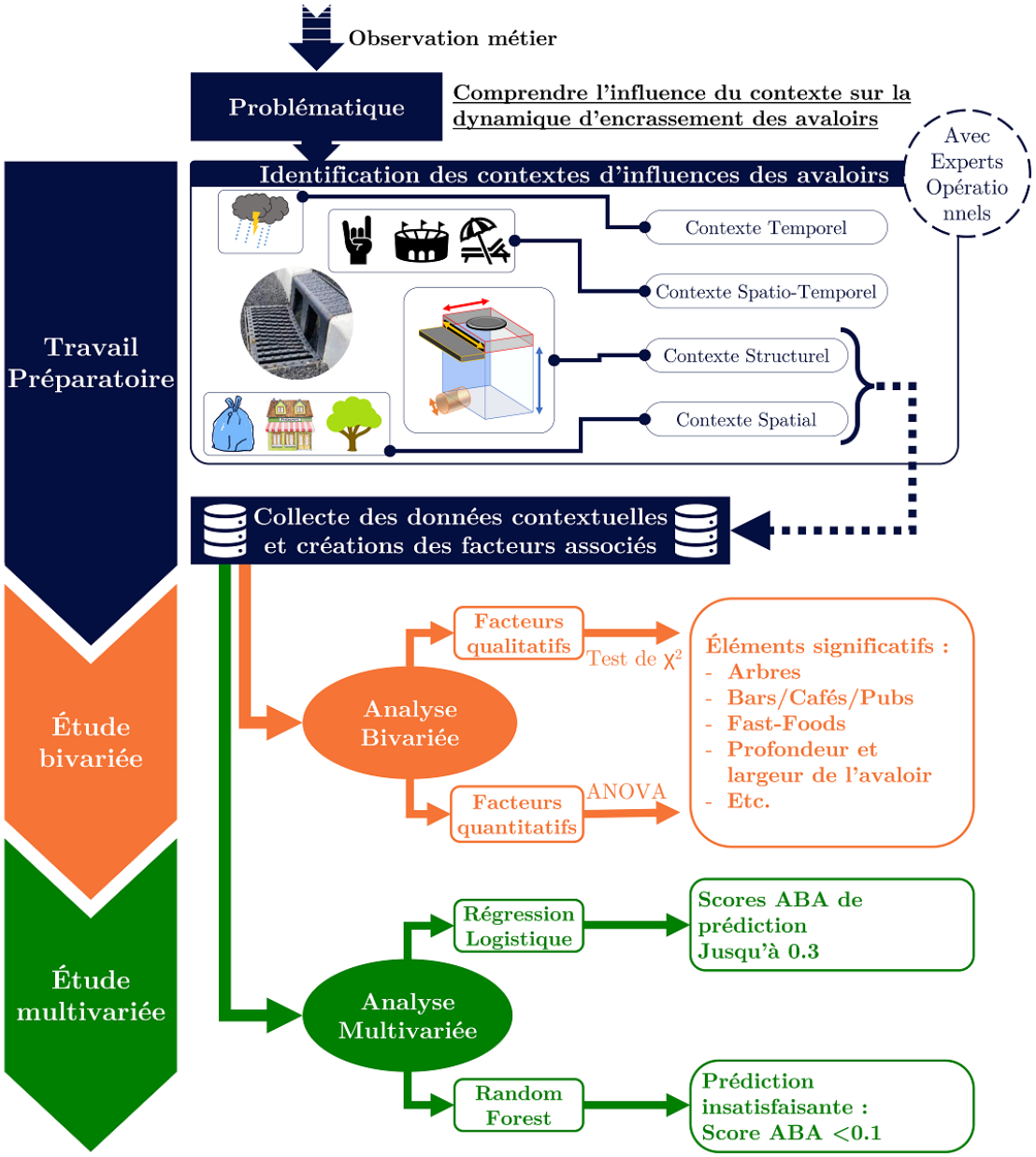

FIGURE 4.11 – Synthèse de l'étude contextuelle.

L'étude se concentre d'abord sur l'identification des éléments contextuels avec l'aide d'experts. Nous avons identifié quatre types d'éléments contextuels : structurels, spatiaux, temporels et spatio-temporels. Les éléments structurels incluent des caractéristiques comme la largeur et la profondeur de l'avaloir, tandis que les éléments spatiaux concernent l'environnement immédiat, comme la proximité d'arbres ou de fast-foods. Les éléments temporels se rapportent, par exemple, aux conditions

météorologiques comme les précipitations, et les éléments spatio-temporels incluent des événements spécifiques comme les manifestations sportives ou culturelles. Dans notre étude, nous nous somme principalement concentrés sur les contextes structurel et spatial. Après une collecte et mise en forme fastidieuse des données contextuelles à partir de la base de données publique, environ 70 facteurs contextuels ont été définis, majoritairement qualitatifs, pour décrire ces éléments contextuels. L'analyse de ces facteurs a révélé une corrélation significative entre certains éléments, comme la présence d'arbres ou de fast-foods, et un encrassement accru des avaloirs. Des facteurs structurels spécifiques, comme le diamètre de l'exutoire, ont également montré une influence notable.

Dans un deuxième temps, une approche multivariée, incluant la régression logistique et le Random Forest, a été utilisée pour étudier l'impact combiné de ces facteurs et essayer de prédire la dynamique d'encrassement des avaloirs à partir de leur contexte. Cependant, les résultats ont été modestes, suggérant soit la simplicité des relations dans les données, soit les limites de leur qualité. La prédiction de la classe d'un avaloir basée uniquement sur son contexte s'est avérée complexe, indiquant éventuellement la nécessité d'explorer des techniques plus avancées ou de collecter des données supplémentaires, notamment sur les contextes temporel et spatio-temporel.

Conclusions générales et perspectives

Conclusions générales

Cette thèse s'inscrit dans le cadre d'un projet mené par le service d'assainissement de Marseille, une filiale de SUEZ, visant à équiper environ 5000 avaloirs de la ville d'un dispositif de mesure du niveau d'encrassement et sa télétransmission vers un centre de stockage et de calcul. Les avaloirs, communément appelés "bouches d'égout", sont des ouvertures situés principalement le long des trottoirs qui ont pour rôle d'absorber les eaux pluviales. Cependant, leur efficacité peut être entravée par l'accumulation de divers déchets et de feuillages, entraînant ainsi des nuisances visuelles, des risques d'inondation en cas de fortes pluies, des dommages aux équipements du réseau d'assainissement, ou encore des conséquences environnementaux négatifs si les déchets sont rejetés en mer. Jusqu'en 2020, la maintenance des avaloirs reposait sur des vérifications visuelles de l'état d'encrassement sur terrain (visites), suivie, si nécessaire, d'un "curage" (nettoyage de l'avaloir). Cette maintenance était basée sur des quotas annuels de 50 000 visites et 25 000 curages, et était être limitée en termes d'éfficacité car elle engendrait des interventions inutiles ou tardives. Afin d'optimiser cette maintenance, en intervenant au bon endroit et au bon moment, la solution retenue consiste à équiper chaque avaloir d'un capteur permettant de suivre à distance et en temps réel son niveau d'encrassement. Ce capteur, placé dans la partie supérieure de l'avaloir et orienté vers le bas, mesure le niveau d'encrassement en calculant le temps de parcours d'une impulsion ultrasonore jusqu'à son écho. Ces mesures sont envoyées vers le centre de stockage et de calcul. Un tel dispositif permet de suivre l'état d'encrassement en temps réel avec une fréquence typique de deux fois par jour mais qui peut être modifiée si besoin. L'analyse des données collectées par ces capteurs révèle une diversité dans la dynamique d'encrassement. Par exemple, il est observé que l'encrassement peut évoluer progressivement ou brusquement, et qu'il peut diminuer même sans curage, un phénomène appelé *lessivages*. De plus, les observations opérationnelles indiquent que le niveau d'encrassement peut fluctuer en fonction de divers facteurs extérieurs

et environnementaux.

L'objectif principal de cette thèse est donc d'étudier et de comprendre la dynamique d'encrassement des avaloirs, en s'appuyant sur les données fournies par les capteurs. Cependant, avant de nous lancer dans cette analyse, une phase essentielle d'exploration, de préparation, et de nettoyage des données a été nécessaire. Ces travaux nous ont permis de constater des valeurs manquantes, résultant à la fois de problèmes de référencement du réseau d'avaloirs et de problèmes de transmission des mesures par les capteurs. Nous avons également traité les mesures incohérentes qui surpassaient la distance physique entre le capteur et le fond de l'avaloir en établissant un seuil. Ce seuil correspond à la plus grande valeur entre la distance maximale de fond et la profondeur de l'avaloir, augmentée d'une marge d'erreur de 20 cm pour prendre en compte les problèmes de référencement, tels que des profondeurs incorrectes ou manquantes, et les problèmes de calibration, comme une calibration inadéquate ou absente. Cette méthode a permis de supprimer les mesures erronées tout en conservant les données potentiellement valides. Nous avons également abordé le problème des mesures redondantes. La configuration des capteurs est telle qu'il devrait normalement y avoir un intervalle d'au moins 10 minutes entre deux mesures. Cependant, nous avons observé des occasions où cet intervalle était plus court que prévu. Dans ces cas, nous avons éliminé les mesures redondantes qui étaient espacées de moins de 10 minutes. Par ailleurs, nous avons constaté que certains capteurs transmettaient des données aberrantes ou incohérentes sur des périodes étendues, possiblement en raison de problèmes d'installation. Ces capteurs sont éliminés de l'étude dès leur identification au cours de notre travail. Un autre aspect important est que la variabilité des mesures semble dépendre de la surface sur laquelle se reflète l'onde ultrasonore émise. On intuite que cette surface, doit être idéalement plane et horizontale pour une mesure précise par le capteur. En pratique, cette surface dépend de la forme et des dimensions du fond de l'avaloir si celui-ci est vide ou bien de la disposition de l'encrassement dans l'avaloir, qui peut varier avec le temps. En plus de la variabilité habituelle, des "outliers" sous forme de pics aberrants ont été observés, divergeant de la tendance générale. Ces anomalies ont été attribuées à la nature de la surface mesurée. Pour détecter ces pics, nous avons proposé un algorithme de détection d'anomalies appelé PPZ (pour Peak Pattern based Z-score). Cet algorithme combine le Z-score avec un score de détection de motifs, OVD (Opposite Variation Detection), conçu spécifiquement pour détecter ce type d'anomalies. Suite à ce processus minutieux d'exploration et de nettoyage des données, nous avons estimé que la qualité des données après ces prétraitrements était suffisante pour aborder notre problématique principale, à savoir, comprendre la dynamique d'encrassement des avaloirs à Marseille.

Comme évoqué précédemment, nous avons observé l'historique des niveaux

d'encrassement des avaloirs et constaté divers comportements. Pour analyser cette dynamique d'encrassement, nous avons regroupé les avaloirs selon leur dynamique en utilisant des algorithmes de classification non supervisée (clustering). Notre démarche, basée sur des attributs (features-based approach), nous a permis d'expérimenter avec diverses combinaisons d'attributs, d'algorithmes, et d'hyperparamètres, tout en s'assurant que les résultats soient interprétables et en quantités réduites pour faciliter l'analyse. Cette étude a concerné environ 2200 avaloirs avec au moins un an de données historiques.

La **méthodologie adoptée** pour cette analyse comprenait plusieurs étapes. La première était la **création d'attributs**, où nous avons généré des attributs empiriques, inférentiels, et basés sur des excursions pour décrire la dynamique d'encrassement tout en conservant la dimension temporelle des données. Ensuite, dans la **sélection des attributs**, nous avons minimisé la redondance en regroupant les attributs en familles en fonction de leur corrélation, et en sélectionnant les plus représentatifs de chaque famille. Le **choix des algorithmes** de clustering a été guidé par la diversité, en sélectionnant des algorithmes basés sur différents principes. Nous avons choisi K-means, Spectral Clustering et DBSCAN pour leur approches complémentaires. Pour chaque combinaison d'attributs et d'algorithme, nous avons **ajusté les hyperparamètres** en utilisant des scores de clustering tels que le Silhouette Score et l'indice de Calinski-Harabasz, pour mesurer la qualité des clusters.

Face à la grande quantité de combinaisons d'attributs, d'algorithmes de clustering et d'hyperparamètres, chacune produisant un résultat de clustering différent, nous avons mis en place une stratégie de **regroupement des résultats similaires** en utilisant le Kappa de Cohen pour évaluer la similarité entre les différents résultats de clustering. Grâce à cette approche, nous avons pu réduire le nombre de résultats de clustering à analyser, en ne visualisant et interprétant qu'un seul résultat représentatif pour chaque groupe de résultats similaires.

Les résultats obtenus ont révélé des tendances intéressantes. Nous avons identifié deux regroupements particulièrement pertinents de 3 et 4 clusters. Une proportion significative d'avaloirs présentait un faible niveau d'activité, tandis que d'autres étaient sensibles à de grandes variations, évoquant une éventuelle sensibilité aux déchets volumineux. Certains clusters indiquaient des avaloirs se remplissant progressivement, tandis qu'un autre regroupait des avaloirs avec des remplissages et pertes réguliers des déchets. Notamment, nous avons constaté que la majorité des avaloirs dans ce dernier cluster n'avaient pas été curés, ou l'avaient été une seule fois en $\sim$ 12 mois, ce qui souligne leur potentiel risque en termes de rejet de déchets.

Ces analyses ont permis d'identifier les principaux comportements des avaloirs et les proportions qu'ils représentent. Compte tenu des observations métier indiquant que certains éléments contextuels peuvent influencer l'encrassement des avaloirs,

nous nous sommes concentrés ensuite sur l'analyse de l'impact de ce contexte sur les différents comportements d'encrassement identifiés, **afin d'approfondir notre compréhension de la dynamique d'encrassement**.

La première étape de ces travaux est **l'identification de ces éléments contextuels** avec l'aide de deux experts métiers. Les éléments identifiés peuvent être liés à la *structure* de l'avaloir, comme la largeur, la profondeur ou la présence de grilles, par exemple. Ils peuvent également concerner l'environnement immédiat de l'avaloir, tels que la proximité d'arbres ou de fast-foods, qui influencent l'encrassement dans des zones *spatiales* spécifiques. Par ailleurs, les conditions météorologiques, et en particulier les précipitations, sont des facteurs *temporels* (en supposant que la pluie soit homogène spatialement) pouvant jouer un rôle important. Enfin, des phénomènes *spatio-temporels*, tels que les événements sportifs ou culturels, peuvent avoir un impact localisé, par exemple, autour d'un stade ou des plages.

Nous nous sommes alors focalisés dans un premier temps sur les deux premiers types de contexte : structurel et spatial. Pour aborder cette problématique, un travail de récolte et de mise en forme des données a été mené afin d'établir une base de données contextuelle des avaloirs. Ainsi, pour chaque élément contextuel identifié, nous avons **créé un ou plusieurs facteurs contextuels**. Parmi les ~ 70 facteurs construits une majorité (~ 50) sont qualitatifs et peuvent décrire la structure de l'avaloir (par exemple, la présence de grille "Vrai ou Faux"), l'entourage proche d'un avaloir (par exemple, la présence d'arbres dans un rayon de 30 mètres autour de l'avaloir), ou encore la zone où se trouve l'avaloir (quartier ou IRIS avec une densité de bars supérieure à celle des autres quartiers ou IRIS). Les facteurs restants sont quantitatifs, comme la profondeur de l'avaloir en cm ou la densité de fast-foods par km^2 dans le quartier.

L'étude approfondie des facteurs contextuels structurels et spatiaux a révélé leur influence significative sur la dynamique d'encrassement des avaloirs. En **étudiant les facteurs construits individuellement**, les premières analyses basées sur des tests statistiques ont d'abord mis en lumière l'association entre la présence d'arbres, de commerces en tout genre, de bars/cafés/pubs et de fast-foods avec une dynamique d'encrassement plus prononcée, reflétée par des risques relatifs élevés. Ces corrélations pourraient refléter une causalité : par exemple, les feuilles des arbres pourraient être absorbées par les avaloirs, ou certains débris comme les canettes et les emballages retrouvés dans les avaloirs pourraient provenir des fast-foods. Ou encore, ces associations pourraient également être dues à des effets indirects. Par exemple, il est concevable que les zones comportant un nombre élevé de bars, cafés, pubs ou de commerces soient plus fréquentées, ce qui pourrait augmenter la probabilité d'incivilités, comme le fait de jeter des détritus dans la rue, qui, à leur tour, peuvent conduire à l'encrassement des avaloirs. Enfin, certains facteurs structurels comme le diamètre de l'exutoire, la largeur et la profondeur de l'avaloir

se sont révélés significatifs. Certaines corrélations peuvent sembler intuitives, comme par exemple le fait que les avaloirs avec de petits diamètres d'exutoire ont tendance à accumuler davantage d'encrassement, probablement car les débris ont plus de mal à être évacués. On peut imaginer que la profondeur et la largeur de l'avaloir affectent la manière dont les déchets tombent et se répartissent dans l'avaloir. Par exemple, on constate que les avaloirs les plus larges sont souvent moins dynamiques, possiblement parce que les débris sont mieux répartis à l'intérieur, réduisant ainsi les variations de la mesure de l'encrassement. Cependant, on constate également que les anomalies dans les données identifiées précédemment semblent fortement corrélées à la profondeur. Une explication pourrait être que la nature des mesures peut être affectée par les dimensions de l'avaloir. Cette piste montre une perspective intéressante à explorer afin d'améliorer la robustesse du dispositif.

Enfin, il est possible que certains facteurs aient peu d'influence individuellement, mais que leur combinaison ait un impact sur la dynamique d'encrassement. Afin de mieux identifier ces facteurs d'influence, ces combinaisons, et d'évaluer le caractère prédictif du contexte sur les clusters étudiés, nous avons opté pour **une approche multivariée**. Dans ce processus, la première méthode appliquée est la régression logistique qui a montré des résultats modestes en termes de performance prédictive avec un score ABA avoisinant les 0.3 au mieux. Ces performances restent supérieures aux résultats obtenus par le Random Forest, qui affiche un score ABA inférieur à 0.1, soulignant ainsi ses limites dans le cadre de cette étude. Ces résultats indiquent potentiellement soit une simplicité des relations dans les données, soit les limites de la qualité des données étudiées.

Perspectives

Alors que cette thèse a exploré de nombreuses facettes des données des capteurs et de leur analyse, un vaste champ de recherches reste à explorer pour approfondir nos résultats. Dans ce qui suit, nous présentons quelques améliorations et perspectives pour étendre le travail réalisé.

Dans le cadre de cette thèse, un effort considérable a été déployé pour nettoyer les données et identifier les capteurs non fiables. Cependant, il est possible que certains problèmes subsistent malgré ces efforts, en particulier parmi les capteurs que nous n'avons pas encore étudiés en détail, car ils ne disposaient pas d'un historique suffisant. Afin de poursuivre ces travaux, un projet spécifique mené par SUEZ est en cours, visant à identifier et résoudre les problèmes résiduels des capteurs. Ce projet a pour but de nettoyer davantage les données historiques et de mettre en place un système de détection en temps réel des anomalies de mesures, afin d'assurer une collecte de données fiable et précise, et effectuer des actions de maintenance de manière proactive.

Parallèlement, il est important d'effectuer une modélisation plus approfondie des mesures des capteurs, en prenant en compte la propagation des ondes ultrasonores et le profil de la surface mesurée. Cette étape est essentielle pour mieux comprendre l'origine des anomalies et améliorer leur détection.

Concernant le clustering, il est bien-sûr possible d'effectuer une recherche plus exhaustive de clusters en introduisant d'autres attributs ou essayant d'autres algorithmes et plus grand nombre d'hyperparamètres. La considération d'autres scores pour la sélection des attributs ou pour le regroupement des résultats de clustering offre également des perspectives pour trouver des comportements nouveaux.

Pour l'analyse contextuelle, l'intégration de nouveaux éléments contextuels, notamment ceux liés aux contextes temporels et spatio-temporels, combinée à l'utilisation d'autres algorithmes de classification, pourrait conduire à une meilleure prédiction des catégories d'avaloirs.

Notre analyse initiale, basée sur environ 2000 avaloirs, a validé l'intérêt de notre démarche. Il serait judicieux d'étendre cette étude en intégrant les données accumulées en 2023, qui n'avaient pas été prises en compte initialement. Bien que l'analyse complète du réseau doive plutôt attendre fin 2024, pour que chaque capteur dispose d'au moins un an d'historique. Enfin, avec une profondeur d'historique accrue, en particulier pour les capteurs disposant désormais de plusieurs années de données, nous pouvons réexaminer de manière plus détaillée l'influence saisonnière.

Enfin, de manière plus globale, nous envisageons deux pistes majeure de recherche. La première concerne la réduction du nombre de capteurs nécessaires au suivi efficace de l'encrassement des avaloirs. L'approche envisagée repose sur l'inférence spatiale, c'est-à-dire la capacité de prédire l'état d'un avaloir en se basant sur des données provenant d'un autre, situé à proximité. Ce concept trouve son utilité dans des situations où, par exemple, les avaloirs d'une même rue présentent des dynamiques d'encrassement corrélées, avec un avaloir se remplissant systématiquement avant un autre. L'idée est d'explorer la possibilité d'utiliser les données d'un avaloir pour inférer ou prédire l'état d'encrassement du second, ce qui pourrait significativement réduire le nombre de capteurs déployés. Une telle avancée permettrait d'une part d'approfondir notre compréhension de la dynamique d'encrassement, et d'autre par permettrait de réduire le nombre de capteur déployés et donc d'optimiser les coûts de maintenance du réseau de capteur.

La deuxième piste concerne la modélisation d'une fréquence de curage optimale. L'objectif est de déterminer une stratégie de maintenance qui maximise ou minimise certains critères spécifiques. Cette approche nécessitera de définir des critères d'efficacité de la maintenance, tels que les coûts, l'impact environnemental, ou la minimisation des nuisances urbaines. Ensuite, il s'agira de modéliser une fréquence de curage adaptée, qui prend en compte ces critères tout en respectant les contraintes opérationnelles et budgétaires. Une telle modélisation aiderait à ordonnancer les

interventions de maintenance de manière plus stratégique et efficace, en s'alignant avec les objectifs globaux du projet de gestion des avaloirs.

Annexes

A Acteurs de la thèse

FIGURE A.1 – Chiffres clefs de Suez en 2022.

Chiffres clefs du SERAMM en 2020.

B Description du capteur

Le capteur utilise un réseau LPWAN pour la transmission des mesures, qui peut être basé sur les technologies LoRa ou Sigfox (utilisée ici). Les passerelles Sigfox sont colocalisées avec des serveurs réseau. Les données collectées sont transférées vers un cloud via le serveur réseau pour être stockées dans une base de données et traitées. Les utilisateurs peuvent accéder aux données via une application web. Plus de détails sur les spécifications du réseau et des capteurs sont fournis ci-dessous [63].

- Précision : ±2 cm
- Portée : 20 cm - 5 m
- Résolution : ±0.5 mm
- Puissance de transission 14 dBm
- Puissance Isotrope Rayonnée Efficace : 16 dBm
- Sensibilité du récepteur : −126 dBm
- Attenne : Interne / Externe (option)
- Fréquence de mesure : 1 à 144 mesures / jour
- Autonomie : jusqu'à 12 ans

C Travaux sur le clustering

C.1 Synthèse des attributs

Les attributs construits peuvent être de différents types, empiriques, inférentiels ou basés sur les excursions. Ils sont décrits ci-dessous. Les **attributs empiriques**

- Amplitude de remplissage de l'avaloir : pour évaluer la taille moyenne des déchets y entrant
  Correspond à la moyenne empirique des variations positives en termes de déchet [8] supérieures à un seuil ζ.
- Fréquence de remplissage : pour évaluer la fréquence à laquelle un avaloir est susceptible de recevoir des déchets
  Correspond à la fréquence relative des variations positives en termes de déchet supérieures à un certain seuil ζ.
- Vitesse moyenne de remplissage
  Correspond à la moyenne des variations positives.

8. on entend par là que les déchets entrent dans l'avaloir

- Influence de la saison
  Chaque variation est associée à la saison correpsondante. Nous comparons les variations moyennes pour chaque saison avec un test ANOVA. La p-value rendue par ce test est utilisée comme attribut.
- Influence des jours de la semaine
  Chaque variation est associée au jour correspondant [9]. Nous comparons les variations moyennes pour chaque jour de la semaine avec un test ANOVA. La p-value rendue par ce test est utilisée comme attribut.
- Influence jours ouvrés / week-end
  Chaque variation est associée au jour correspondant. Nous comparons ensuite les variations moyennes entre les jours ouvrés et les jour de week-end avec un test ANOVA. La p-value rendue par ce test est utilisée comme attribut.
- Percentiles de la distribution des variations positives
  Nous avons retenus les percentiles 25, 50, 75 et 90.

Les **attributs inférentiels** postulent la présence de trois lois différentes une loi gaussienne $\mathcal{G}_V$ reflétant la variabilité intrinsèque du signal ; une deuxième loi gaussienne $\mathcal{G}_D$ pour les variations liées à la dynamique de l'avaloir, comme le remplissage ou le lessivage ; et enfin, une loi uniforme $\mathcal{U}$ pour les variations exceptionnelles ou extrême. Les paramètres liés à cette inférence, obtenus en utilisant un algortihme EM sont les suivant :

- $\mu_{\mathcal{G}_V}$, la moyenne de la variabilité intrinsèque de nos mesures. Nous l'avons fixé à 0.
- $\sigma_{\mathcal{G}_V}$, l'écart-type de la variabilité intrinsèque de nos mesures. Ce paramètre est initialisé à partir d'un tirage aléatoire d'une loi normale de moyenne 1 et d'écart-type 1, noté $\sim \mathcal{N}(1, 1)$.
- $\mu_{\mathcal{G}_D}$, la moyenne des variations liées aux déchets entrant/sortant de l'avaloir. Ce paramètre est initialisé à partir d'un tirage aléatoire d'une loi $\sim \mathcal{N}(4, 1)$.
- $\sigma_{\mathcal{G}_D}$ l'écart-type des variations liées aux déchets entrant/sortant de l'avaloir. Ce paramètre est initialisé à partir d'un tirage aléatoire d'une loi $\sim \mathcal{N}(1, 1)$.
- $a_{\mathcal{U}}$ la variation maximale liée à une diminution de l'encrassement, fixé à $-D_{\max}$
- $b_{\mathcal{U}}$ la variation maximale liée à une augmentation de l'encrassement, fixé à $D_{\max}$
- les probabilités d'appartenance à chacune des lois.

Dans les travaux de clustering, nous utilisons ces paramètres comme attributs, à l'exception des paramètres liés au bruit de mesure (paramètres liés à la loi

9. de sorte que, par exemple, la variation du lundi au mardi soit associée au mardi

gaussienne $\mathcal{G}_V$), ainsi que les variations maximales $a_{\mathcal{U}}$ et $b_{\mathcal{U}}$ car nous voulons regrouper les avaloirs en fonction de leur dynamique, et non de leur profondeur. L'algortihme EM, permettant d'estimer

Les **attributs basés sur les excursions** sont basés les périodes pendant lesquelles les mesures dépassent un seuil prédéfini, ce qui permet de conserver la dimension temporelle des données. Le seuil choisi correspond à 30% de la distance de fond $D_{\max}$.

- ε_N compte le nombre de fois où le niveau minimal d'encrassement de l'avaloir sur une journée dépasse le seuil choisi.
- ε_T évalue la durée maximale en jours consécutifs pendant laquelle le niveau d'encrassement est resté supérieur au seuil choisi.

C.2 Transformation de Yeo-Johnson

La transformation de Yeo-Johnson est une méthode statistique utilisée pour stabiliser la variance et normaliser les distributions de données. Elle est définie différemment pour les valeurs positives et négatives de la variable d'entrée x. Soit x une variable et λ le paramètre de transformation, la transformation de Yeo-Johnson est définie comme suit :

Pour $x \geq 0$, la transformation est :

$$y(x) = \begin{cases} \frac{(x+1)^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0, \\ \log(x+1) & \text{si } \lambda = 0. \end{cases}$$

Pour $x < 0$, la transformation est :

$$y(x) = \begin{cases} -\frac{-(x+1)^{2-\lambda} + 1}{2-\lambda} & \text{si } \lambda \neq 2, \\ -\log(-x+1) & \text{si } \lambda = 2. \end{cases}$$

avec λ un paramètre calibré de manière à maximiser la log-vraisemblance de la distribution résultante, rendant ainsi la distribution des données transformées aussi proche que possible d'une distribution gaussienne.

C.3 Combinaisons d'attributs et hyperparamètres

Les sous-ensembles d'attributs sont établis afin de tester les différentes combinaisons décrites ci-dessous :

- Avec ou sans les attributs correspondant à des p-values (par exemple, l'attribut testant l'influence des jours de la semaine).
- Avec ou sans les attributs basés sur les excursions.

- En choisissant soit un seul, soit deux attributs représentant chacune des familles $\mathcal{F}_1$ et $\mathcal{F}_2$.

À cela, nous ajoutons quelques combinaisons supplémentaires basées sur l'intuition, telles que l'utilisation exclusive d'attributs empiriques ou d'attributs inférentiels basés sur des tests statistiques (comme les p-values). Au total, cela représente environ une douzaine de sous-ensembles d'attributs.

Ensuite, pour chacun de ces sous-ensembles d'attributs, nous testons différentes combinaisons pour effectuer le clustering. Nous expérimentons notamment chacun des trois algorithmes retenus : K-means, Spectral Clustering et DBSCAN. Pour chaque sous-ensemble d'attributs et pour chaque algorithme, nous calibrons les hyperparamètres en utilisant deux scores de clustering : le score Silhouette et l'indice de Calinski-Harabasz. Nous visons à sélectionner deux ou trois hyperparamètres pour chaque configuration. Au final, cela correspond à plus d'environ 70 expériences, chaque expérience correspondant à une combinaison spécifique de sous-ensemble d'attributs, d'algorithme et d'hyperparamètres.

D Éléments contextuels

De façon plus globale, nous avons établi la liste des éléments contextuels susceptibles d'affecter la dynamique d'encrassement des avaloirs à Marseille. Cette liste est basée sur des connaissances terrain et des expertises métier. Ces éléments peuvent être liés à la *structure* de l'avaloir, comme la largeur, la profondeur ou la présence de grilles, par exemple. Ils peuvent également concerner l'environnement immédiat de l'avaloir, tels que la proximité d'arbres ou de fast-foods, qui influencent l'encrassement dans des zones *spatiales* spécifiques. Par ailleurs, les conditions météorologiques, et en particulier les précipitations, sont des facteurs *temporels*[10] pouvant jouer un rôle important. Enfin, des phénomènes *spatio-temporels* sont également pris en compte.
La liste exhaustive des éléments contextuels est détaillée ci-dessous :

- Éléments structurels :
  - dimensions de l'avaloir : profondeur, largeur, longueur, etc.
  - caractéristiques plus spécifiques des avaloirs, comme la présence de barreaudages, de grilles, etc.
  - équipements particuliers comme les paniers, les bavettes (membrane en caoutchouc pour lutter contre les mauvaises odeurs), les clapets (sorte de filtre pour retenir les déchets), etc.
- Éléments spatiaux :

10. En supposant que la pluie soit homogène spatialement.

  - bacs à ordures (ménagères ou tri sélectif) ;
  - arbres ;
  - fast-foods ;
  - commerces ;
  - bars et autres pubs ;
  - type d'espace public (place publique, parking, etc.) ;
  - type d'habitat à proximité (zone résidentielle, commerciale, etc.) ;
  - topographie de la ville ;
  - pente de la route ;
  - état de la route ;
  - arrêts de bus, arrêts de tram ;
  - écoles ;
  - supermarchés.
- Éléments temporels :
  - précipitations (si l'on considère qu'elles sont homogènes spatialement) ;
  - vent.
- Éléments spatio-temporels :
  - marchés ;
  - évènements sportifs et culturels ;
  - mode de nettoyage des rues choisi par la Direction de la Propreté Urbaine (DPU) de Marseille : nettoyage manuel ou automatisé (véhicule de nettoyage dédié) ; nettoyage à sec ou humide (karcher et autres jets d'eau). On peut parfois même observer que lorsque ces nettoyages sont manuels, certains agents poussent les déchets dans les avaloirs.

E Cartographie IRIS

IRIS (*Îlots Regroupés pour l'Information Statistique*) est un système de zonage territorial créé par l'INSEE en France, qui découpe le territoire en environ 16000 zones. Ces unités sont définies en concertation avec les collectivités locales et peuvent être révisées périodiquement pour tenir compte de l'évolution démographique et géographique. Les critères de découpage tiennent compte de la taille de la population (un IRIS compte en général entre 1800 et 5000 habitants), de la cohérence socio-économique et des limites géographiques naturelles ou administratives. Ce système vise à permettre des analyses statistiques fines au niveau local. Un IRIS peut représenter un quartier dans une grande ville ou une petite commune dans son intégralité. Ces unités sont utilisées pour la diffusion de statistiques et sont utiles

pour les décideurs politiques, les urbanistes et les chercheurs. La Figure E.1 illustre le découpage IRIS de l'hypercentre de la ville et ses alentours.

FIGURE E.1 – Exemple du découpage IRIS de l'hypercentre de Marseille et ses alentours.

F Facteurs contextuels

Parmi les éléments contextuels identifiés, nous nous sommes concentrés d'abord sur les éléments structurels et spatiaux. Nous avons collecté et préparé les données afin de construire des facteurs (binaires ou quantitatifs) pour décrire le contexte d'une avaloir.

Facteurs binaires

Les facteurs spatiaux et les rayons d'étude associés sont listés ci-dessous.

Facteur spatial étudié	**Rayon étudié**
arbre	3 m et 10 m
bar/café/pub	10 m et 50 m
fast-food	10 m et 50 m
parking	10 m et 100 m
arrêt de transport en commun	10 m et 50 m
école	10 m et 100 m
marché	100 m
supermarché/hypermarché	10 m et 100 m
parc/jardin	10 m et 100 m
bacs à ordures	5 m, 10 m, 30 m, et 50 m

TABLE 10 – Liste des facteurs contextuels spatiaux.

Parmi les autres facteurs binaires construits on peut retrouver la présence de :

- plus d'une marquise
- d'au moins une grille
- de barreaudages

Enfin, parmi les 101 quartiers (respectivement les 325 IRIS) en tout, on s'intéresse plus précisément aux 25% de quartiers (respectivement IRIS) ayant le plus grand nombre de bars/cafés/pubs (respectivement de fast-foods, d'hôtels, de shops, et de restaurants). Quelques exemples de facteurs créés pour évaluer si l'avaloir se trouve dans ces quartiers (ou IRIS) sont listés ci-dessous.

Facteur créé)	**Description**
Is-Shop-Quarter	Variable binaire qui évalue si l'avaloir se trouve dans un quartier faisant partie des 25% des quartiers comptant le plus grand nombre de *shops*. Plus précisément, cela est équivalent à examiner les quartiers ayant 34 shops ou plus.
Is-Shop-IRIS	Variable binaire qui évalue si l'avaloir se trouve dans un IRIS faisant partie des 25% d'IRIS comptant le plus grand nombre de *shop*. Plus précisément, cela correspond à examiner les IRIS ayant 12 shops ou plus.
Is-Fast-Food-Quarter	Variable binaire qui évalue si l'avaloir se trouve dans un quartier faisant partie des 25% des quartiers comptant le plus grand nombre de fast-foods, cela est équivalent à examiner les quartiers ayant 4 fast-foods ou plus.
Is-Fast-Food-IRIS	Variable binaire qui évalue si l'avaloir se trouve dans un IRIS faisant partie des 25% d'IRIS comptant le plus grand nombre de fast-foods, cela est équivalent à examiner les IRIS ayant 2 fast-foods ou plus.

TABLE 11 – Exemple de facteurs basés sur le nombre d'éléments contextuels dans un quartier ou dans un IRIS.

De manière similaire, on s'interesse également aux 25% quartiers (respectivement IRIS) ayant la plus grande densité de bars/cafés/pubs par km^2 (respectivement de fast-foods, d'hôtels, de shops, et de restaurants). Quelques exemples sont listés ci-dessous.

Élement dont la présence est étudié au sein du quartier (ou IRIS)	**Description**
Is-Shop-Quarter	Variable binaire qui évalue si l'avaloir se trouve dans un quartier faisant partie des 25% des quartiers ayant la plus grande densité de *shops*. Plus précisément, cela est équivalent à examiner les quartiers ayant environ 34 shops par km^2 ou plus.
Is-Shop-IRIS	Variable binaire qui évalue si l'avaloir se trouve dans un IRIS faisant partie des 25% d'IRIS ayant la plus grande densité de *shop*. Plus précisément, cela correspond à examiner les IRIS ayant environ 48 shops par km^2 ou plus.
Is-Fast-Food-Quarter	Variable binaire qui évalue si l'avaloir se trouve dans un quartier faisant partie des 25% des quartiers ayant la plus grande densité de fast-foods, cela est équivalent à examiner les quartiers ayant environ 3 fast-foods par km^2 ou plus.
Is-Fast-Food-IRIS	Variable binaire qui évalue si l'avaloir se trouve dans un IRIS faisant partie des 25% d'IRIS ayant la plus grande densité de fast-foods, cela est équivalent à examiner les IRIS ayant environ 4 fast-foods par km^2 ou plus.

TABLE 12 – Exemple de facteurs basés sur la densité d'éléments contextuelles dans un quartier ou dans un IRIS.

Facteurs quantitatifs

Parmi les facteurs quantitatifs, on peut retrouver :

- la longueur, la largeur, la profondeur, le diamètre de l'exutoire, etc.
- le nombre de bar/café/pub (respectivement fast-food, restaurant, *shop*, hôtel) dans le quartier (respectivement l'IRIS) où se trouve l'avaloir
- la densité de bar/café/pub (respectivement fast-food, restaurant, *shop*, hôtel) dans le quartier (respectivement l'IRIS) où se trouve l'avaloir

G Balanced Accuracy d'un modèle indépendant des données observées

La Balanced Accuracy (BA) est un score utilisé pour évaluer les performances des modèles de classification, particulièrement utile dans les cas où les classes sont déséquilibrées. Elle est calculée comme la moyenne des taux de vrais positifs pour chaque classe. Dans un contexte multiclasse, le True Positive Rate (TPR) pour une classe spécifique correspond au ratio des instances correctement identifiées comme appartenant à cette classe par rapport au nombre total d'instances qui appartiennent réellement à cette classe, indépendamment des autres classes.

Dans un problème de classification à N classes. Pour chaque classe i, la probabilité qu'une instance appartienne réellement à cette classe est notée $P(\text{réel} = i)$. Similairement, la probabilité qu'un modèle indépendant du problème, attribue une instance à la classe i est noté $P(\text{modèle} = i)$.

Comme on suppose que la classe réelle d'une instance et la classe prédite par le modèle sont indépendantes, la probabilité conjointe qu'une instance soit réellement dans la classe i et que le modèle la classe aussi dans i est donnée par

$$P(\text{réel} = i \,\text{et modèle} = i) = P(\text{réel} = i) \cdot P(\text{modèle} = i)$$

Le TPR pour la classe i est le ratio des instances correctement classées par le modèle sur le nombre total d'instances réellement dans la classe i, soit dans la limite d'un nombre infini d'observations :

$$\begin{aligned}
\text{TPR}_i &= \frac{P(\text{réel} = i \,\text{et modèle} = i)}{P(\text{réel} = i)} \\
&= \frac{\cancel{P(\text{réel} = i)} \cdot P(\text{modèle} = i)}{\cancel{P(\text{réel} = i)}} \\
&= P(\text{modèle} = i)
\end{aligned}$$

La Balanced Accuracy (BA) est la moyenne des précisions pour toutes les classes :

$$\begin{aligned}
BA &= \frac{1}{N} \cdot \sum_{i=1}^{N} \text{TPR}_i \\
&= \frac{1}{N} \cdot \sum_{i=1}^{N} P(\text{modèle} = i)
\end{aligned}$$

or comme $P(\cup_{i=1}^{N} \text{modèle} = i) = \sum_{i=1}^{N} \text{TPR}_i = 1$ [11], la BA vaut :

11. Car, pour chaque instance le modèle prédit une unique classe i. Autrement dit, la famille $\{\forall i \in 1, ..., N |$ Le modèle prédit la classe i$\}$ est un système complet d'évènements.

$$BA = \frac{1}{N} \cdot 1 = \frac{1}{N}$$

La Balanced Accuracy d'un modèle de classifiant les instances indépendamment du problème est donc de $\frac{1}{N}$.

Bibliographie

[1] Z. Yazdanfar, A. Sharma, Urban drainage system planning and design–challenges with climate change and urbanization : a review, Water Science and Technology 72 (2) (2015) 165–179. (Cited on page 13.)

[2] L. Atzori, A. Iera, G. Morabito, The internet of things : A survey, Computer networks 54 (15) (2010) 2787–2805. (Cited on page 13.)

[3] D. Uckelmann, M. Harrison, F. Michahelles, Architecting the internet of things, Springer Science & Business Media, 2011. (Cited on page 13.)

[4] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, Y. Portugali, Smart cities of the future, The European Physical Journal Special Topics 214 (2012) 481–518. (Cited on page 13.)

[5] A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, Internet of things for smart cities, IEEE Internet of Things journal 1 (1) (2014) 22–32. (Cited on page 13.)

[6] A. Kulkarni, S. Sathe, Healthcare applications of the internet of things : A review, International Journal of Computer Science and Information Technologies 5 (5) (2014) 6229–6232. (Cited on page 13.)

[7] L. Da Xu, W. He, S. Li, Internet of things in industries : A survey, IEEE Transactions on industrial informatics 10 (4) (2014) 2233–2243. (Cited on page 13.)

[8] S. O. Olatinwo, T.-H. Joubert, Enabling communication networks for water quality monitoring applications : A survey, IEEE Access 7 (2019) 100332–100362. (Cited on page 13.)

[9] N. N. Kasat, P. D. Gawande, A. D. Gawande, Smart city solutions on drainage, unused well and garbage alerting system for human safety, in : International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-19), Nagpur, India, 2019, pp. 1–6. (Cited on page 13.)

[10] R. Rayhana, Y. Jiao, A. Zaji, Z. Liu, Automated vision systems for condition assessment of sewer and water pipelines, IEEE Transactions on Automation Science and Engineering 18 (4) (2021) 1861–1878. (Cited on page 13.)

[11] H. Lee, K. Calvin, D. Dasgupta, G. Krinner, A. Mukherji, P. Thorne, Synthesis report of the ipcc sixth assessment report (ar6), Intergovernmental Panel on Climate Change, Geneva, Switzerland (2023). (Cited on page 13.)

[12] S. Yao, Y. Zhao, A. Zhang, S. Hu, H. Shao, C. Zhang, L. Su, T. Abdelzaher, Deep learning for the Internet of Things, Computer 51 (5) (2018) 32–41. (Cited on page 13.)

[13] S. Helal, F. C. Delicato, C. B. Margi, S. Misra, M. Endler, Challenges and opportunities for data science and machine learning in IoT systems – a timely debate : Part 2, IEEE Internet of Things Magazine 4 (2) (2021) 46–50. (Cited on page 13.)

[14] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, M. S. H. Sunny, Application of big data and machine learning in smart grid, and associated security concerns : A review, IEEE Access 7 (2019) 13960–13988. (Cited on page 13.)

[15] A. C. D. S. Junior, R. Munoz, M. D. L. A. Quezada, A. V. L. Neto, M. M. Hassan, V. H. C. D. Albuquerque, Internet of Water Things : A remote raw water monitoring and control system, IEEE Access 9 (2021) 35790–35800. (Cited on page 13.)

[16] Q. F. Hassan, Internet of Things A to Z : Technologies and Applications, Wiley IEEE Press, 2018. (Cited on page 13.)

[17] D. Zhang, Sewer system control using artificial intelligence, hydraulic model, and Internet of Things, Ph.D. thesis, Norwegian University of Life Sciences (2019). (Cited on page 14.)

[18] U. Raza, P. Kulkarni, M. Sooriyabandara, Low power wide area networks : An overview, IEEE Communications Surveys & Tutorials 19 (2) (2017) 855–873. (Cited on page 27.)

[19] A. Ikpehai, B. Adebisi, K. M. Rabie, K. Anoh, R. E. Ande, M. Hammoudeh, H. Gacanin, U. M. Mbanaso, Low-power wide area network technologies for internet-of-things : A comparative review, IEEE Internet of Things Journal 6 (2) (2019) 2225–2240. (Cited on page 27.)

[20] A. Shakil, M. A. Khalighi, P. Pudlo, C. Leclerc, D. Laplace, F. Hamon, A. Boudonne, Outlier detection in non-stationary time series applied to sewer network monitoring, Internet of Things 21 (2023) 100654. (Cited on page 34.)

[21] M. Braei, S. Wagner, Anomaly detection in univariate time-series : A survey on the state-of-the-art, ArXiv (2020). (Cited on page 42.)

[22] A. Blázquez-García, A. Conde, U. Mori, J. A. Lozano, A review on outlier/anomaly detection in time series data, ACM Comput. Surv. 54 (3) (apr 2021). (Cited on page 42.)

[23] B. Justusson, Median filtering : Statistical properties, in : Two-Dimensional Digital Signal Processing II, Springer, 1981, pp. 161–196. (Cited on page 42.)

[24] R. J. Hyndman, G. Athanasopoulos, Forecasting : principles and practice, OTexts, 2018, available online at https://otexts.com/fpp3/expsmooth.html. (Cited on page 43.)

[25] M. C. Dani, F.-X. Jollois, F. Cassiano, M. Nadif, Adaptive Threshold for Anomaly Detection Using Time Series Segmentation, ICONIP (Nov. 2015). (Cited on page 43.)

[26] W. Wei, Time Series Analysis : Univariate and Multivariate Methods, 2nd edition, 2006, Addison-Wesley, 2006. (Cited on page 43.)

[27] J. Hochenbaum, O. S. Vallis, A. Kejariwal, Automatic anomaly detection in the cloud via statistical learning, CoRR abs/1704.07706 (2017). (Cited on page 43.)

[28] J. Ma, S. Perkins, Time-series novelty detection using one-class support vector machines, in : International Joint Conference on Neural Networks, Vol. 3, 2003, p. 1741–1745. (Cited on page 43.)

[29] Z. Ferdousi, A. Maeda, Unsupervised outlier detection in time series data, in : 22nd International Conference on Data Engineering Workshops (ICDEW), Atlenta, GA, USA, 2006, p. 121. (Cited on page 43.)

[30] K. Peker, Subsequence time series (sts) clustering techniques for meaningful pattern discovery, in : International Conference on Integration of Knowledge Intensive Multi-Agent Systems, Waltham, MA, USA, 2005, pp. 360–365. (Cited on page 43.)

[31] O. I. Provotar, Y. M. Linder, M. M. Veres, Unsupervised anomaly detection in time series using lstm-based autoencoders, in : 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), Kyiv, Ukraine, 2019, pp. 513–517. (Cited on page 43.)

[32] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, 2001. (Cited on pages 44, 108, 109 et 115.)

[33] P. Mahalanobis, On the generalized distance in statistics, in : Proceedings National Institute of Science of India, Vol. 49, 1936, pp. 234–256. (Cited on page 48.)

[34] B. Wang, W. Shi, Z. Miao, Confidence analysis of standard deviational ellipse and its extension into higher dimensional euclidean space, PLoS ONE 10 (3) (2015) 60–67. (Cited on page 49.)

[35] M. Nixon, A. Aguado, Feature Extraction and Image Processing for Computer Vision, Elsevier Science, 2019. (Cited on page 55.)

[36] T. Acharya, A. Ray, Image Processing : Principles and Applications, Wiley, 2005. (Cited on page 55.)

[37] R. Szeliski, Computer Vision : Algorithms and Applications, Texts in Computer Science, Springer International Publishing, 2022. (Cited on page 55.)

[38] J. Canny, A computational approach to edge detection, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8 (6) (1986) 679–698. (Cited on page 56.)

[39] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Transactions on Information Theory 8 (2) (1962) 179–187. (Cited on page 56.)

[40] R. H. Shumway, D. S. Stoffer, Time Series Analysis and Its Applications, Springer, 2000. (Cited on page 56.)

[41] A. Nielsen, Practical Time Series Analysis : Prediction with Statistics and Machine Learning, O'Reilly Media, 2019. (Cited on page 56.)

[42] O. J. Dunn, Confidence intervals for the means of dependent, normally distributed variables, Journal of the American Statistical Association 54 (287) (1959) 613–621. (Cited on page 64.)

[43] S. Holm, A simple sequentially rejective multiple test procedure, Scandinavian journal of statistics (1979) 65–70. (Cited on page 64.)

[44] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate : a practical and powerful approach to multiple testing, Journal of the Royal statistical society : series B (Methodological) 57 (1) (1995) 289–300. (Cited on page 65.)

[45] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the royal statistical society : series B (methodological) 39 (1) (1977) 1–22. (Cited on page 66.)

[46] B. Butcher, B. J. Smith, Feature Engineering and Selection : A Practical Approach for Predictive Models, Taylor & Francis, 2020. (Cited on page 71.)

[47] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning : A new perspective, Neurocomputing 300 (2018) 70–79. (Cited on page 71.)

[48] I.-K. Yeo, R. A. Johnson, A new family of power transformations to improve normality or symmetry, Biometrika 87 (4) (2000) 954–959. (Cited on page 73.)

[49] F. Pedregosa, G. Varoquaux, Gramfort, et al., Scikit-learn : Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830. (Cited on pages 75 et 107.)

[50] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Transactions on neural networks 16 (3) (2005) 645–678. (Cited on page 75.)

[51] D. Xu, Y. Tian, A comprehensive survey of clustering algorithms, Annals of Data Science 2 (2015) 165–193. (Cited on page 76.)

[52] A. E. Ezugwu, A. M. Ikotun, et al., A comprehensive survey of clustering algorithms : State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects, Engineering Applications of Artificial Intelligence 110 (2022) 104743. (Cited on page 76.)

[53] C. Aggarwal Charu, K. Reddy Chandan, Data clustering : algorithms and applications, CRC press, 2013. (Cited on page 76.)

[54] S. Landau, M. Leese, D. Stahl, B. Everitt, Cluster Analysis, Wiley Series in Probability and Statistics, Wiley, 2011. (Cited on page 76.)

[55] L. Hubert, P. Arabie, Comparing partitions, Journal of classification 2 (1985) 193–218. (Cited on page 78.)

[56] N. X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison : is a correction for chance necessary ?, in : Proceedings of the 26th annual international conference on machine learning, 2009, pp. 1073–1080. (Cited on page 78.)

[57] J. Cohen, A coefficient of agreement for nominal scales, Educational and psychological measurement 20 (1) (1960) 37–46. (Cited on page 78.)

[58] W. M. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical association 66 (336) (1971) 846–850. (Cited on page 78.)

[59] E. B. Fowlkes, C. L. Mallows, A method for comparing two hierarchical clusterings, Journal of the American statistical association 78 (383) (1983) 553–569. (Cited on page 78.)

[60] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, biometrics (1977) 159–174. (Cited on page 79.)

[61] OpenStreetMap contributors, Planet dump retrieved from https ://planet.osm.org , https://www.openstreetmap.org (2017). (Cited on page 91.)

[62] INSEE, Découpage iris, https://www.insee.fr/fr/metadonnees/definition/c1523, accessed : 2023-09-20 (2016). (Cited on page 92.)

[63] Greencityzen, Hummbox level ultrasonic : Technical specifications (2022). (Cited on pages 103 et 127.)

[64] R. J. Little, D. B. Rubin, Statistical analysis with missing data, Vol. 793, John Wiley & Sons, 2019. (Cited on page 106.)

[65] P. Refaeilzadeh, L. Tang, H. Liu, Cross-validation, Encyclopedia of database systems (2009) 532–538. (Cited on page 110.)

[66] L. Breiman, Classification and regression trees, Routledge, 2017. (Cited on page 115.)