Local Evaluation of Time Series Anomaly Detection Algorithms

Alexis Huet
Huawei Technologies Co., Ltd.
France
alexis.huet@huawei.com

Jose Manuel Navarro
Huawei Technologies Co., Ltd.
France
jose.manuel.navarro@huawei.com

Dario Rossi
Huawei Technologies Co., Ltd.
France
dario.rossi@huawei.com

ABSTRACT

In recent years, specific evaluation metrics for time series anomaly detection algorithms have been developed to handle the limitations of the classical precision and recall. However, such metrics are heuristically built as an aggregate of multiple desirable aspects, introduce parameters and wipe out the interpretability of the output. In this article, we first highlight the limitations of the classical precision/recall, as well as the main issues of the recent event-based metrics – for instance, we show that an adversary algorithm can reach high precision and recall on almost any dataset under weak assumption. To cope with the above problems, we propose a theoretically grounded, robust, parameter-free and interpretable extension to precision/recall metrics, based on the concept of "affiliation" between the ground truth and the prediction sets. Our metrics leverage measures of duration between ground truth and predictions, and have thus an intuitive interpretation. By further comparison against random sampling, we obtain a normalized precision/recall, quantifying how much a given set of results is better than a random baseline prediction. By construction, our approach keeps the evaluation local regarding ground truth events, enabling fine-grained visualization and interpretation of algorithmic results. We compare our proposal against various public time series anomaly detection datasets, algorithms and metrics. We further derive theoretical properties of the affiliation metrics that give explicit expectations about their behavior and ensure robustness against adversary strategies.

CCS CONCEPTS

• **General and reference** → **Evaluation**; **Metrics**; • **Mathematics of computing** → **Time series analysis**; • **Computing methodologies** → **Anomaly detection**.

KEYWORDS

time series; anomaly detection; evaluation; metrics; precision; recall

ACM Reference Format:
Alexis Huet, Jose Manuel Navarro, and Dario Rossi. 2022. Local Evaluation of Time Series Anomaly Detection Algorithms. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539339

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539339

1 INTRODUCTION

Time series anomaly detection is the field consisting in detecting elements of a time series that behave differently from the rest of the data. This field attracted interest in recent years with the rise of monitoring systems collecting a large amount of data over time, mainly for the purpose of troubleshooting and security. Many scientific domains are involved: water control industrial systems [8, 24], Web traffic [15, 31], servers of Internet companies [21, 26], spacecraft telemetry [10], and also medicine or robotics [30, 1]. Due to the nature of the series, each anomaly (referred as an *event* in the context of time series) can be a point in time (point-based anomaly) or occupy a range of consecutive samples (range-based anomaly). The detection is performed in a supervised or in a unsupervised way, but the resulting performance of the algorithm is generally always assessed against ground truth labels that have been previously collected (either in controlled environments or labeled by experts in the field). This assessment is realized with *evaluation metrics* taking as input both the ground truth and the predicted labels, and outputting one or multiple scores. The most common metrics for anomaly detection are the classical precision and recall, computed by comparing the predicted and the ground truth outputs for each sample. In the usual terminology, the positive samples refer to the samples that are predicted as positive, and are partitioned into the true positives (TP, positive samples that are also anomalous in the ground truth) and false positives (FP). Likewise, the samples predicted as negative are partitioned into false negatives (FN) and true negatives (TN). The *precision* measures the proportion $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ of positive predicted samples that are correct, whereas the *recall* measures the proportion $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ of positive ground truth samples that have been retrieved. Those classical metrics are convenient for the tasks that regard each sample separately, however this does not hold for time series datasets, where the time component is intrinsically continuous. Researchers developing detection algorithms have realized this challenge during the evaluation process and have come up with metrics fitting their specific use-case: range precision/recall [27] for evaluating the Greenhouse algorithm [17], time-aware precision/recall [11] for evaluating the HAI dataset [24], Numenta benchmark [16] for evaluating the Numenta corpus [1], etc.

In this paper, we first provide a comprehensive picture of the limitations of the classical metrics, along with the different directions of research that have been explored to handle them. Against this background, we introduce a new pair of precision/recall metrics named the *affiliation metrics* – that exhibit a series of important properties as they are theoretically principled, parameter-free, robust against adversary predictions, retain a physical meaning (as they are connected to quantities expressed in time units), and are locally interpretable (allowing to troubleshoot detection at individual event level). Summarizing our main contributions:

- We show that existing range-based metrics for anomaly detection are easily gamed by adversary predictions: this complements dataset flaws outlined in [30], and concur in creating an "illusion of progress".
- We introduce the affiliation metrics, an extension of the classical precision/recall for time series anomaly detection that is local, parameter-free, and applicable generically on both point and range-based anomalies.
- We produce closed-form expectations of the affiliation metrics in theoretical scenarios, proving their robustness against adversary strategies.
- We contrast the affiliation metrics to existing metrics in real datasets, and further show local interpretability at the event level, giving visual clues for algorithmic comparison.

In the following sections, we detail the background for extending the metrics for time series (Sec. 2) before introducing the affiliation metrics (Sec. 3). We then evaluate theoretical and practical properties of the proposed metrics (Sec. 4). We conclude by discussing the application scope of the affiliation metrics (Sec. 5 and Sec. 6).

2 BACKGROUND AND MOTIVATION

We first introduce the limitations of the classical metrics for time series anomaly detection, which led to the design of new metrics that we briefly overview. We next illustrate the main limits of those proposed metrics, and sum up the main desirable goals for proper metrics definition, which motivated our work in the first place.

Limitations of the classical metrics for time series tasks. As summarized in Tab. 1 and illustrated in Fig. 1, two main limitations have been observed in the literature for dealing with time series tasks (e.g., point or range anomaly detection, segmentation, or change point detection). The (A) *unawareness of temporal adjacency* prevents the metric from valuing the proximity between the samples. For instance, a prediction closely located in time to a ground truth label is adding both a FP and a FN samples instead of being considered as a TP, even for a one sample miss [6, 22, 16, 7]. Similarly, the predictions located closely after the end of a ground truth event (dubbed as "ambiguous samples" by Hwang et al. [11]) are immediately penalized without any tolerance. The other aspect concerns the (B) *unawareness of the events durations*, which relates to evaluation of the individual samples without considering each event as a single unit. Adverse consequences include the overrating of long events, in the sense that correctly detecting such event will be rewarded much more than correctly detecting another single-sample outlier [27, 11].

Recent time series evaluation metrics. To cope with the above problems, numerous evaluation metrics have been recently introduced to better handle the time component, that can be grouped into three main categories: (i) *distance-based* metrics, (ii) *window-based* metrics, and (iii) metrics specific to *range-based* anomalies.

The first direction to handle near detection, i.e. limitation (A), has been to employ direct measurement of the *distances* between the elements from the two sets to derive a single score. This score is usually measured as a total deviation distance and is meant to be minimized. Since the Hausdorff distance is sensitive to the presence of any outlier, it is not suitable for evaluating the time series prediction tasks [4, 12]. Further work have therefore carried out modified Hausdorff distances [12] built on metrics introduced for computer vision tasks [3, 2].

Table 1: Limitations of the classical precision and recall for evaluating time series tasks as exposed in the literature.

	Limitation	Aspect	Mentions
(A)	Unawareness of the temporal adjacency	Inter-events	[6, 22, 16, 7, 11]
(B)	Unawareness of the events durations	Intra-event	[27, 11]

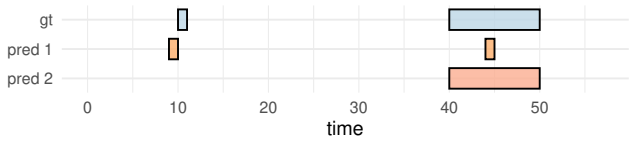


Figure 1: The classical precision/recall of predictions 1 and 2 against ground truth are 0.50/0.09 **and** 1.00/0.91**, illustrating resp. limitations (A), since each ground truth event is approximately detected but the scores are low, and (B), since only a single event is correctly detected but the scores are high.**

An orthogonal direction of research to handle limitation (A) has been to surround each ground truth event by a *window*. Each window containing a predicted element is considered as a TP, relaxing the difficulty to obtain it compared to the classical matrix of confusion. The precision and recall are then deduced from this new counting. For point anomaly detection, Gensler and Sick [6] propose to count the predictions within a ground truth window as a TP, but only once for each window. The same principle has been used in the context of change point detections [28]. The scoring of the ambiguous samples located after the anomalous point or range is also performed from ground truth windows: both in the Numenta Anomaly Benchmark (NAB) scoring [16] and in the time-aware precision/recall (TaP/TaR) metrics [11], predictions are mapped to a score based on a decaying sigmoid function. Finally, Scharwächter and Müller [22] compute the TP differently depending on the selected point of view, either from the predictions or from the ground truth.

To handle limitation (B), *range-based* metrics have recently been proposed [27, 11, 31, 14]. The range precision/recall (RP/RR) [27] and the TaP/TaR [11] are designed for anomaly detection events. The common idea consists in rewarding both the presence and the size of an overlap between the predicted and the ground truth events. For those metrics, each event is considered as a single unit irrespective of its length. Finally, the point adjust metrics [31] and an extension [14] have been introduced to ease the scoring of the range-based events: the computation consists in sticking to the classical metrics, after initially extending each TP sample to the whole corresponding ground truth event. As [31, 14] however do not deal with limitation (A) nor (B), this yield to a recent adaptation named F1-composite metric [5].

Characteristics and limitations of the existing metrics. The characteristics of the existing metrics are summarized in Tab. 2. Most of the metrics (except distance based) take the form of a

Table 2: Characteristics of the recent metrics extending the classical precision/recall for time series anomaly detection.

(Class) Metric	P/R form?	mainly for	handles (A)	handles (B)	# param.
(i) distance [4, 12, 7]	✗	point	✓	✗	0
(ii) window [6, 16, 22]	✓	point	✓	✗	1
RP/RR [27]	✓	range	✗	✓	4
(iii) TaP/TaR [11]	✓	range	✓	✓	3
point adjust [31, 14]	✓	range	✗	✗	0
(—) affiliation (this work)	✓	both	✓	✓	0

precision and recall pair. Except TaP/TaR [11], no metric handles both limitations (A) and (B). Distance-based and windowed metrics are mostly limited to point anomalies and cannot handle (B), while RP/RR and point adjust metrics are not able to handle limitation (A), e.g. a one sample miss, since the overlap is empty in this case.

Remarkably, with the exception of distance-based and point adjust metrics, *one or multiple parameters* have to be selected to control several aspects of the scoring process, adding a number of parameters in the anomaly detection pipeline, likely leading to a lack of generality. While their purpose can be understood, fine-tuning or overspecializing the metric may be counterproductive for real use-cases. Conversely while most of those parameters have default values, their setting is not always trivial. For instance, to handle (A) both TaP/TaR and window-based metrics select a window size which is tuned for each dataset and drastically impacts the evaluation [22, 25]. It even prevents the well-definition of the metrics when two windows overlap (i.e. leading to a precision possibly greater than one, as mentioned in [28], and occurring but not discussed for [6, 16, 11]).

Interpretability and robustness against adversary algorithms or naive random predictions. A core issue finally concerns the *interpretability* of the recent range-based metrics. For instance RP/RR and TaP/TaR derive per event quantities, averaged over the whole dataset. Individual quantities are however difficult to understand, as they combine multiple aspects: existence of an overlap, proportion of the overlap, relative positions, or ambiguous samples. Additionally, each predicted event is considered as an individual element, which is detrimental for the global meaning of the score, since the number of predicted events and their positions are not controlled: the resulting metric is not considering each region equally, that is any cluster of predictions in a specific region can impact globally the final score. This lack of locality allows the development of adversary algorithms reaching both high recall and high precision, as we show in Sec. 4.2.

Another aspect related to interpretability concerns the lack of statistical properties constraining the behavior of the metrics. Overestimation of the scores has been shown for window metrics using Monte Carlo simulations [22], while for the point adjust metrics it has been recently pointed out that "even a random anomaly score can easily turn into a state-of-the-art time series anomaly detection method" [14] – while in our proposal, random scores define the lower bound baseline as we formalize in Sec. 4.4 and demonstrate in Appendix C.

Design targets. To sum up, a convenient precision/recall pair for time series anomaly detection should include the following desirable targets, that have not been jointly addressed in prior art:

- handle the limitations (A) and (B) introduced by the presence of time (only addressed by TaP/TaR);
- parameter-free definition (only available with distance and point adjust metrics);
- expressiveness of the scores, in the sense that a slight improve of the predictions should result in a slight improve of the scores (only addressed by the distance metrics);
- local interpretability of the scores (not addressed so far);
- existence of statistical bounds of the scores (only addressed by simulation for window-based metrics).

3 LOCAL EVALUATION BASED ON AFFILIATION

In this section, we introduce and describe the reasoning of the *affiliation metrics* for evaluating anomalous events. Three concepts are developed. First, we define a directed average distance between sets, to measure how far the events are one from each other (Sec. 3.1). Then, each prediction is affiliated to the closest ground truth event, allowing a local perspective and maintaining the interpretability even for outlying predictions (Sec. 3.2). Finally, the observed temporal distances are locally converted into probabilities, by comparing them against random sampling, and are averaged into a precision/recall pair (Sec. 3.3). The practical aspects are detailed at the end of the section (Sec. 3.4).

We limitedly detail description for anomalous events consisting of ranges (particularization to the case of point anomalies is deferred to the Appendix B). An anomalous event is described by a continuous time interval $[t_{\text{start}}, t_{\text{stop}})$ with $t_{\text{stop}} > t_{\text{start}}$. Both prediction and ground truth are represented by a set of disjoint anomalous events, respectively noted $\text{pred}_1, \dots \text{pred}_m$ and $\text{gt}_1, \dots \text{gt}_n$.

3.1 Average distance between sets

The distance from a point x to a set Y is defined, as usual, by: $\text{dist}(x, Y) := \min_{y \in Y} |x - y|$. For measuring the distance from a set X to another, we consider the *average directed distance* defined by: $\text{dist}(X, Y) := \frac{1}{|X|} \int_{x \in X} \text{dist}(x, Y) dx$. For the corner cases, we have $\text{dist}(X, \varnothing) = +\infty$ for nonempty X, and we keep $\text{dist}(\varnothing, Y)$ undefined for all Y. This function is not a metric in the mathematical sense because it does not satisfy symmetry nor the triangle inequality, however is nonnegative and, for the case where X, Y are sets of disjoint anomalous events, verifies $\text{dist}(X, Y) = 0 \Leftrightarrow (X \subset Y \text{ and } X \neq \varnothing)$. It has been introduced as a part of the modified Hausdorff distance [3] used in computer vision.

This directed distance has been selected for smoothness and interpretability reasons. First, contrary to the Hausdorff metric or to a simple threshold based on a window size, it satisfies smooth variation [3, 12] since each sample contributes to the total score. Then, it has a clear interpretation as an average and retains a physical meaning as a time. Additionally, the distance is not converted into an undirected one – such as taking the maximum over the two directions in the modified Hausdorff distance – to prevent the dilution of the interpretation by an additional layer. Indeed, our main idea in using this function is to relate the directed distance

from prediction to ground truth events to a precision, and the one from the predicted events to the ground truth to a recall.

This idea is illustrated through the example in Fig. 2a. The prediction comprises three events whereas the ground truth is a single event. From the predicted events to the ground truth (left of Fig. 2a), the directed distance is short, since 80% of the prediction area matches with the ground truth while the remaining predicted event (accounting for 20%) is at a distance of 1min30s in average. In total, the directed distance is 18s. This short distance is interpreted as a good precision of the predictions. Improving the precision would be possible by removing the last predicted event and would lead to an average distance of zero corresponding to a perfect precision. From the ground truth to the predicted events (right of Fig. 2a), the directed distance is computed in the other direction, giving a directed distance of 76.5s. This distance is interpreted as a recall, currently expressed in time. In that case, removing the last predicted event would not change the directed distance nor the recall.

3.2 Local affiliation to the closest ground truth

The time axis is partitioned by assigning each time t to the closest ground truth event gt_j, $j \in [\![1, n]\!]$. The resulting partition consists of n intervals, and the j-th one I_j is called the *zone of affiliation* of the ground truth event j. On each zone of affiliation, the ground truth and predictions belonging to it are obtained and the average directed distances described in Sec. 3.1 are computed to retrieve the *individual precision/recall distances*. Formally, the union of all predictions is noted $\mathrm{pred} := \bigcup_{i=1}^{m} \mathrm{pred}_i$ and the formulas are given, for $j \in [\![1, n]\!]$, as follows:

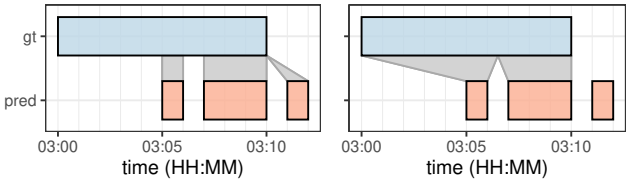
$$\mathrm{D}_{\mathrm{precision}_j} := \mathrm{dist}\left(\mathrm{pred} \cap I_j,\ \mathrm{gt}_j\right), \tag{1}$$

$$\mathrm{D}_{\mathrm{recall}_j} := \mathrm{dist}\left(\mathrm{gt}_j,\ \mathrm{pred} \cap I_j\right). \tag{2}$$

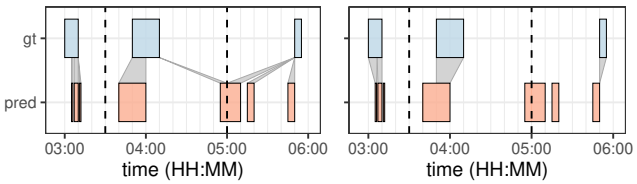
The choice made by isolating each zone of affiliation from the ground truth perspective (for both precision/recall) is driven by the difference of reliability expected regarding the ground truth and the predictions. For the ground truth, upstream control measures can be taken in place to assess its quality, for instance by ensuring that each labeled event corresponds to a separate anomaly that would need to be identified. On the other hand, there should not be expectation for the shape of the predicted events, particularly on their distribution along the time line. As a consequence, taking the perspective of the ground truth is preventing the possibility for multiple short predictions stacked in a local area to bias globally the evaluation metric (cf. Sec. 4.2). Overall, each individual distance is related to a single ground truth and can be interpreted locally.

A practical derivation of the individual precision/recall distances is presented for the example shown in Fig. 2b. First, the total timeline is cut into the zones of affiliation: $(-\infty, 3{:}30)$, $[3{:}30, 5{:}00)$, $[5{:}00, +\infty)$. Then the individual precision/recall distances are computed on each zone. For the first zone, it corresponds exactly to the case shown in Fig. 2a, giving 18s / 76.5s for respectively the precision and the recall. The results are 11min30s / 2min30s for the second zone and 31min15s / 2min30s for the third one.

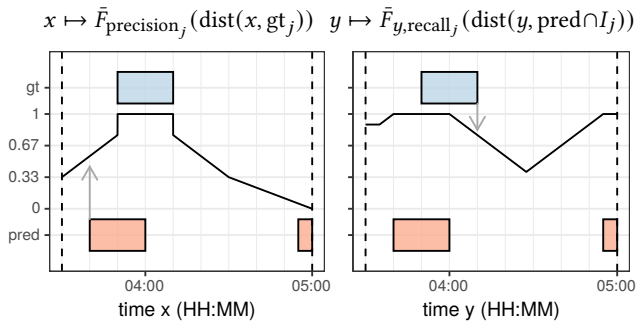
Individual distances are providing a summarized view of each ground truth event expressed in a meaningful unit (i.e., time), and can be used as they are by practitioners and domain experts. However, for comparing the different ground truth events of a dataset, it would be desirable to convert each individual value to the $[0, 1]$ range, by assuming that each ground truth event is equally important (as opposed to the classical metrics that consider each sample as equally important). This normalization step is presented in Sec. 3.3.



(a) *Average distance between sets*: example of the directed distance computed from predicted events to ground truth (left) and from ground truth to predicted events (right).



(b) *Local affiliation to the closest ground truth event*: example resulting in zones delimited by the dashed lines. The zones are similar for both directions: precision (left), and recall (right).



(c) *Comparison against random sampling*: example for converting each predicted sample to a precision score (left) and each ground truth sample to a recall score (right).

Figure 2: Illustration of the three steps for computing the affiliation metrics.

3.3 Comparison against random sampling

The final normalization step consists in replacing each distance measured at the sample level with a probability $[0, 1]$, by comparing this distance against the random sampling of a prediction. We here introduce the main concept of this step, but we point out that the survival functions and the integrals involved in the computations have closed-form expressions, that are detailed in Appendix A.

Individual precision probability. As previously expressed, each affiliation zone is considered separately. On each zone, the ground truth gt_j is fixed and a random prediction is made: this prediction X is a random variable corresponding to a single point

in time uniformly sampled within the affiliation zone. For the precision, the distance from X to the ground truth gt_j is a random variable with a cumulative distribution function $F_{\mathrm{precision}_j}$ (that only depends on four time elements: the current zone and ground truth intervals). The value of an observed distance $d \geq 0$ against this random prediction is defined by the survival function:

$$\bar{F}_{\mathrm{precision}_j}(d) := 1 - F_{\mathrm{precision}_j}(d-). \tag{3}$$

Applying this function on each predicted sample, we derive the *individual precision probability* (see also Appendix A.2.1) as follows:

$$P_{\mathrm{precision}_j} := \frac{1}{|\mathrm{pred} \cap I_j|} \int_{x \in \mathrm{pred} \cap I_j} \bar{F}_{\mathrm{precision}_j}(\mathrm{dist}(x, \mathrm{gt}_j))dx. \tag{4}$$

Individual recall probability. For the recall, the distance depends on the considered ground truth sample $y \in \mathrm{gt}_j$. Knowing it, along with the affiliation zone interval, the distance from y to X is a random variable with a distribution F_{y,recall_j}. As before, the value of an observed distance is defined by the survival function, and by applying this function on each ground truth sample, we derive the *individual recall probability* as follows (see also Appendix A.2.2):

$$P_{\mathrm{recall}_j} := \frac{1}{|\mathrm{gt}_j|} \int_{y \in \mathrm{gt}_j} \bar{F}_{y,\mathrm{recall}_j}(\mathrm{dist}(y, \mathrm{pred} \cap I_j))dy. \tag{5}$$

Illustrative example. The computation for the second affiliation zone of Fig. 2b is detailed with the help of Fig. 2c. For the precision, the function $x \mapsto \bar{F}_{\mathrm{precision}_j}(\mathrm{dist}(x, \mathrm{gt}_j))$ is deduced from the survival distribution (left of Fig. 2c). We observe that any predicted sample located inside the ground truth area would get a value of 1, while this value decreases to 0 as the distance to the ground truth event increases. For instance, the value of the predicted time 3:40 is 0.556 (illustrated by a gray arrow). Taking the average over all the predicted elements, we obtain an individual precision probability of 0.672. For the recall, the function $y \mapsto \bar{F}_{y,\mathrm{recall}_j}(\mathrm{dist}(y, \mathrm{pred} \cap I_j))$ is computed for each ground truth sample y (right of Fig. 2c). Any ground truth element inside the predicted area has a value of 1, that decreases as the distance to the predicted elements increases.

In the case of y = 04:10, the evaluation gives a value of 0.778 (illustrated by a gray arrow). Taking the average over all the ground truth elements, we obtain an individual recall probability of 0.944.

Interpretation and understanding. The baseline for the random prediction has been selected with the lightest possible a priori on the predicted samples, namely by selecting a single predicted sample randomly in each affiliation zone. In particular, no information about the shape, the distribution, or the number of the predicted events is used.

Some observations can be directly made from the construction. First, we note that a precision of 1 is equivalent to pred $\subset$ gt, while a recall of 1 corresponds to gt $\subset$ pred, which is in alignment with the classical metrics. Then, a precision or a recall lower than 0.5 (resp. around 0.5) is interpreted as doing worse than (resp. as bad as) a random prediction, meaning that the predictions made have not been able to provide information about the location of the event. The exact formulation of those interpretations constraining the behavior of the affiliation metrics are formalized through properties in Sec. 4.4.

Averaging of the individual precision/recall probabilities. The *precision/recall* are defined as the mean over the defined individual probabilities. Following the corner cases about the average distance between sets, an individual precision probability is undefined only when there is not any predicted element affiliated to gt_j. We let $S := \{j \in [\![1, n]\!] \; ; \; \mathrm{pred} \cap I_j \neq \emptyset\}$, and obtain:

$$P_{\mathrm{precision}} := \frac{1}{|S|} \sum_{j \in S} P_{\mathrm{precision}_j}, \quad P_{\mathrm{recall}} := \frac{1}{n} \sum_{j=1}^{n} P_{\mathrm{recall}_j}. \tag{6}$$

3.4 Practical settings

In real settings, the affiliation metrics can be applied on evenly or unevenly-spaced time series. The ground truth and the predictions have the form of a binary vector of same length N, where the anomalous samples are indicated by 1 and the normal ones by 0, each index i corresponding to a timestamp $t(i)$, as depicted in Tab. 3.

Table 3: Illustration of the practical input shape for calculating the affiliation metrics, here corresponding to Fig. 2a

index i	1	2	3	4	5	6	7	8
gt	1	1	1	1	1	0	0	0
pred	0	0	1	0	1	0	1	0
$t(i)$	3:00	3:02	3:05	3:06	3:07	3:10	3:11	3:12

A consistent way to convert those indexes into range-based events is to match any positive index i to the corresponding interval $[t(i), t(i+1))$. The last timestamp $t(N+1)$ is clear for evenly spaced measures, otherwise it needs to be selected. In the example of Tab. 3, it gives the events pred_1 = [3:05, 3:06), pred_2 = [3:07, 3:10), pred_3 = [3:11, 3:12) and gt_1 = [3:00, 3:10). It corresponds to the example of Fig. 2a. Given those intervals, the individual precision/recall distances are available. With the additional knowledge of the total range $[t(1), t(N+1)]$, the precision/recall probabilities can be computed. Moreover, due to the closed-form expressions developed in Appendix A, their implementation is computationally efficient. Further practical considerations for reproducibility, including pointers to code, etc. are discussed in Appendix D.

4 EVALUATION AND PROPERTIES

The evaluation of the affiliated metrics is performed against the range-based anomaly detection metrics RP/RR and TaP/TaR, along with the classical sample-based metrics on a set of algorithms and datasets (Sec. 4.1). We first show that adversary predictions easily fool range metrics, whereas affiliated metrics are local and robust (Sec. 4.2). Second, we show that this local construction further allows a detailed per-event interpretation and comparison of the results given by the anomaly detection algorithms (Sec. 4.3). Finally, we formalize (Sec. 4.4) and prove (Appendix C) theoretical properties of the affiliated metrics in typical prediction scenarios.

4.1 Evaluation settings

Benchmark anomaly detection algorithms. As our primary aim is to contrast metrics for algorithmic evaluation, we rely on the

anomaly detection algorithms that were selected by the previous authors [27, 11]. For the datasets Machine-Temp, NYC-Taxi, and Twitter-AAPL, the predicted events are deduced from three algorithms: Greenhouse [17], LSTM-AD [20], both based on neural networks, and one implemented in the Luminol library based on time series bitmaps [29, 18] (Luminol TSB). For the SWaT dataset, two unsupervised algorithms are selected (iForest [19] and OCSVM [23]) along with a neural network approach (seq2seq [13]).

Datasets. Similarly, we select the four time series datasets used in [27, 11] for the comparison. Three of them are those used by authors of the RP/RR metrics [27], and taken from the publicly available Numenta Anomaly Benchmark Data Corpus [1]: *Machine-Temp* (temperature sensor data of an industrial machine), *NYC-Taxi* (number of New York City taxi passengers), and *Twitter-AAPL* (collection of Twitter mentions of the ticker symbol AAPL). The other dataset has been used by authors of the TaP/TaR metrics [11], and is a secure water treatment testbed known as *SWaT* [8]. A short description of the number of samples, percentage of anomalies, and number of anomalous events is available in Tab. 4.

4.2 Adversary predictions

Adversary algorithm. We design an algorithm to deceive metrics that consider each predicted event as a unit, by aggregating numerous predictions within a local region in order to impact the score globally. Following the definition coined and discussed by Wu and Keogh [30], an anomalous event is said *trivial* if it can be identified with a single elementary line of code (e.g., a threshold)[1]. The function applied to derive a trivial event for the selected datasets is reported in Tab. 4. Regarding SWaT, since we do not have direct access to the raw data, we simply consider one of the events, for instance the first one containing 940 samples, as trivial.

Knowing any such trivial event, the adversary predictions are defined in two steps: (i) *within the trivial event*, set the maximum possible number of predicted events by alternating positive and negative samples. (ii) *outside the trivial event*, label all the samples as positive. We denote this methodology for producing the predicted events as *adversary algorithm*[2].

The construction of the adversary is shown for the NYC-Taxi dataset in Fig. 3. First, given the raw values of the series, a trivial event is identified (it does not need to cover entirely the true event). Then, the adversary algorithm is applied to produce the predictions, resulting in thirteen predicted events.

Trivial and adversary predictions are reported along with the other anomaly detection algorithms in Tab. 5.

Results for the RP/RR and TaP/TaR metrics. The second and third rows of Tab. 5 show that the adversary algorithm beats by far any other tested algorithms, although this adversary algorithm being unable to provide any informative content about the position of the anomalies, since almost all the samples are labelled positively.

[1]Note that the threshold can be on the original data or from a new time series derived from it using basic primitive operations, such as a moving average of the series.

[2]We stress that similar adversary strategies can be defined for other metrics non included in the evaluation, such as point adjust F1-composite metric [5]. The adversary in this case would (i) keep the trivial event of duration T and (ii) add N additional predictions of length $\frac{T}{10N}$ regularly spaced over the rest of the interval, so that the precision is at least $T/(T + N(T/(10N))) = 1/(1 + 1/10) \approx 0.91$, while the recall is 1 for sufficiently large N.

Table 4: Description of the datasets and of the corresponding base trivial events for the adversary algorithm.

	Machine Temp	NYC Taxi	Twitter AAPL	SWaT
samples	17682	2307	11889	449919
anom.	6%	27%	7%	12%
# events	2	3	2	35
trivial	$Y_t < 40$	$Y_t < 1250$	$Y_t > 12000$	First event

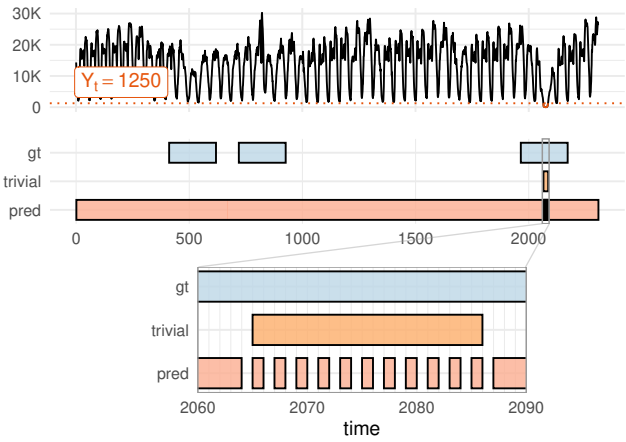


Figure 3: Construction of the adversary predictions for the NYC-Taxi dataset, with a zoom on the interval $t \in [2060, 2090)$ containing the trivial event $Y_t < 1250$. The predicted events (bottom row) are calculated using the adversary algorithm given a trivial event (middle row). The predictions are evaluated against the ground truth (top row).

In this way, those algorithms are unable to provide reliable results in a conceivable situation.

The results of the evaluation for the adversary algorithm can be retrieved theoretically, for a dataset containing n ground truth events and where the trivial event is cut into k pieces. First, $n - 1$ events are fully recalled, while the remaining one overlaps 50% of the trivial event, giving a recall around $1 - \frac{1}{2n}$ for RP/RR and at least $1 - \frac{1}{4n}$ for TaP/TaR (using default parameters). Then, the precision is perfect for k pieces and uncontrolled for the two remaining ones, giving a total precision of at least $\frac{k}{k+2}$. Globally, both precision and recall are close to 1 for k and n sufficiently large.

Results for the classical and affiliation metrics. We observe in Tab. 5 that the classical and the affiliation metrics are not sensitive to the adversary algorithm.

For the classical metrics, that operate at the *sample level*, the trivial algorithm has a precision of one while the recall covers the proportion of correctly identified *samples*. The adversary algorithm increases the recall but drastically impacts the precision, since most of the elements are now FP.

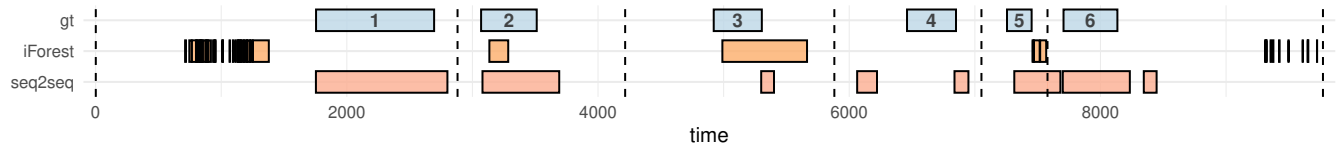
For the affiliation metrics, that operate at the *event level*, the trivial algorithm has also precision of one while the recall covers the proportion of correctly identified *events* (e.g. $1/35 \approx 0.03$ for

Table 5: Comparison of the metrics using datasets and algorithm predictions selected by [27, 11], against a trivial prediction and an adversary algorithm. Each cell shows the precision/recall/F1-score w.r.t. a certain metric, algorithm, and dataset. The algorithm reaching the best F1-score is shown in bold for each metric and dataset.

Metric	Algorithm	Dataset Machine-Temp	Dataset NYC-Taxi	Dataset Twitter-AAPL	Algorithm	Dataset SWaT
Classical (sample-based precision/recall)	trivial	**1.00/0.34/0.50**	1.00/0.03/0.07	**1.00/0.13/0.23**	trivial	1.00/0.02/0.03
	adversary	0.05/0.83/0.10	**0.27/0.98/0.42**	0.06/0.93/0.12	adversary	0.12/0.99/0.21
	Greenhouse	0.33/0.42/0.37	0.23/0.43/0.30	0.50/0.06/0.11	iForest	**0.30/0.74/0.43**
	LSTM-AD	0.06/1.00/0.12	0.24/0.49/0.32	0.24/0.13/0.17	OCSVM	0.17/0.85/0.28
	Luminol TSB	0.10/0.04/0.06	0.15/0.02/0.04	0.37/0.07/0.11	seq2seq	0.59/0.25/0.35
RP/RR	trivial	1.00/0.46/0.63	1.00/0.18/0.31	1.00/0.31/0.48	trivial	1.00/0.03/0.06
	adversary	**0.99/0.75/0.85**	**0.88/0.85/0.86**	**0.96/0.75/0.85**	adversary	**1.00/0.99/0.99**
	Greenhouse	0.15/0.57/0.24	0.23/0.52/0.32	0.26/0.51/0.35	iForest	0.04/0.52/0.08
	LSTM-AD	0.03/1.00/0.06	0.27/0.51/0.35	0.10/0.51/0.17	OCSVM	0.14/0.61/0.23
	Luminol TSB	0.08/0.50/0.14	0.14/0.34/0.20	0.24/0.51/0.32	seq2seq	0.35/0.66/0.46
TaP/TaR	trivial	1.00/0.42/0.59	1.00/0.02/0.03	1.00/0.06/0.12	trivial	1.00/0.03/0.06
	adversary	**0.99/0.96/0.97**	**0.93/1.00/0.96**	**0.96/1.00/0.98**	adversary	**1.00/0.99/1.00**
	Greenhouse	0.17/0.47/0.25	0.32/0.64/0.43	0.42/0.04/0.07	iForest	0.05/0.40/0.09
	LSTM-AD	0.04/1.00/0.07	0.36/0.66/0.47	0.13/0.08/0.10	OCSVM	0.17/0.55/0.26
	Luminol TSB	0.10/0.02/0.04	0.23/0.02/0.03	0.26/0.04/0.07	seq2seq	0.44/0.65/0.52
Affiliation	trivial	1.00/0.50/0.66	1.00/0.30/0.46	1.00/0.49/0.66	trivial	1.00/0.03/0.06
	adversary	0.49/1.00/0.66	**0.54/1.00/0.70**	0.50/1.00/0.67	adversary	0.53/1.00/0.69
	Greenhouse	**0.71/0.99/0.83**	0.51/0.99/0.67	**0.78/0.98/0.87**	iForest	0.52/0.84/0.64
	LSTM-AD	0.50/1.00/0.67	0.51/1.00/0.67	0.66/0.99/0.79	OCSVM	0.65/0.70/0.68
	Luminol TSB	0.54/0.99/0.70	0.38/0.79/0.51	0.73/0.98/0.83	seq2seq	**0.86/0.79/0.83**

Table 6: Tabulated and visual per-event comparison between iForest and seq2seq on the SWaT dataset for the first six ground truth events, using the affiliation metrics. The affiliation zones are indicated with dashed lines.

Algorithm	Mean of 35 events	Ev. 1	Ev. 2	Ev. 3	Ev. 4	Ev. 5	Ev. 6
iForest	0.52/0.84/0.64	0.37/0.53/0.44	**1.00/0.91/0.95**	**0.76/0.99/0.86**	NaN/0/NaN	0.38/0.60/0.46	0.09/0.21/0.12
seq2seq	**0.86/0.79/0.83**	**0.96/1.00/0.98**	0.86/1.00/0.93	0.73/0.78/0.75	**0.39/0.71/0.50**	**0.71/0.97/0.82**	**0.88/1.00/0.94**



the SWaT dataset). The adversary algorithm increases the recall but reduces the precision to 0.50, meaning that the predictions are not better compared to a random prediction (by definition of the construction of the metric, as developed in Sec. 4.4).

Regarding the other algorithms, we consider for instance Greenhouse and observe that it did not provide informative predictions for the NYC-Taxi (affiliation metrics with a precision around 0.50), while performing better for both Machine-Temp and Twitter-AAPL. For this latter, Greenhouse performs fewer predictions that do not cover the whole ground truth events but are overlapping or close to them, explaining the small classical recall (0.06) compared to the high affiliated recall (0.98). This also holds for the SWaT dataset, for which we give a detailed event level interpretation in Sec. 4.3.

4.3 Event-level comparison

Since affiliation metrics have a local significance, each anomalous event can be analyzed individually from both precision and recall viewpoints, which is not possible with the other range-based metrics. An example of the results obtained using *iForest* and *seq2seq* for the 6 initial events of the 35 overall anomalous events of the SWaT dataset is shown in Tab. 6. The illustration intentionally mimics the one reported in [11], where however the judgement of algorithmic performance at individual events level is left to the eye of the reader. In contrast, affiliation precisely quantifies in an unbiased and unequivocal manner the detection performance of each event, offering a new light in the algorithmic evaluation through this new detailed view.

For events 1 and 6 of Tab. 6, iForest misses the ground truth events by far, while seq2seq predictions locate near and within the anomalous region. This translates into a poor precision/recall for iForest (worse than a random prediction for the precision, as bad as a random prediction for the recall) while for seq2seq the precision is good, with a perfect recall. Event 2 depicts the situation with a perfect precision (for iForest) or an almost perfect recall (for seq2seq), leading to similar performance since in both cases the main region has been correctly identified. For event 3, iForest has better identified the ground truth, even if both predictions are better than a random prediction. For event 4, the prediction is either missing (for iForest) or poor (for seq2seq). For event 5, most the predictions made by iForest are close to the limit of the affiliation zone, impacting the overall precision, whereas seq2seq also captures most of the ground truth event inducing a boost in the precision and in the recall.

Overall, seq2seq gives better results for 21 events (13 for both precision and recall, and 8 only for the precision), equivocal results for 4 events (increase of the precision while the recall decreases, or the contrary) and a worse performance for 2 events. The remaining events are either completely undetected by iForest (for 2 events) or by seq2seq (for 6 events).

Globally, iForest, OCSVM, and adversary algorithms do not provide better results compared to a random guess. Locally, e.g. for events 2 and 3, the iForest algorithm is better (in terms of F1-score) compared to seq2seq. This behavior highlights the difference in the algorithm to deduce the anomalies (per sample for iForest and in-context for seq2seq) and can help for understanding the strengths and weaknesses of the algorithms and design better ones (for example for doing ensembles).

In an operational perspective, non-normalized distances can complement the understanding for each event by providing the individual time distances for precision/recall (optionally considering directionality, i.e., in case from a practical viewpoint an early detection is preferable to a late one).

4.4 Theoretical properties

In addition with the practical results, we provide theoretical properties supporting the correct behavior of the affiliation metrics. Since each affiliated zone is considered independently, we consider a single ground truth event gt_j included in the affiliation zone I_j. We let $p = |\mathrm{gt}_j|/|I_j|$ the proportion taken by the ground truth event within its affiliation zone, which is also the proportion of positive samples. As we consider an anomalous detection task, we are expecting rare events and $p \ll 1$ in most of the cases.

In the following, we derive a closed-form of the metrics in three scenarios: first (i) when the whole interval is predicted as anomalous, then (ii) for a random prediction within the affiliation zone, and finally (iii) for a single prediction located at specific locations on the interval. Details of the proofs are available in the Appendix C.

Predicting the whole interval as anomalous. In this case, the precision and recall are given by (cf. Appendix C.1 for proof):

$$P_{\text{precision}} = \frac{1}{2} + \frac{p^2}{2}, \quad P_{\text{recall}} = 1. \tag{7}$$

In the $p \ll 1$ regime, the precision is close to $1/2$ which corresponds to a poor detector (as poor as a random predictor). For all values of p, this behavior can be put in parallel with the classical precision/recall, in which predicting all samples would give precision of p and a recall of 1.

Expected precision and recall given a single random prediction. The expected precision and recall are given by:

$$P_{\text{precision}} = \frac{1}{2} + \frac{p^2}{2}, \quad P_{\text{recall}} = \frac{1}{2}. \tag{8}$$

This property confirms that scores around $1/2$ corresponds to a random detector (cf. Appendix C.2 for proof). In this case, the classical precision/recall would give a precision of p and a recall close to 0 assuming a large number of samples.

Single prediction at a defined position. We consider different single predictions located at four different positions on the affiliation zone: (a) the border of the affiliation zone, (b) the position halfway between the border and the ground truth event, (c) the first element of the event, and (d) the center of the event. The latter case (d) corresponds to the best possible single-element prediction. Since the position of the ground truth event also impacts the results, for simplification we assume it centered within the affiliation zone.

We defer a derivation of closed-form expression for (a)-(d) in Appendix C.3, and report here an intuitive example. The scores as a function of p are reported in Fig. 4, along with an illustration of the four positions for $p = 1/5$. For $p \ll 1$, we observe that both precision and recall are 0 for distant predictions (a), and increase until reaching 1 for close predictions (c, d). In the other regimes, the precision is always 1 for overlapping predictions (c, d) but the recall decreases as p increases, representing the impossibility for a single prediction to reach a perfect recall of a large event. For instance, the results for the best single-element prediction (d) are given by (with $x_+ := \max(0, x)$ the positive part):

$$P_{\text{precision}} = 1, \quad P_{\text{recall}} = 1 - \frac{p}{2} + \frac{1}{2p}\left(p - \frac{1}{2}\right)_+^2. \tag{9}$$

5 DISCUSSION

We finally discuss the proposed affiliation metrics along complementary aspects.

Shape of the ground truth labels. The most important factor towards a correct evaluation is the ground truth labels themselves. The intended purpose and a description of the labeling strategy are necessary requirements, but other factors impact the quality of the evaluation. First, the actual output should represent an anomaly detection task, i.e. labeled events need to be rare. Furthermore, we expect that each event corresponds to a single anomaly. For this purpose, the merge of fragmented events related to a single anomaly have been proposed [30]. However, the amplitude of the labeled zone may remain imprecise: even for controlled experiments [8, 24] for which the start date is known, the time at which the system returns to a steady state remains subjective. By design, the affiliation metrics are less sensitive to precise labeling compared to the previous range or window-based metrics.

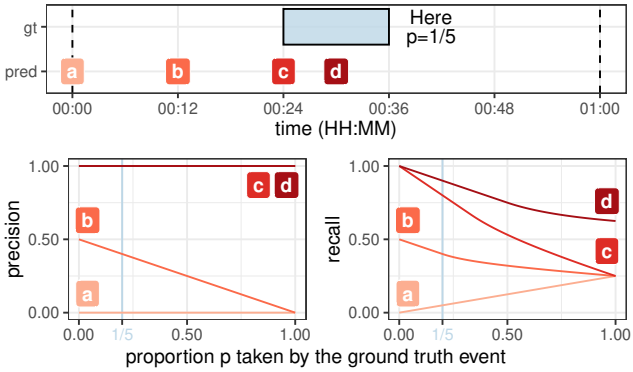


Figure 4: Given a ground truth centered in the affiliation zone and filling a proportion p, result of the affiliated metrics for a *single-point* prediction located at the: (a) border of the affiliation zone, (b) position halfway between the border and the event, (c) first element of the event, (d) center of the event.

Expressiveness of the metrics. The affiliation metrics have been built to handle the two main limitations of the classical precision/recall. As for algorithmic comparison, we have shown that they are expressive enough to perform a fair comparison without the need for introducing additional parameters. The modification of the metrics to handle additional features (such as focusing on overlaps only, or giving more importance to the beginning of an event) is possible by modifying the survival functions – but, at the same time, this is not encouraged. The main reason is the additional layer of complexity needed to select those parameters, in the context of small volume of labeled events, would render evaluation arbitrary.

Theoretical properties. The affiliation metrics focus on a single aspect of the evaluation: the assessment of the proximity between the predicted and the ground truth labels, which is the primary aspect to compare anomaly detection algorithms from a research point of view. Summarizations of the precision/recall as a single quantity that exist for the classical metrics (such as the F-score or the Average Precision) are straightforward to derive for the affiliation metrics. In order to properly characterize the behavior of such summarized metrics, further research is needed to gather theoretical bounds on the variance.

Practical deployment. From a deployment perspective, complementary measures need to be considered, similarly to those introduced by Gensler and Sick [6], such as the number of predicted events in each segmentation zone and the direction tendency of the predictions. In this context, the trade-offs between those measures remain in the hand of the field expert. In order to assist decisions (e.g., algorithm selection and score thresholding), further development of an interactive visualization library leveraging the affiliation metrics would be beneficial.

6 CONCLUSION

For evaluating time series anomaly detection tasks, we proposed a precision/recall pair that handles the limitations encountered with the classical metrics. Contrary to the existing metrics, it is generic (parameter-free, applicable on all datasets), and local (each ground truth event is considered separately). In turn, locality makes it both expressive (possible to break down the final score into individual interpretable and visualizable bricks) and robust (e.g., not sensitive to adversary predictions). Finally, its construction makes it both theoretically principled, as well as practically useful – overall, we hope that the research community will find them a useful contribution for the unbiased evaluation of time series anomaly detection tasks.

REFERENCES

[1] Subutai Ahmad et al. 2017. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262, 134–147.
[2] Michel Marie Deza and Elena Deza. 2009. Encyclopedia of distances. In Springer.
[3] Marie-Pierre Dubuisson and Anil K. Jain. 1994. A modified Hausdorff distance for object matching. In *IEEE ICPR*. Volume 1, 566–568.
[4] Osamu Fujita. 2013. Metrics based on average distance between sets. *Japan Journal of Industrial and Applied Mathematics*, 30, 1, 1–19.
[5] Astha Garg et al. 2021. An evaluation of anomaly detection and diagnosis in multivariate time series. *IEEE Trans. on Neural Networks and Learning Systems*.
[6] André Gensler and Bernhard Sick. 2014. Novel criteria to measure performance of time series segmentation techniques. In *LWA*. Citeseer, 193–204.
[7] Shaghayegh Gharghabi et al. 2017. Matrix profile VIII: domain agnostic online semantic segmentation at superhuman performance levels. In *IEEE ICDM*.
[8] Jonathan Goh et al. 2016. A dataset to support research in the design of secure water treatment systems. In *CRITIS*. Springer, 88–99.
[9] Alexis Huet et al. 2022. Affiliation precision/recall library. https://doi.org/10.6084/m9.figshare.19131425. (2022).
[10] Kyle Hundman et al. 2018. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *ACM SIGKDD*, 387–395.
[11] Won-Seok Hwang et al. 2019. Time-series aware precision and recall for anomaly detection: considering variety of detection result and addressing ambiguous labeling. In *ACM CIKM*, 2241–2244.
[12] Nick James et al. 2020. Novel semi-metrics for multivariate change point analysis and anomaly detection. *Physica D: Nonlinear Phenomena*, 412.
[13] Jonguk Kim et al. 2019. Anomaly detection for industrial control systems using sequence-to-sequence neural networks. *CyberICPS*.
[14] Siwon Kim et al. 2021. Towards a rigorous evaluation of time-series anomaly detection. *arXiv preprint arXiv:2109.05257*.
[15] Nikolay Laptev et al. 2015. S5 - a labeled anomaly detection dataset, v1.0 (16M). https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70. (2015).
[16] Alexander Lavin and Subutai Ahmad. 2015. Evaluating real-time anomaly detection algorithms – the Numenta anomaly benchmark. In *IEEE ICMLA*.
[17] Tae Jun Lee et al. 2018. Greenhouse: a zero-positive machine learning system for time-series anomaly detection. *SysML*.
[18] LinkedIn. 2018. Luminol. https://github.com/linkedin/luminol. (2018).
[19] Fei Tony Liu et al. 2008. Isolation forest. In *IEEE ICDM*, 413–422.
[20] Pankaj Malhotra et al. 2015. Long short term memory networks for anomaly detection in time series. In *ESANN*. Volume 89, 89–94.
[21] Hansheng Ren et al. 2019. Time-series anomaly detection service at Microsoft. In *ACM SIGKDD*, 3009–3017.
[22] Erik Scharwächter and Emmanuel Müller. 2020. Statistical Evaluation of Anomaly Detectors for Sequences. In *ACM SIGKDD Workshop on Mining and Learning from Time Series (KDD MiLeTS)*.
[23] Bernhard Schölkopf et al. 2001. Estimating the support of a high-dimensional distribution. *Neural computation*, 13, 7, 1443–1471.
[24] Hyeok-Ki Shin et al. 2020. HAI 1.0: HIL-based augmented ICS security dataset. In *USENIX Workshop on Cyber Security Experimentation and Test*.
[25] Nidhi Singh and Craig Olinsky. 2017. Demystifying Numenta anomaly benchmark. In *IEEE IJCNN*.
[26] Ya Su et al. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *ACM SIGKDD*, 2828–2837.
[27] Nesime Tatbul et al. 2018. Precision and recall for time series. *NeurIPS*.
[28] Charles Truong et al. 2020. Selective review of offline change point detection methods. *Signal Processing*, 167, 107299.
[29] Li Wei et al. 2005. Assumption-free anomaly detection in time series. In *SSDBM*. Volume 5, 237–242.
[30] Renjie Wu and Eamonn Keogh. 2021. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE TKDE*.
[31] Haowen Xu et al. 2018. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In *WWW*, 187–196.

A CLOSED-FORM OF THE AFFILIATION METRICS

For the ground truth event $\text{gt}_j =: [a, b)$ located within the affiliation zone $I_j =: [A, B)$, we let respectively m and M the shortest and largest distance from the event to the borders of the zone:

$$m := \min(a - A, B - b) \quad \text{and} \quad M := \max(a - A, B - b).$$

A.1 Survival functions

The survival function for the precision defined by Eq. 3 is given by $\bar{F}_{\text{precision}_j}(0) = 1$ and, for $d \in (0, M]$:

$$\bar{F}_{\text{precision}_j}(d) = 1 - \frac{|\text{gt}_j| + \min(d, m) + d}{|I_j|}. \tag{10}$$

The explanation is as follows: taking uniformly at random an element within the affiliation zone, the probability to obtain a distance of zero is $|\text{gt}_j|/|I_j|$, while outside the decrease of the survival function is in $2d$ on $(0, m]$ and in d on $(m, M]$, because two elements exist for each distance before m, and only one after. For a ground truth sample $y \in \text{gt}_j$, we let:

$$m_y := \min(y - A, B - y) \quad \text{and} \quad M_y := \max(y - A, B - y).$$

The survival function for the recall is given, for $d \in [0, M_y]$, by:

$$\bar{F}_{y,\text{recall}_j}(d) = 1 - \frac{\min(d, m_y) + d}{|I_j|}. \tag{11}$$

A.2 Closed-form of the integrals

The closed-form of the integrals is straightforward (integral of piecewise linear functions) but include many cases due to the presence of the min and max functions. The expressions are derived in the following paragraphs. All the cases have been considered in the Python implementation [9].

A.2.1 **Integral over the samples for the precision.** For the precision, each predicted interval located within the affiliation zone is cut into three pieces (possibly empty), corresponding to the portion before, within, and after the ground truth event. We consider the case of a single predicted interval pred located after the ground truth event (thus not intersecting it). The distance from the interval to the ground truth event goes from $d_{\min} = \text{pred}_{\text{start}} - b$ to $d_{\max} = \text{pred}_{\text{stop}} - b$. We also consider either the case (a) $d_{\max} \leq m$ or (b) $m \leq d_{\min}$. For the remaining case $m \in (d_{\min}, d_{\max})$, the predicted interval is cut again into two smaller pieces, one verifying (a) and the other (b). The integral over the samples is given by:

$$\mathcal{A} := \int_{x \in \text{pred}} \bar{F}_{\text{precision}_j}(\text{dist}(x, \text{gt}_j))dx = \int_{d_{\min}}^{d_{\max}} \bar{F}_{\text{precision}_j}(z)dz. \tag{12}$$

We replace the survival function using (10). The linear part gives $\int_{d_{\min}}^{d_{\max}} z dz = (d_{\max} - d_{\min})(d_{\max} + d_{\min})/2 = |\text{pred}| d_{\text{center}}$, with d_{center} the distance reached for the point at the middle of the predicted interval. For the two cases (a) and (b), we deduce the same expression:

$$\frac{\mathcal{A}}{|\text{pred}|} = 1 - \frac{|\text{gt}_j| + \min(d_{\text{center}}, m) + d_{\text{center}}}{|I_j|}. \tag{13}$$

A.2.2 **Integral over the samples for the recall.** For the recall, we cut the ground truth interval into a finite partition such that each piece $\text{gt}_{j,k}$ is either fully included in the predictions (and the integral is immediate) or has a unique closest prediction $t_{\text{pivot},k} \in I_j$ located at the border or outside the piece. Furthermore, all elements of a piece should be closer either to the closest prediction or to the border. We consider the case of a closest prediction located after the ground truth event, so that the distance from an element $y \in \text{gt}_{j,k}$ to the predictions is: $t_{\text{pivot},k} - y$.

The integral over the samples is given by:

$$\mathcal{B} := \int_{y \in \text{gt}_{j,k}} \bar{F}_{y,\text{recall}_j}(\text{dist}(y, \text{pred} \cap I_j))dy \tag{14}$$

$$= \int_{y \in \text{gt}_{j,k}} \bar{F}_{y,\text{recall}_j}(t_{\text{pivot},k} - y)dy. \tag{15}$$

We replace the survival function using Eq. 11. We define y_{center} the middle point of $\text{gt}_{j,k}$, $d_{\text{pivot}} = |t_{\text{pivot},k} - y_{\text{center}}|$ its distance to the pivot prediction, and $m_{\text{center}} = \min(y_{\text{center}} - A, B - y_{\text{center}})$ its closest distance to the border. Using those notations, we obtain on the one side the linear part:

$$\int_{y \in \text{gt}_{j,k}} (t_{\text{pivot},k} - y)dy = |\text{gt}_{j,k}| d_{\text{pivot}} \tag{16}$$

and on the other side:

$$\int_{y \in \text{gt}_{j,k}} \min(t_{\text{pivot},k} - y, m_y)dy = |\text{gt}_{j,k}| \min(d_{\text{pivot}}, m_{\text{center}}). \tag{17}$$

Combining those elements, we arrive at the following expression:

$$\frac{\mathcal{B}}{|\text{gt}_{j,k}|} = 1 - \frac{\min(d_{\text{pivot}}, m_{\text{center}}) + d_{\text{pivot}}}{|I_j|}. \tag{18}$$

B PARTICULARIZATION TO POINT ANOMALIES

The case of point anomalies corresponds to express each anomaly at date t as the limit of an event $[t, t + \varepsilon)$ when $\varepsilon \to 0$. It is therefore immediate to restate Eq. 4 as follows (with pred corresponding here to the point predictions within the affiliation zone I_j):

$$P_{\text{precision}_j} = \frac{1}{\#\text{pred}} \sum_{x \in \text{pred}} \bar{F}_{\text{precision}_j}(\text{dist}(x, \text{gt}_j)) \tag{19}$$

and Eq. 5, since gt_j is now a single point, as:

$$P_{\text{recall}_j} = \bar{F}_{\text{gt}_j,\text{recall}_j}(\text{dist}(\text{gt}_j, \text{pred})). \tag{20}$$

The forms of the survival functions do not change (in Eq. 10, the term $|\text{gt}_j|$ is replaced by 0).

C PROOF OF THE PROPERTIES

The generic method for proving the properties is to find a cut of the intervals satisfying the conditions expressed in Appendix A, and to apply (13) and (18).

C.1 Predicting the whole interval as anomalous

The recall is 1 because the ground truth interval is included in the predictions. For the precision, we suppose that $m = a - A$ (the other case is symmetric). The predicted interval $[A, B)$ is cut into $[A, a) \cup [a, b) \cup [b, b+m) \cup [b+m, B)$. The Eq. 13 is applied on each part, giving four areas:

$$\mathcal{A}_{[A,a)} = \frac{(a-A)(B-b)}{B-A}\,,\ \mathcal{A}_{[a,b)} = b-a\,,\ \mathcal{A}_{[b,b+m)} = \mathcal{A}_{[A,a)},$$
$$\mathcal{A}_{[b+m,B)} = \frac{1}{2}\frac{((B-b)-(a-A))^2}{B-A}.$$

It leads to the following expression (corresponding to the Eq. 7):

$$P_{\text{precision}} = \frac{\mathcal{A}_{[A,a)} + \mathcal{A}_{[A,a)} + \mathcal{A}_{[A,a)} + \mathcal{A}_{[A,a)}}{B-A} = \frac{1}{2} + \frac{1}{2}\left(\frac{b-a}{B-A}\right)^2. \tag{21}$$

C.2 Expected precision and recall given a single random prediction

The expected recall is expressed as the average of (5) over $t \in I_j$:

$$\begin{aligned}\mathbb{E}\left[P_{\text{recall}}\right] &= \frac{1}{|\text{gt}_j|}\frac{1}{|I_j|}\int_{t\in I_j}\int_{y\in \text{gt}_j} \bar{F}_{y,\text{recall}_j}(\text{dist}(y,t))dydt \\ &= \frac{1}{|\text{gt}_j|}\frac{1}{|I_j|}\int_y\int_t 1 - \frac{\min(|t-y|, m_y) + |t-y|}{|I_j|}dtdy\end{aligned}$$

where on the second line, we use the Fubini's theorem, which facilitates the computation of the inner integral for each fixed y:

$$\begin{aligned}R(y) &:= \int_t 1 - \frac{\min(|t-y|, m_y) + |t-y|}{|I_j|}dt \\ &= |I_j| - \frac{1}{|I_j|}\left[\int \min(|t-y|, m_y) + |t-y|dt\right] \\ &= |I_j| - \frac{1}{|I_j|}\left[2\int_0^{m_y} 2z dz + \int_{m_y}^{M_y}(m_y + z)dz\right].\end{aligned}$$

For the last line, we observe that for $t \in [y - m_y, y + m_y]$, the distance $|t-y|$ goes from 0 to m_y (and this happens two times, on the left and on the right), while for $t \in [m_y, M_y]$, the distance goes from m_y to M_y. After computing the integral, we end up with a quantity which does not depend on y:

$$R(y) = |I_j| - \frac{1}{2|I_j|}(m_y + M_y)^2 = \frac{|I_j|}{2}, \tag{22}$$

and finally: $\mathbb{E}\left[P_{\text{recall}}\right] = 1/2$. For the expected precision, the computations are identical to those leading to Eq. 21.

C.3 Single prediction at a defined position

To ease the understanding, we consider the case $|I_j| = [A, B) = [0, 1)$. Since $\text{gt}_j = [a, b)$ is centered in the affiliation zone and of proportion p, we have: $[a, b) = [1/2 - p/2, 1/2 + p/2)$. We detail the case (d) of a prediction located at the position $1/2$ and corresponding to Eq. 9.

If $p \leq 1/2$, the cut $[1/2 - p/2, 1/2) \cup [1/2, 1/2 + p/2)$ verifies the necessary conditions. By applying Eq. 18 on each part, we obtain: $\mathcal{B}_{[1/2-p/2,1/2)} = \mathcal{B}_{[1/2,1/2+p/2)} = (p/2)(1 - p/2)$ so that $P_{\text{recall}} = 1 - p/2$.

If $p > 1/2$, the ground truth region is partitioned into four areas: $[1/2 - p/2, 1/4) \cup [1/4, 1/2) \cup [1/2, 3/4) \cup [3/4, 1/2 + p/2)$, from which we compute:

$$\begin{aligned}\mathcal{B}_{[1/4,1/2)} = \mathcal{B}_{[1/2,3/4)} &= (1/4)(3/4), \\ \mathcal{B}_{[1/2-p/2,1/4)} = \mathcal{B}_{[3/4,1/2+p/2)} &= (p/2 - 1/4)(1/2),\end{aligned}$$

so that $P_{\text{recall}} = 1/2 + 1/(8p)$. By combining those two cases, we end up with Eq. 9. Globally, the same method is applied for the positions (a), (b), (c), and we get the curves represented in Fig. 4:

$$\text{(a)}\quad P_{\text{precision}} = 0, \qquad P_{\text{recall}} = \frac{p}{4}, \tag{23}$$

$$\text{(b)}\quad P_{\text{precision}} = \frac{1}{2} - \frac{p}{2}, \qquad P_{\text{recall}} = \frac{1}{2} - \frac{p}{2} + \frac{25}{64p}\left(p - \frac{1}{5}\right)_+^2, \tag{24}$$

$$\text{(c)}\quad P_{\text{precision}} = 1, \qquad P_{\text{recall}} = 1 - p + \frac{16}{9p}\left(p - \frac{1}{3}\right)_+^2, \tag{25}$$

$$\text{(d)}\quad P_{\text{precision}} = 1, \qquad P_{\text{recall}} = 1 - \frac{p}{2} + \frac{1}{2p}\left(p - \frac{1}{2}\right)_+^2. \tag{26}$$

D REPRODUCIBILITY

The affiliation metrics have been implemented using the standard Python 3 library and is available at [9]. The implementation leverages the closed-form highlighted in Appendix A and follows the process of Sec. 3.4: the binary inputs are converted into events before being separated into affiliation zones. On each segment, the metrics are computed and the output consists of the precision/recall scores as well as the individual distances and probabilities.

Reliability of the code has been checked through unit tests. Additionally, the numerical results related to the affiliation metrics obtained in Sec. 4 are directly reproducible by typing the following: `python -m unittest discover`. On a Windows 10 machine using an Intel(R) Core(TM) i7-8650U processor running at 1.90 GHz with 16 GB of RAM, the whole tests took 8 seconds, including the computation of the affiliation metrics for the SWaT [8] dataset containing 449919 samples and 35 ground truth events, against up to 472 predicted events.