



**HAL**  
open science

# Validation of wastewater data using artificial intelligence tools and the evaluation of their performance regarding annotator agreement

Imane Zidaoui, Cédric Wemmert, Matthieu Dufresne, Claude Joannis, Sandra Isel, Jonathan Wertel, José Vazquez

## ► To cite this version:

Imane Zidaoui, Cédric Wemmert, Matthieu Dufresne, Claude Joannis, Sandra Isel, et al.. Validation of wastewater data using artificial intelligence tools and the evaluation of their performance regarding annotator agreement. *Water Science and Technology*, inPress, 87 (12), 10.2166/wst.2023.174. hal-04116074

**HAL Id: hal-04116074**




**<https://hal.science/hal-04116074>**

Submitted on 28 Jun 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Validation of wastewater data using artificial intelligence tools and the evaluation of their performance regarding annotator agreement

Imane Zidaoui <sup>a,b,\*</sup>, Cédric Wemmert <sup>c</sup>, Matthieu Dufresne<sup>b</sup>, Claude Joannis<sup>d</sup>, Sandra Isel<sup>b</sup>, Jonathan Wertel<sup>b</sup> and José Vazquez <sup>a</sup>

<sup>a</sup> Department of Fluid Mechanics, ICube Laboratory, 2 Rue Boussingault, Strasbourg 67000, France

<sup>b</sup> 3D EAU, 3 Quai Kléber, Strasbourg 67000, France

<sup>c</sup> Data Science and Knowledge Department, ICube Laboratory, 300 Bd Sébastien Brant, Illkirch-Graffenstaden 67400, France

<sup>d</sup> CJ Conseil, 37 Rue du Coteau, Nantes 44100, France

\*Corresponding author. E-mail: imane.zidaoui@3deau.fr

 IZ, 0000-0002-9667-7061; CW, 0000-0002-4360-4918; JV, 0000-0001-9065-8559

### ABSTRACT

To prevent the pollution of water resources, the measurement and the limitation of wastewater discharges are required. Despite the progress in the field of data acquisition systems, sensors are subject to malfunctions that can bias the evaluation of the pollution flow. It is therefore essential to identify potential anomalies in the data before any use. The objective of this work is to deploy artificial intelligence tools to automate the data validation and to assess the added value of this approach in assisting the validation performed by an operator. To do so, we compare two state-of-the-art anomaly detection algorithms on turbidity data in a sewer network. On the one hand, we conclude that the *One-class SVM* model is not adapted to the nature of the studied data which is heterogeneous and noisy. The *Matrix Profile* model, on the other hand, provides promising results with a majority of anomalies detected and a relatively limited number of false positives. By comparing these results to the expert validation, it turns out that the use of the *Matrix Profile* model objectifies and accelerates the validation task while maintaining the same level of performance compared to the annotator agreement rate between two experts.

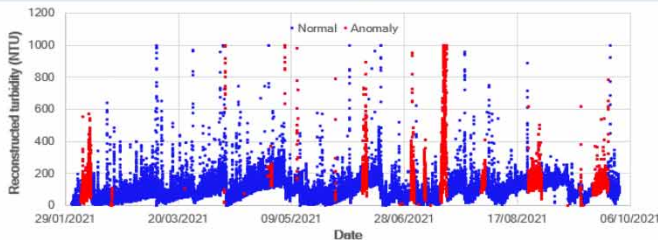
**Key words:** annotator agreement, artificial intelligence, data validation, matrix profile, one-class SVM, wastewater

### HIGHLIGHTS

- The subjective nature of manual validation of wastewater data limits the agreement rate between different experts.
- Artificial intelligence approaches based on binary clusterings, such as OC-SVM, are not adapted to anomaly detection in turbidity data.
- Matrix profile is being evaluated on wastewater data for the first time and shows promising results.

## GRAPHICAL ABSTRACT

## Validation of wastewater data using artificial intelligence tools and the evaluation of their performance regarding annotator agreement



Despite the significant advancements in data acquisition technology, it is still essential to identify anomalies in the massive amount of data collected through IoT sensors to ensure effective wastewater management.



An operator may spend up to two hours validating a month's chronicle from a single sensor, indicating the significant amount of time and effort required for manual validation



Manual validation is subjective due to validation bias and variations in expert judgment, which can cause discrepancies between different experts and/or the same expert at different times.



Despite their good feedback, some AI models, such as *OC-SVM*, may not be suitable for detecting anomalies in turbidity data due to its variability based on meteorological conditions



The *Matrix Profile* model enables faster and more objective validation without compromising accuracy, as confirmed by its performance compared to the agreement rate between two experts.

Authors: I. Zidaoui, C. Wemmert, M. Dufresne, C. Joannis, S. Isel, J. Wertel and J. Vazquez

### 1. INTRODUCTION

To limit the pollution of natural resources, the stakeholders in the water sector must be mobilized for pragmatic management of wastewater facilities and rigorous control of wastewater discharges. Hence, it is important to know the functioning and structural state of the wastewater system and to prevent or identify malfunctions in order to orient and implement an operating and/or investment programme that responds to a logic of continuous improvement and sustainable management (van Daal *et al.* 2017). Overflow monitoring allows sustainable management based on continuous metrology (flow measurements) or discontinuous metrology with periodicity (sampling to establish pollution balances, etc.). Information is thus obtained on the direct discharges of the wastewater system during rainy events. Considering the number of equipped points, a large and diversified park of sensors should be operated and maintained. The plethora of sensors in place with a high sampling frequency and acquisition systems with innovative and wireless data transfer technologies leads to the integration of wastewater systems in a smart city perspective (Dogo *et al.* 2019b). The goal is therefore to optimize the strategies for processing and exploiting the massive amount of measurements made within the framework of the Internet of Things (IoT) in order to get the most out of them. This approach is part of a quality process whose aim is to ensure that the results obtained through the instrumentation of the various facilities are appropriate.

Thus, the reliability of the data is essential, and its validation is indispensable. Despite the progress made in the field of sensors and acquisition systems, the charged and aggressive nature of wastewater leads to problems of inconsistency or loss of data, which invalidates some of the measures. Possible failures include clogging and fouling of the probe, displacement of the sensor by the flow, transmission errors, and others. These problems affect the quality of the stored data. They impact the recorded measurement and can occur in different aspects: missing data, measurement blockage or saturation, sensor drift, noisy signal, or others (Methnani 2012). In order to ensure the quality of the measured data, they must be systematically validated.

In the field of artificial intelligence (AI), data validation refers to fault or anomaly detection. It is a discipline of data mining that is widely studied and deployed in different domains, such as intrusion detection, fraud detection, health monitoring, and industry manufacturing. It consists in identifying elements or events that are significantly different from usual data.

Therefore, the objectives of this study are manifold:

- discuss the use of an automatic AI-based validation system for large sewer measurement data sets;
- provide a comparison between two state-of-the-art algorithms;

- highlight the shortcomings and limitations of the 'ground truth' validation;
- recommend future research directions in this field.

This paper is organized as follows. A review of relevant work from the literature is discussed in Section 2. Sections 3 and 4 describe our proposed methodology for the validation of wastewater quality data and the different used approaches. Experiments and results are discussed in Section 5, and Section 6 presents the study's conclusion and perspectives for future work.

## 2. RELATED WORK

Invalid data is a straightforward consequence of system failures and harsh operating conditions. To address the problem at the source, the measuring device must be adapted to the measurement range and placed in a position that allows accurate measurements without significant disturbances (Campisano *et al.* 2013). An acquisition system capable of functioning properly under different weather conditions is also necessary. In addition, a preventive maintenance procedure and self-cleaning systems are required whenever possible. Nevertheless, experiences have shown that despite all these routines and efforts to optimize the reliability of sensors in sewerage networks, they remain insufficient, failures are inevitable and consequently corrupted data as well (Saber *et al.* 2015).

Several 'trivial' anomalies such as sampling problems, out of range, missing data, sensor blockage, or saturation can be detected via simple parametric tests. These tests can be implemented on specific supervision software or spreadsheets (Van Bijnen & Korving 2008). This (pre-)validation can be completed by statistical techniques or physically based models allowing to create and exploit a virtual redundancy (Piatyszek *et al.* 2000; Alferes *et al.* 2013). Most of these statistical models rely on hypotheses made on the characterization of the data and assume a stationary process. This assumption is far from being valid in sewerage networks which are highly dynamic on a daily and seasonal basis (Mourad & Bertrand-Krajewski 2002).

Thus, the pre-validation process is often completed by an expert validation. The latter requires the intervention of an operator to evaluate the overall likelihood of the results obtained (Mourad & Bertrand-Krajewski 2002; Branislavljević *et al.* 2010; Therrien *et al.* 2020). This manual procedure is mainly based on numerical models or visualization tools. The use of numerical simulation models requires the presence of an explicit model that allows to statistically evaluate the deviation between the measurement and the expected value based on the model-driven behaviour of the system (Venkatasubramanian *et al.* 2003). Such models are used for the validation of hydrometric data using the network's 1D models, but their calibration is difficult for pollution data. Moreover, these models are time-consuming and require a thorough knowledge of the operating conditions of the network and the correlation between these different parameters. In contrast, visualization-based validation requires specific validation criteria that rely on prior knowledge of the measurand dynamics and available exogenous information. Mourad & Bertrand-Krajewski (2002) formulate and evaluate seven criteria for the validation of hydrometric rainfall intensity and flow data. Zidaoui *et al.* (2022) provide validation procedures for turbidity data with material redundancy. This approach rapidly reaches its limits, especially with the increase in database volume to be processed. This task is also daunting for the experts involved, given the time it takes and the repetitive process. A reorientation of the workforce towards engineering tasks where its added value is valuable is more judicious. Moreover, this approach is also subject to human error and subjectivity which are difficult to overcome (Mourad & Bertrand-Krajewski 2002). This subjectivity will be discussed and evaluated later in this work (cf. Section 3.3).

All these arguments plead for the investigation of new approaches that allow to optimize the process of data validation in sewerage networks by automating this task and reducing the time needed.

AI has already proven its efficiency for data validation in different disciplines. Consequently, it is naturally one of the levers for action and one of the paths to be investigated. There are several examples of automatic data validation based on AI in the field of hydrology, particularly for the assessment of the quality of rivers and drinking water (Zhang *et al.* 2017; Muharemi *et al.* 2019). Support vector machine (SVM), logistic regression (LR), and artificial neural networks (ANN) are some of the most commonly used AI approaches for anomaly detection in water quality data (Dogo *et al.* 2019a). *One-Class Support Vector Machine* (OC-SVM) is a reviewed version of the SVM algorithm, adapted to the imbalanced database where the anomalies are a minority. This algorithm was tested on data from a drinking water treatment testbed with 51 measured parameters (Inoue *et al.* 2017). The results are encouraging, especially for their limited computational time. Furthermore, the latest research in the field of temporal data mining has led to the development of the *Matrix Profile* model which is deemed to be 'the state-of-the-art anomaly detection technique for continuous time series' (Ferebans *et al.* 2020). This

model has so far never been evaluated for wastewater quality data. Nevertheless, being domain agnostic and given that its results in other applications have proven its performance and its ability to identify anomalies, it seems important to test it in our context.

### 3. METHODS

To summarize, this work aims to investigate AI approaches and evaluate their ability to optimize the data validation process. The evaluation will focus on two algorithms that result from the state-of-the-art analysis, namely *One-class SVM* and *Matrix Profile*. Their performance is compared to the results of the expert validation applied to quality data from a wastewater network.

#### 3.1. Proposed algorithms

##### 3.1.1. One-class support vector machine

OC-SVM was proposed by Schölkopf *et al.* (2001) for the detection of anomalies by considering that the majority of the training data are in a first class while the anomalies form a second class. Zhang *et al.* (2008) adapted this model to temporal data and evaluated its performance on telecommunication network data.

The basic idea of the OC-SVM algorithm is to project the input data into a characteristic space of higher dimension than the input data using an appropriate kernel function (see Figure 1(a)) and then build a decision function to best separate the two classes by maximizing the distance between them (see Figure 1(b)). Two hyperparameters must be defined in order to run the algorithm: the parameter  $\mu$  is the fraction between points inside the established boundary (valid data) and those outside the boundary (anomalies) (see Figure 1(c)). The parameter  $\sigma$  controls the nonlinearity of the decision function.

##### 3.1.2. Matrix profile

*Matrix Profile* is a relatively new algorithm for time series analysis, introduced in 2016 by Eamonn Keogh (University of California Riverside) and Abdullah Mueen (University of New Mexico) (Yeh *et al.* 2016). Its principle is to compare subsequences of the time series to itself by computing the Euclidean distance between each pair of subsequences of a given length. A naive way would be to compute the complete distance matrix of all pairs of subsequences in a time series and then to evaluate the minimum value of each column. However, this approach is not feasible due to the computational complexity and storage capacity. Hence, the interest of Matrix Profile algorithm which optimizes these calculations and adapts the method to large databases.

The recognition of anomalies via *Matrix Profile* is based on a vector of minimum Euclidean distances, called the matrix profile, between each pair of subsequences in a time series (see Figure 2). The subsequences with the largest distances correspond to anomalies. The Matrix Profile algorithm thus depends on two hyperparameters: the length of the subsequence  $w$  and the number (or ratio) of outliers  $k$ .

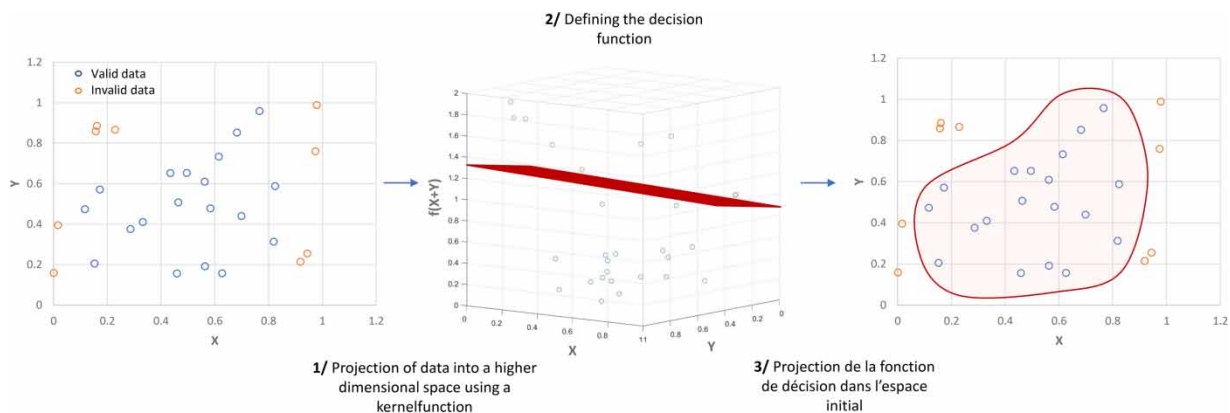
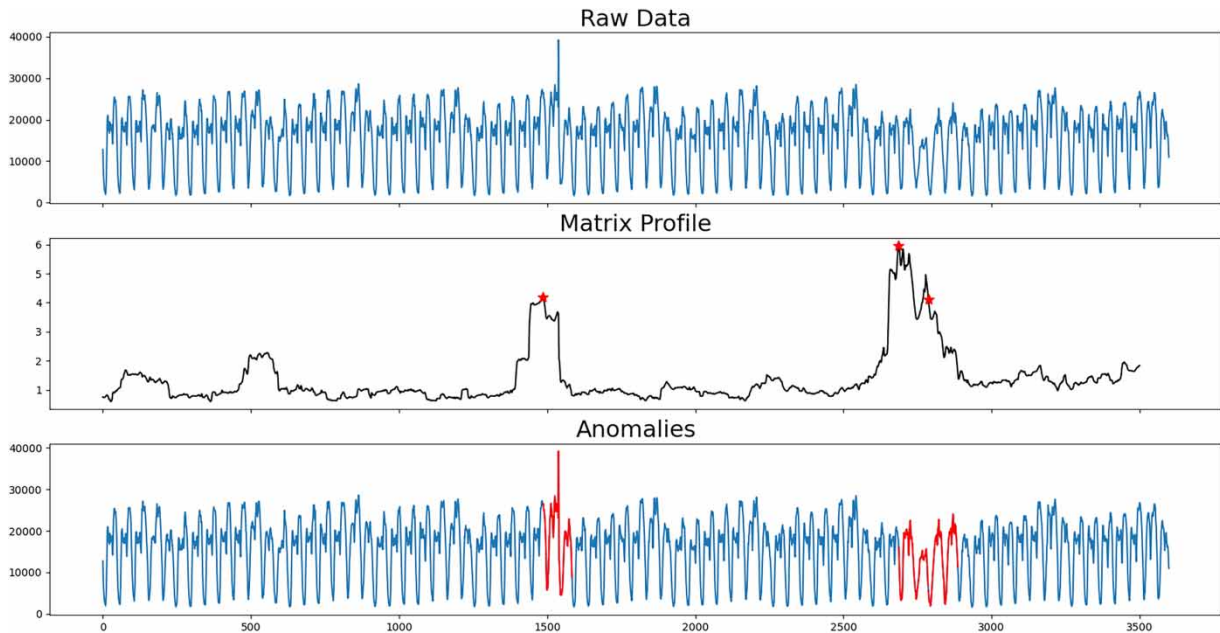


Figure 1 | OC-SVM principle.



**Figure 2** | Matrix Profile principle. Large distances are anomalous events. Please refer to the online version of this paper to see this figure in colour: <https://dx.doi.org/10.2166/wst.2023.174>.

### 3.2. Case study: wastewater quality database

In order to evaluate the AI algorithms, an urban wastewater network database was used. This database comes from a city in France of about 46,000 PE. For confidentiality reasons, the site is anonymized. The instrumentation operation consisted in equipping six overflow chambers since February 2021 with continuous measurement probes for turbidity and conductivity.

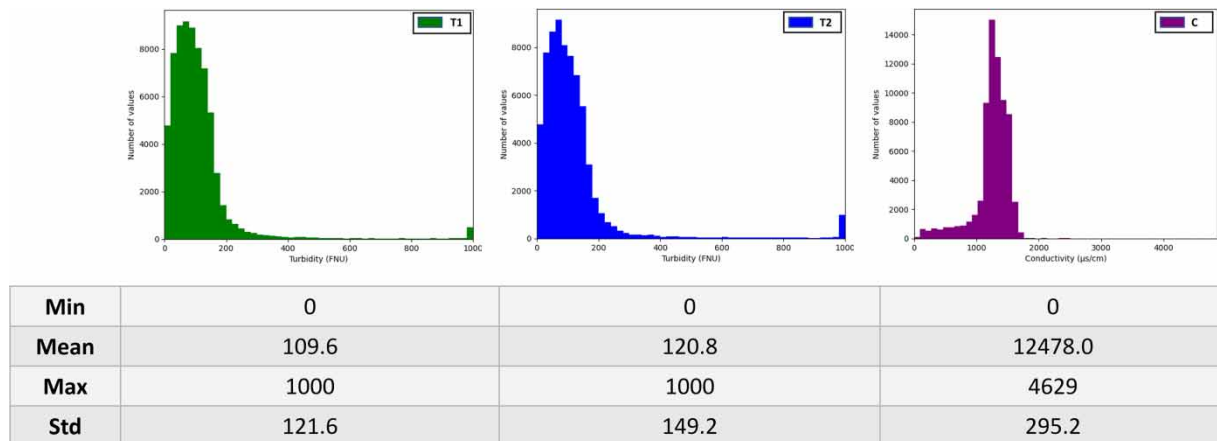
In general, the turbidity signal shows a significant dynamic. This dynamic corresponds to a real variability of the effluent quality. Moreover, the signal can also be disturbed by artefacts, for example abnormal peaks, which can lead to an overestimation of the average values. Thus, physical turbidity redundancy was implemented. Redundancy here means the installation of two identical sensors in close proximity. This makes the measurement more reliable by detecting inconsistencies and, if necessary, identifying the suspect sensor. On the other hand, conductivity probes are sufficiently reliable that no sensor redundancy is proposed. A probe is installed in the chamber of each of the six interceptors. For operational constraints, a value is registered every 5 min. This is an average of the values measured every 20 s (Figure 3).

The main objective of the operation is to progressively achieve the most reliable evaluation of the rejected flows by systematically validating the collected data which constitute the database of this work with the help of an automatic pre-validation combined with a manual expert validation.

### 3.3. Ground truth

The evaluation of the performance of different AI algorithms and their efficiency in identifying anomalies implies the presence of a reference validation. This is referred to as ground truth. For this, domain experts are contracted to validate the data manually. This validation consists in visualizing the data acquired by the different probes while varying the analysis scale from a few hours to several weeks. Multivariate analyses can also be performed in order to compare the different measurements with each other and to evaluate their coherence and their representativeness of the network functioning. Validating a month's chronicle of data can take up to 2 h full time for a trained expert.

This manual approach takes advantage of human skills in recognizing characteristic patterns and remembering past faults. However, it also inherits its shortcomings which are the subjectivity of the operator and human error. To be completely free of subjectivity, the experiment should be conducted under fully controlled conditions, where defects can be accurately tracked. However, *Versini et al. (2015)* show that the behaviour of turbidity data in experimental laboratory circumstances is different from that of a sewer network. According to *Wu & Keogh (2020)*, the use of synthetic data represents one of the flaws in the evaluation of anomaly detection models. The other alternative is to use the results of the validation by the expert as the best



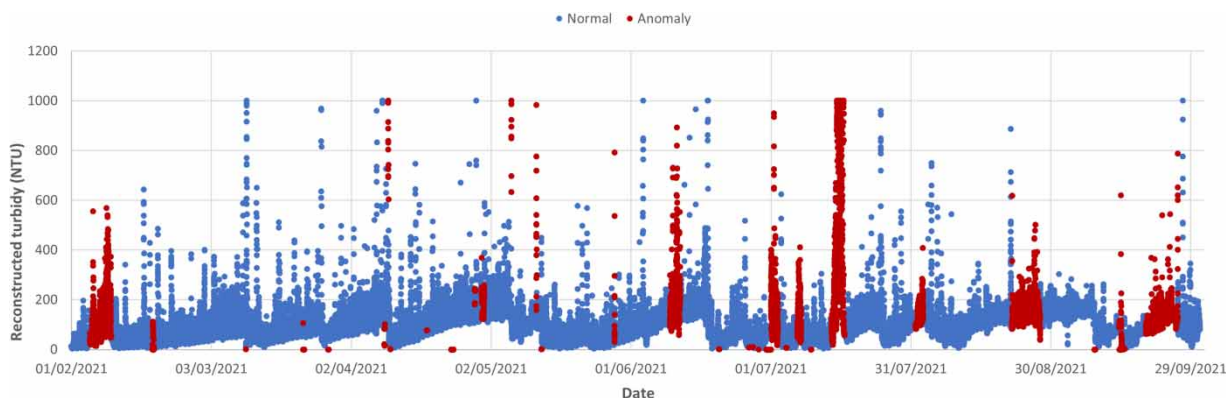
**Figure 3** | Exploratory analysis of the measured data: the two turbidities (T1 in green and T2 in blue) and the conductivity (C in purple). Please refer to the online version of this paper to see this figure in colour: <https://dx.doi.org/10.2166/wst.2023.174>.

available approximation of the ground truth. Hence, in order to evaluate the sensitivity of the ground truth to annotator's subjectivity, a comparison of the validation performed by two operators is conducted. Both validators are used to dealing with pollution data and are familiar with the dynamics of sewerage networks. A pre-validation tool is set up in common for both annotators. This tool selects suspicious periods (Zidaoui *et al.* 2022). Then, the experts' mission consists in, independently, attributing a binary label (valid/invalid) to each period, with the possibility of segmenting it or extending it if exogenous elements justify so. Figure 4 represents the validation results of the turbidity data such as operated by one of the experts.

### 3.4. Performance metrics

Choosing the right metrics is a crucial step in evaluating the performance of a machine learning model. Often, a single metric is not enough to truly judge a model. The most classical is accuracy. However, it makes no difference between correctly classified valid and invalid data. In our problem, data are highly unbalanced, meaning that anomalies are a minority which leads to high accuracies even if invalid data is not identified.

Therefore, metrics suitable for binary classification of unbalanced databases (anomalies being minor) are used. The confusion matrix compares the results of the algorithm (predicted class) to the reference (true class). A positive label denotes an anomaly (invalid data). The standard template for binary confusion matrix is shown in Table 1 as a  $2 \times 2$  matrix with four types of results.



**Figure 4** | Data validation results operated by an expert (ground truth).

**Table 1** | Confusion matrix definitions

		Model	
		Predicted Positive (PP)	Predicted Negative (PN)
Reference	Positive (P)	True Positive TP	False Positive FP
	Negative (N)	False Negative FN	True Negative TN

From the confusion matrix, different metrics can be calculated:

- Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall (sensitivity) is the ratio of correctly predicted positive observations to all observations in actual class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- *F1* score is the weighted average of precision and recall. Therefore, this score takes both false positives and false negatives into account.

$$F1 \text{ score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

## 4. EXPERIMENTS

### 4.1. Data pre-processing

In real-life scenarios, time series coming from sensors are particularly subject to noise and outliers. The aim of anomaly detection models is to identify outliers. However, time series data requires some groundwork before it can be mapped by machine learning algorithms. The crucial issues are temporal irregularity and missing data. When it comes to time series analysis, it is important to have a constant timestamp. Resampling is then an important technique to ensure a round-the-clock frequency. Consequently, before any modelling, time series are aligned to get fixed timestamps, here a frequency of 5 min is chosen, which corresponds to the pre-set frequency for data acquisition.

Moreover, most anomaly detection algorithms are sensitive to the presence of missing data. Time series models require that there are no gaps in data along the time index. Consequently, missing data need to be replaced with judiciously chosen values before fitting a model. The replacement of values in this context is known as ‘imputation’. There are multiple imputation algorithms in the literature. The aim here is to replace the gaps so that the algorithms that are sensitive to their presence can run but still identify it as an anomaly later on. Replacement with zeros is then chosen.

As mentioned before, the original OC-SVM model was not designed for time series. Thus, a pre-processing of the data must take place. This consists in segmenting the data into sequences of predefined length  $L$  and feeding the model with subsequences instead of single points. This operation adds a parameter to the OC-SVM model which is the size of the sequence to be injected. *Matrix Profile* performs this operation internally without the need for pre-processing since the window size is already one of its parameters.

### 4.2. Model fitting

Model fitting is used to identify the optimal parameters of the model which is used for anomaly detection. An optimization procedure consists in defining a search space with the possible values for each hyperparameter. Different optimization algorithms can be used to explore this space. Here, we use a grid search. A finite combination list of values for each hyperparameter is specified, and all combinations are evaluated to identify the optimal combination (see Table 2).



**Table 2** | Grid search for model fitting

	Parameters	Grid Search	N° of combinations
OC-SVM	$\sigma$ : linearity coefficient $\mu$ : contamination rate $L$ : subsequence length	(start = 0.005, stop = 10, step = 0.2) (start = 5%, stop = 25%, step = 0.2%) [30 min, 1 h, 3 h, 12 h, 24 h]	2,500
Matrix Profile	$w$ : window length $k$ : anomaly ratio	(start = 2, stop = 72, step = 2 hours) (start = 5%, stop = 20%, step = 0.5%)	1,050

The algorithms are all implemented in Python. The processing time depends on the hardware and the processing optimizations. The training and testing of the models were performed on the hardware listed in [Table 3](#).

## 5. RESULTS AND DISCUSSION

### 5.1. Annotator agreement

The objective of this section is to quantify the validation variance between two experts and their agreement rates. To do so, we computed the lower bound on the mean classification error rate relative to the ‘true’ labels for binary classification and two labellers according to the following equation ([Smyth 1996](#)):

$$\bar{e} \geq \frac{n_l}{2N}$$

where  $\bar{e}$  is the mean classification error rate,  $N$  is the length of the database, and  $n_l$  is the number of items where the two experts disagree. If they disagree on all items, their mean error rate will be of 0.5, whereas it is of 0 if they agree on all the elements. According to [Smyth \(1996\)](#), if the lower bound on  $\bar{e}$  is greater than 10%, the validation is therefore inaccurate, and the quality of the expert labelling process must be reconsidered. For the case study database, Smyth’s lower error bound was found to be  $\bar{e} \geq 3.4\%$ . We concluded that the results of the expert validation were acceptable.

In addition, to evaluate the inter-variability among experts, we calculated the pairwise metrics (Section 3.4) between labellers in order to take into account the imbalanced nature of the database ([Lampert et al. 2016](#)). [Table 4](#) synthesizes the results. The overall  $F1$  score was 0.658.

The imbalance between precision and recall and the low score of the former showed that one expert was more likely to invalidate than the other. It is a consequence of the so-called validation bias, which is the tendency of an annotator to prefer one decision over another when there is doubt. Another cause of discrepancy is the random variation of an expert in relation to the other or even regarding himself. As a matter of fact, a revalidation of the same data by the same expert later does not lead to exactly the same result.

Moreover, great variability was observed between different measure sites (see [Table 5](#)).

In order to interpret these scores, the results of the validation of the two sites are shown in [Figure 5](#). We observed that for site 2, the anomalies were more obvious and therefore the disagreement between the two experts concerned more likely the delimitation of the defect. As for site 1, the two experts agreed on very few anomalies, 27% abnormal timesteps only.

**Table 3** | Hardware specification

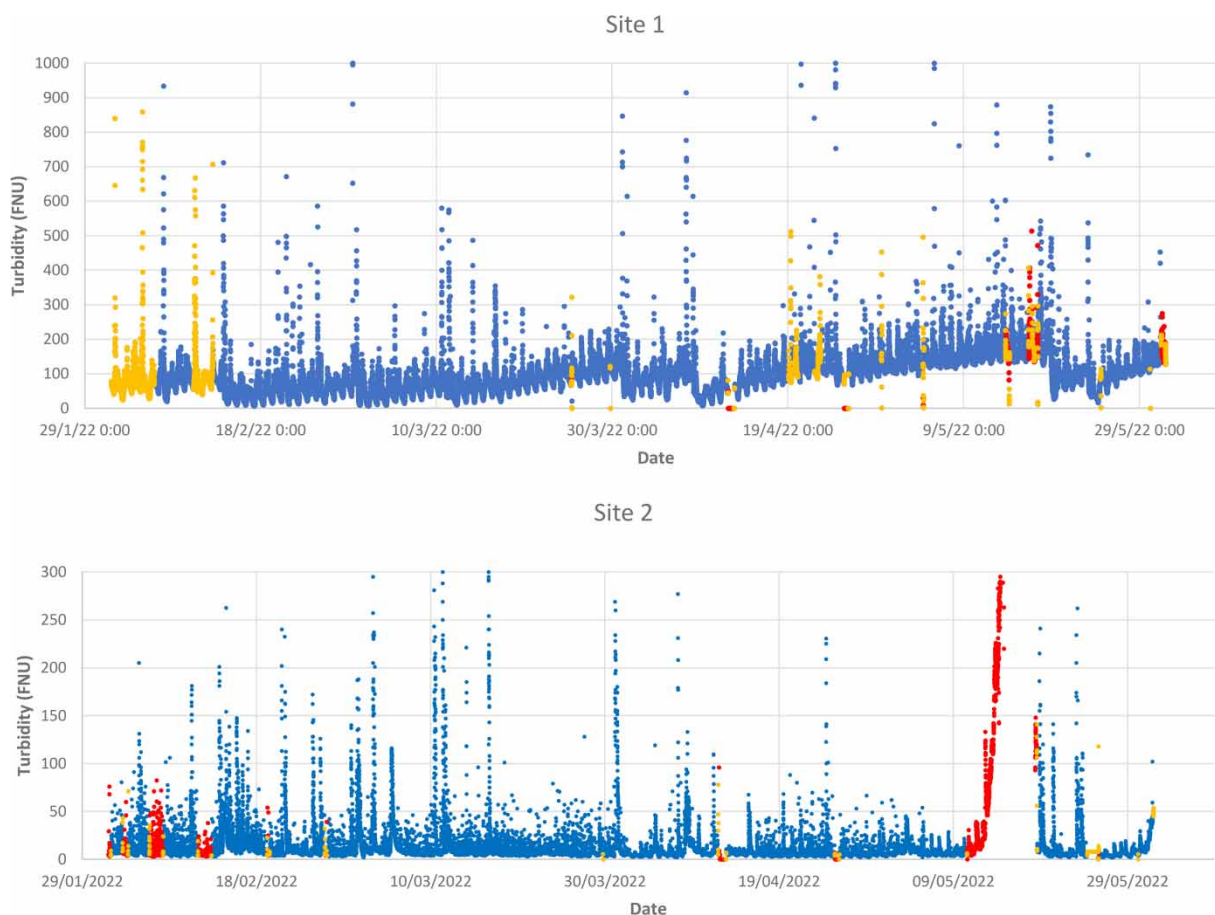
CPU model	AMD EPYC 7502P 32-Core Processor
RAM	16.4 GB
GPU	NVIDIA GeForce RTX 3090, 24 GB

**Table 4** | Overall pairwise performance metrics between labellers

Precision	Recall	F1 score
0.514	0.913	0.658

**Table 5** | Difference of performance metrics between labellers and for different datasets

	Precision	Recall	F1 score
Site 1	0.274	0.892	0.420
Site 2	0.697	0.891	0.927



**Figure 5** | Validation of turbidity data for two different sites and by two different labellers. Colour describes the level of agreement: blue for normal data, red for common defects and orange for unique anomalies (identified by one of the annotators only). Please refer to the online version of this paper to see this figure in colour: <https://dx.doi.org/10.2166/wst.2023.174>.

For the evaluation of the performance of machine learning models, we considered the consensus between the two experts, in such a way as to keep only the most accurate defects and to avoid introducing a human bias. For further use, this approach would impose a permanent double validation to tune the AI model, which is very costly.

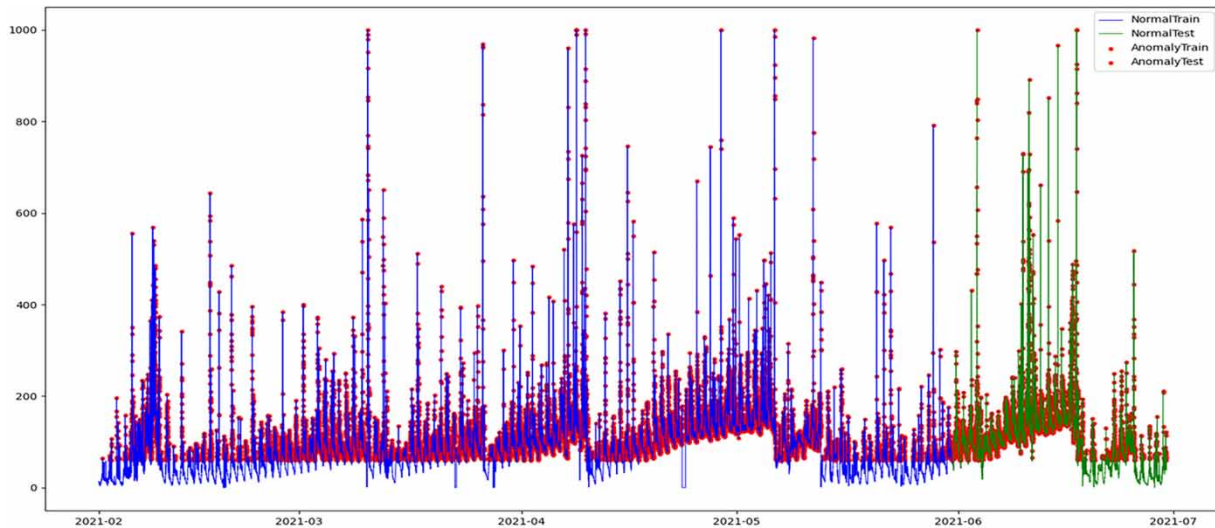
In the case of an unsupervised approach (which is the case for the two proposed models), this bias must be considered when interpreting the performance results of the model. By contrast, the problem is more critical in the case of a supervised approach, since the model will embed the bias introduced by human subjectivity. Indeed, supervised models use labelled data as input for learning and adapt their parameters in order to obtain the same results as output, whereas unsupervised models deduce functions to describe the internal structures from unclassified data. Thus, in view of the invalidity of a gold standard validation free of subjectivity, it is more appropriate to go for unsupervised approaches.

## 5.2. Models' performance

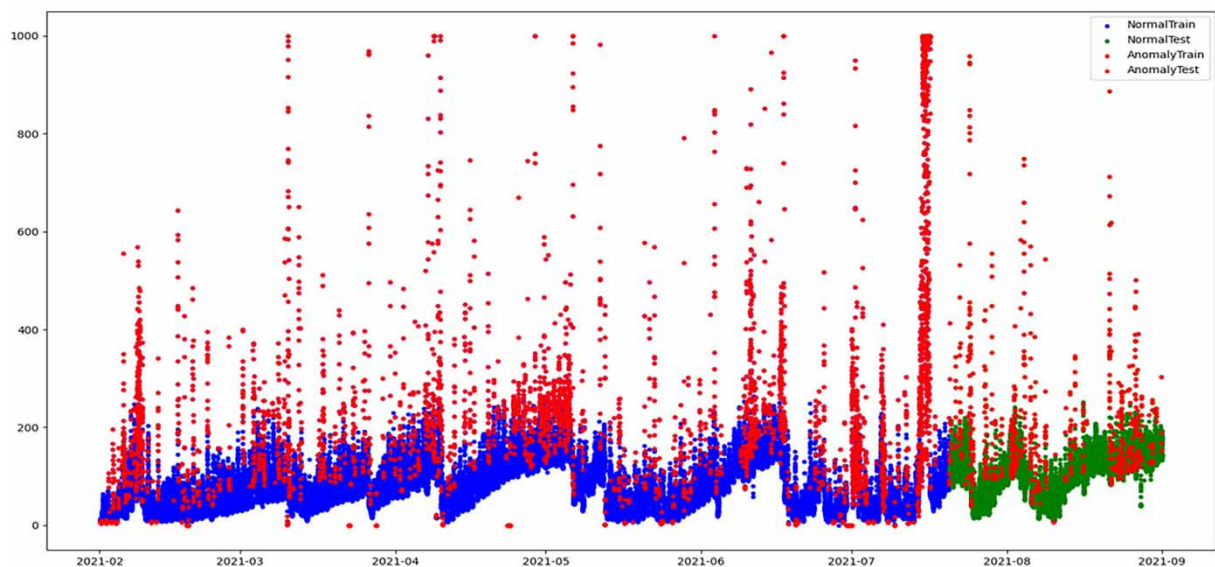
The principle of *OC-SVM* is to define an empirical relation which describes the normal data and consequently any element which does not meet this property is considered as abnormal. However, if we inject the whole data set into the model, there is

a risk of over-fitting the parameters. Thus, the data set was divided into two groups. Training data for the adjustment of the different parameters and test data which allowed us to evaluate the generalization of the model to new data. Figure 6 shows the results of applying the OC-SVM model to turbidity data without segmentation. It can be seen that the classification of normal and abnormal data was based on simple thresholding (linear function). Indeed, OC-SVM did not manage to deal with the temporal dimension, hence the interest in data windowing by using temporal subsequences instead of the raw chronicle.

Figure 7 shows the results of the data validation for the optimal window size of 30 min. The results were less emphatic than what we had before. Nevertheless, all rainy weather was identified as anomalous. In fact, by its principle, OC-SVM managed to identify the majority class which in our case is the dry weather as being the normal class, but all the rest was considered to be abnormal. Quantitatively, the *F1* score in this case was 0.46 for the training data and 0.16 for the test data. As the model



**Figure 6** | Results of turbidity data validation using OC-SVM and without data windowing. Please refer to the online version of this paper to see this figure in colour: <https://dx.doi.org/10.2166/wst.2023.174>.



**Figure 7** | Results of turbidity data validation using OC-SVM and with a window of 30 min. Please refer to the online version of this paper to see this figure in colour: <https://dx.doi.org/10.2166/wst.2023.174>.

could not learn the internal structure of the training data, it had trouble generalizing. The precision and recall in both cases were lower than 0.5, which means that the model was more wrong than right. Therefore, despite its results in related work, *OC-SVM* is not suitable for turbidity data in wastewater systems where the dynamics of the measurand depend on weather conditions.

On the other hand, the decision function for *Matrix Profile* is already predefined; namely the Euclidean distance and therefore the model does not require a two-stage process: training and testing. Figure 8 shows the results of validation of turbidity data. We observed that the anomalies were much better localized compared to the expert validation (see Section 3.3). No bias related to date, value or event type was observed. Indeed, by its definition, the Matrix Profile model is close to the validation by an operator, which consists in ‘memorizing’ the normal patterns and invalidating only what is out of the ordinary. Therefore, the Matrix Profile algorithm did not filter out rainy weather data if it had already been encountered in the data record.

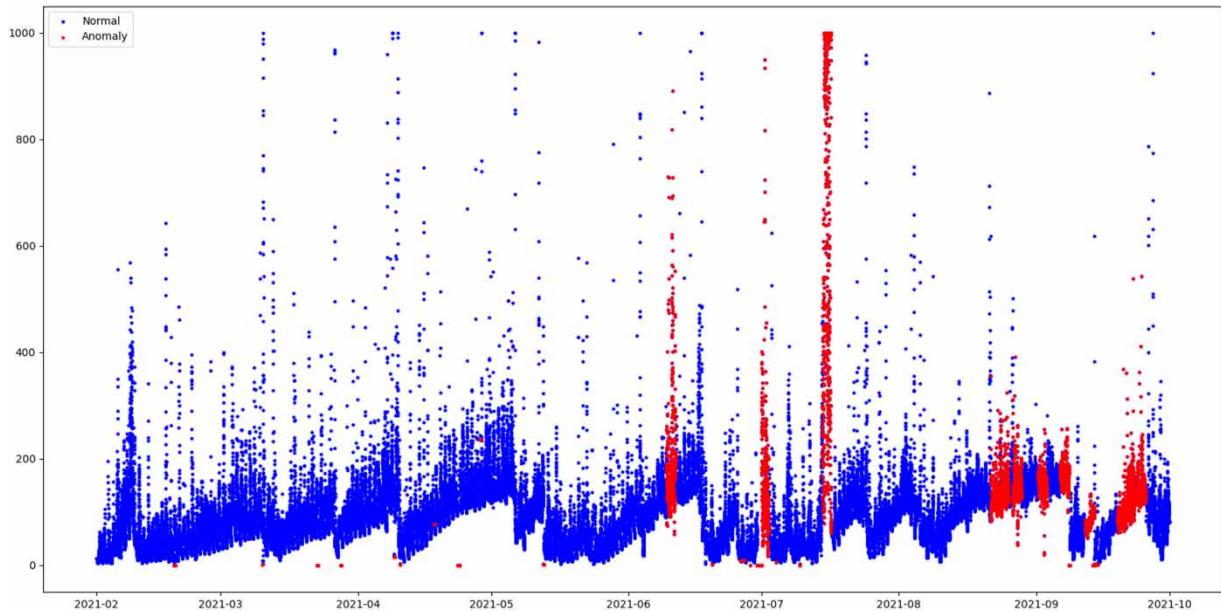


Figure 8 | Results of turbidity data validation using Matrix Profile.

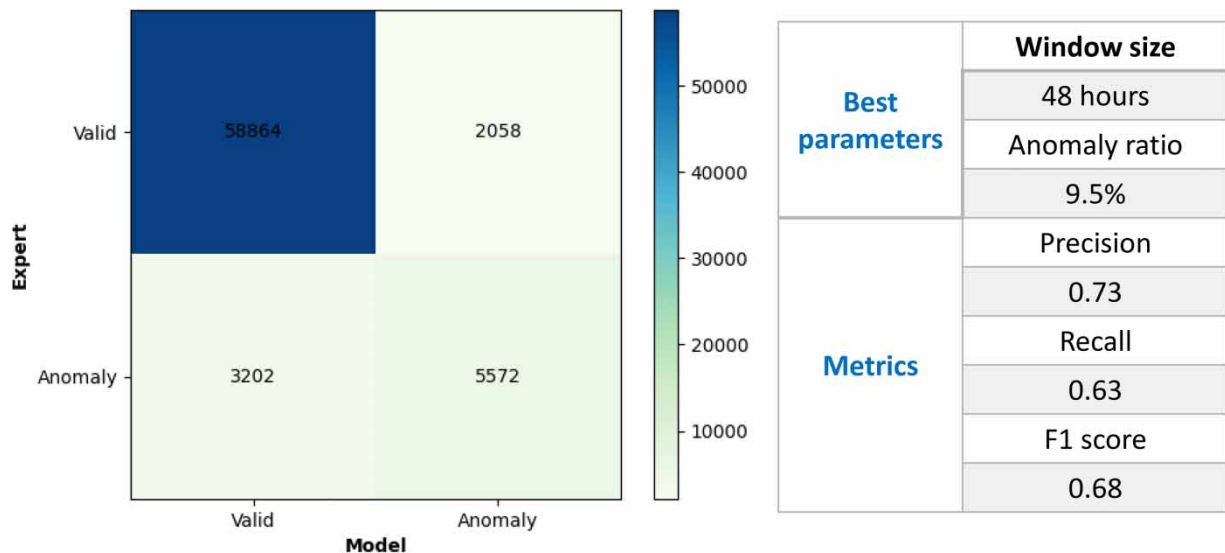


Figure 9 | Best hyperparameters and performance metrics for Matrix Profile.

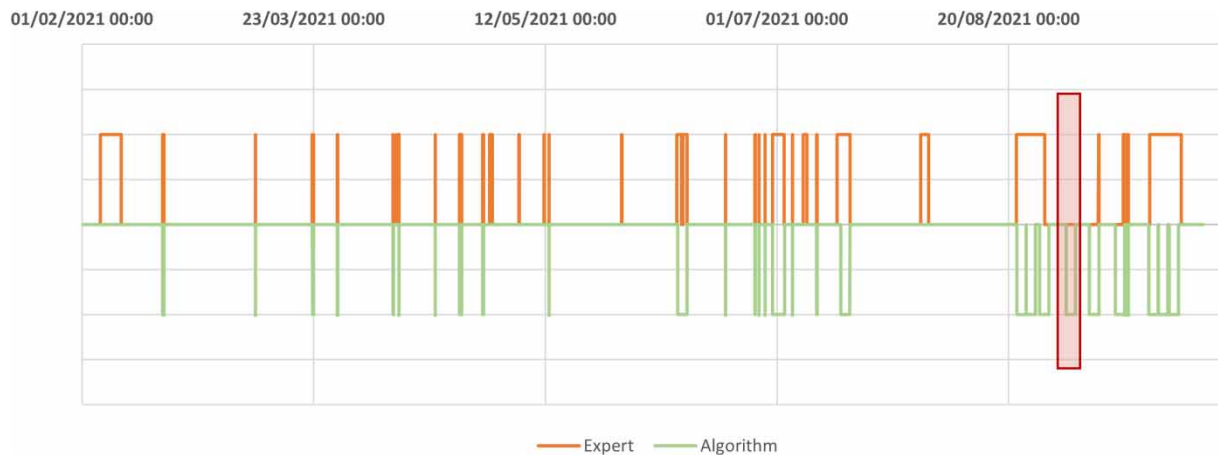
Figure 9 summarizes the results obtained. The  $F1$  score was 0.679.

We realized that the  $F1$  score of the model was of the same order of magnitude as the inter-variability between the two experts. It could be said that *Matrix Profile* obtained results equivalent to the two experts while considering their subjectivity. But moreover, the model allowed us to make objective decisions based on a rational metric which is the Euclidean distance between the subsequences, and this was within a much shorter duration than that required for an operator: we went from over 60 min to a few seconds.

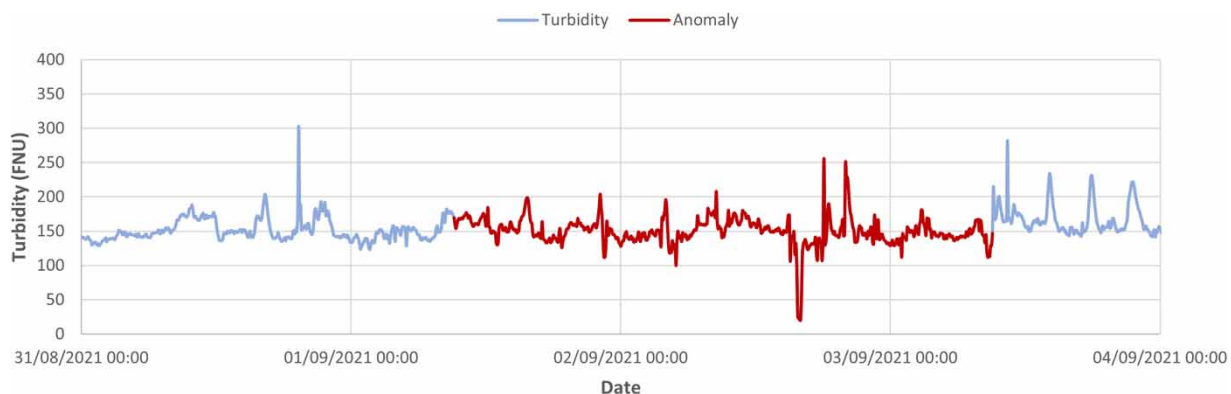
Figure 10 compares the results of the validation done by the expert and the Matrix Profile algorithm. We observed that the only *per se* false positive period identified by the algorithm was from the first of September to the third of the same month (red frame in Figure 10). Otherwise, the rest of the false positives were rather related to the delimitation of the defects.

Figure 11 shows the chronicle of the filtered turbidity during the period in question. We could clearly see a particular pattern of turbidity with more or less important sudden drops. Using domain knowledge, this behaviour can be assimilated to a defect and therefore it may be an error of the expert who ‘forgot’ to identify this period as being abnormal. We, therefore, deal with a false negative in the reference.

On the other hand, the false negatives were related to several factors, starting with the delimitation of the defects, just like false positives. Moreover, we noted the sensitivity of *Matrix Profile* to its parameters. By setting a window size of 48 h, we were unable to identify short defects. And with a fixed anomaly rate, we forced the algorithm to choose the most flagrant anomalies, and which may not be the one identified by the expert. Moreover, it should be noted that the expert performs his validation in a multivariable approach by analysing the measurements of the two turbidimeters as well as the



**Figure 10** | Boolean comparison of data validation by a domain expert and matrix profile. Please refer to the online version of this paper to see this figure in colour: <https://dx.doi.org/10.2166/wst.2023.174>.



**Figure 11** | Zoom on a turbidity anomaly as identified by matrix profile.

conductometer. However, the model Matrix Profile only has access to the turbidity reconstructed from the two sensors (Zidaoui *et al.* 2022).

## 6. CONCLUSIONS

In view of the environmental stakes, the reliability of the data measured in a wastewater network is crucial. However, given the operating constraints, the remediation of this problem is tedious and costly; hence, the interest to optimize it. Today, data validation is mainly done via automatic validation at the supervision level and/or manual validation performed by an expert. The comparison of two AI algorithms was carried out on pollution sensor data sets. The OC-SVM model proves to be unsuitable for the nature of the turbidity data in sewerage networks where different groups can be observed depending on the meteorological conditions (dry weather and rainy weather with different intensities). Thus, multi-clustering models can be better adapted to this type of data. On the other hand, the use of the Matrix Profile algorithm leads to promising results and an *F1* score of 68%, comparable to the score of one expert compared to another. In that sense, in addition to being fast, the use of this model allows it to free from the subjectivity of the domain experts and to objectify the data validation process. Being unsupervised, the Matrix Profile model is able to provide accurate results. However, the evaluation of its performance is biased by the ground truth validation bias which is the result of a subjective approach. Thus, in reality, and given its principle, the model can be considered to have more encouraging results.

Such results can be further optimized by considering a validation pool with more experts. The objective is to have more consistency in the anomalies. The idea is to consider an odd number of experts in order to avoid having middle sequences where we cannot decide on their nature: valid or invalid. Moreover, the validation of the model has been performed on a typical site, and a generalization of the approach must be performed for different rates of anomalies. How will the model behave on a site where the sensors are often malfunctioning? And on the other hand, would it be able to identify very few anomalies; an anomaly rate of around 1% for example? Finally, a multivariate approach by taking advantage of conductivity and redundancy data is also conceivable: the goal is to identify certain contextual anomalies.

## ACKNOWLEDGEMENTS

This work is part of a doctoral thesis involving the company 3D EAU and the ICube laboratory of the University of Strasbourg. We are therefore particularly grateful to these two entities for their funding and scientific support.

## DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

## CONFLICT OF INTEREST

The authors declare there is no conflict.

## REFERENCES

- Alferes, J., Lynggaard-Jensen, A., Munk-Nielsen, T., Tik, S., Vezaro, L., Sharma, A. K., Mikkelsen, P. S. & Vanrolleghem, P. A. 2013 *Validating data quality during wet weather monitoring of wastewater treatment plant influents*. *Proceedings of the Water Environment Federation* 4507–4520. <https://doi.org/10.2175/193864713813686060>.
- Branisavljević, N., Prodanović, D. & Pavlović, D. 2010 *Automatic, semi-automatic and manual validation of urban drainage data*. *Water Science and Technology* **62**, 1013–1021. <https://doi.org/10.2166/wst.2010.350>.
- Campisano, A., Cabot Ple, J., Muschalla, D., Pleau, M. & Vanrolleghem, P. A. 2013 *Potential and limitations of modern equipment for real time control of urban wastewater systems*. *Urban Water Journal* **10**, 300–311. <https://doi.org/10.1080/1573062X.2013.763996>.
- Dogo, E. M., Nwulu, N. I., Twala, B. & Aigbavboa, C. 2019a *A survey of machine learning methods applied to anomaly detection on drinking-water quality data*. *Urban Water Journal* **16**, 235–248. <https://doi.org/10.1080/1573062X.2019.1637002>.
- Dogo, E. M., Salami, A. F., Nwulu, N. I. & Aigbavboa, C. O., 2019b *Blockchain and Internet of Things-based technologies for intelligent water management system*. In: *Artificial Intelligence in IoT* (Al-Turjman, F., ed.). *Transactions on Computational Science and Computational Intelligence*. Springer International Publishing, Cham, Switzerland, pp. 129–150. [https://doi.org/10.1007/978-3-030-04110-6\\_7](https://doi.org/10.1007/978-3-030-04110-6_7).
- Feremans, L., Vercruyssen, V., Cule, B., Meert, W., Goethals, B., 2020 *Pattern-based anomaly detection in mixed-type time series*. In: *Machine Learning and Knowledge Discovery in Databases* (Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M. & Robardet, C., eds). *Lecture Notes in Computer Science*. Springer International Publishing, Cham, Switzerland, pp. 240–256. [https://doi.org/10.1007/978-3-030-46150-8\\_15](https://doi.org/10.1007/978-3-030-46150-8_15).

- Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M. & Sun, J. 2017 Anomaly detection for a water treatment system using unsupervised machine learning. In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, New Orleans, LA, USA, pp. 1058–1065. <https://doi.org/10.1109/ICDMW.2017.149>.
- Lampert, T. A., Stumpf, A. & Gancarski, P. 2016 An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. *IEEE Transactions on Image Processing* **25**, 2557–2572. <https://doi.org/10.1109/TIP.2016.2544703>.
- Methnani, S. 2012 *Diagnostic, reconstruction et identification des défauts capteurs et actionneurs: application aux stations d'épurations des eaux usées (Diagnosis, Reconstruction and Identification of Sensor and Actuator Faults: Application to Wastewater Treatment Plants)*. PhD Thesis, University of Toulon (France), National School of Engineers of Sfax (Tunisie).
- Mourad, M. & Bertrand-Krajewski, J. L. 2002 A method for automatic validation of long time series of data in urban hydrology. *Water Science and Technology* **45**, 263–270.
- Muharemi, F., Logofătu, D. & Leon, F. 2019 Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication* **3**, 294–307. <https://doi.org/10.1080/24751839.2019.1565653>.
- Piatyszek, E., Voignier, P. & Graillot, D. 2000 Fault detection on a sewer network by a combination of a Kalman filter and a binary sequential probability ratio test. *Journal of Hydrology* **230**, 258–268. [https://doi.org/10.1016/S0022-1694\(00\)00213-4](https://doi.org/10.1016/S0022-1694(00)00213-4).
- Saberi, A. 2015 *Automatic Outlier Detection in Automated Water Quality Measurement Stations. Electrical Engineering Master Report*. University of Laval, Québec, Canada.
- Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. & Williamson, R. C. 2001 Estimating the support of a high-dimensional distribution. *Neural Computation* **13**, 1443–1471. <https://doi.org/10.1162/089976601750264965>.
- Smyth, P. 1996 Bounds on the mean classification error rate of multiple experts. *Pattern Recognition Letters* **17**, 1253–1257. [https://doi.org/10.1016/0167-8655\(96\)00105-5](https://doi.org/10.1016/0167-8655(96)00105-5).
- Therrien, J.-D., Nicolăi, N. & Vanrolleghem, P. A. 2020 A critical review of the data pipeline: how wastewater system operation flows from data to intelligence. *Water Science and Technology* **82**, 2613–2634. <https://doi.org/10.2166/wst.2020.393>.
- Van Bijnen, M. & Korving, H. 2008 Application and results of automatic validation of sewer monitoring data. *Presented at the 11th International Conference on Urban Drainage*, Edinburgh, UK, p. 9.
- van Daal, P., Gruber, G., Langeveld, J., Muschalla, D. & Clemens, F. 2017 Performance evaluation of real time control in urban wastewater systems in practice: review and perspective. *Environmental Modelling & Software* **95**, 90–101. <https://doi.org/10.1016/j.envsoft.2017.06.015>.
- Venkatasubramanian, V., Rengaswamy, R., Yin, K. & Kavuri, S. N. 2003 A review of process fault detection and diagnosis. *Computers & Chemical Engineering* **27**, 293–311. [https://doi.org/10.1016/S0098-1354\(02\)00160-6](https://doi.org/10.1016/S0098-1354(02)00160-6).
- Versini, P.-A., Joannis, C. & Chebbo, G. 2015 *Guide technique sur le mesurage de la turbidité dans les réseaux d'assainissement (Guide for Measuring Turbidity in Wastewater Systems), Guides et protocoles*. ONEMA, Vincennes, France.
- Wu, R. & Keogh, E. J., 2020 *Current Time Series Anomaly Detection Benchmarks Are Flawed and Are Creating the Illusion of Progress*. <https://doi.org/10.48550/ARXIV.2009.13807>
- Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Silva, D. F., Mueen, A. & Keogh, E. 2016 Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, Barcelona, Spain, pp. 1317–1322. <https://doi.org/10.1109/ICDM.2016.0179>
- Zhang, R., Zhang, S., Lan, Y. & Jiang, J. 2008 Network Anomaly Detection Using One Class Support Vector Machine. Computer Science, Hong Kong, China.
- Zhang, J., Zhu, X., Yue, Y. & Wong, P. W. H. 2017 A real-time anomaly detection algorithm/or water quality data using dual time-moving windows. In *2017 Seventh International Conference on Innovative Computing Technology (INTECH)*. IEEE, Luton, UK, pp. 36–41. <https://doi.org/10.1109/INTECH.2017.8102421>
- Zidaoui, I., Joannis, C., Wertel, J., Isel, S., Wemmert, C., Vazquez, J. & Dufresne, M. 2022 Utilisation de l'intelligence artificielle pour la validation des mesures en continu de la pollution des eaux usées (*Use of artificial intelligence for validation of wastewater pollution monitoring data*). *TSM* **11**, 39–51. <https://doi.org/10.36904/tsm/202211039>.

First received 28 November 2022; accepted in revised form 25 May 2023. Available online 7 June 2023