ZAAS: Zonal Aware Anomaly Score for Time Series

Nabil Ait Said[1], **Elies Gherbi**[1], **Faouzi Adjed**[1], **Achraf Kallel**[1]

[1] IRT SystemX, 2 Boulevard Thomas Gobert 91120 Palaiseau, France
{nabil.ait-Said, elies.gherbi, faouzi.adjed, achraf.kallel}@irt-systemx.fr

Abstract

Time series anomaly detection plays a critical role across domains from industrial monitoring to cybersecurity use cases. But its evaluation remains challenging. Traditional window level F_1 score overweights long anomaly intervals, while heuristic "point-adjusted" variants introduce bias by extending single detection across entire zones. We propose a Zone Normalized F_1, which treats each true and each predicted anomaly interval as a unit, macro-averaging precision and recall over intervals rather than windows. This eliminates length bias and yields a fairer comparison of detectors. We formalize the metric, illustrate its behavior on toy and real examples, and show how it complements existing protocols.

Introduction

In industrial sectors such as cybersecurity, manufacturing, energy, healthcare, and financial services, anomaly detection is essential for maintaining operational integrity and efficiency. Evaluating AI models for anomaly detection in these fields requires careful consideration of various factors, including the potential impact of detected anomalies and the required interventions.

The choice of evaluation metrics plays a crucial role in assessing model performance, yet inappropriate metric selection can significantly distort the assessment of a model's performance, potentially leading to severe operational consequences. A poorly chosen metric might overestimate the model's accuracy, creating a false sense of security while critical anomalies remain undetected. As a result, the reliability and trustworthiness of the anomaly detection system are compromised, affecting decision-making processes and potentially leading to a decline in overall system performance and safety.

These metric selection challenges become even more pronounced when dealing with temporal data streams. Anomaly detection in time-series data presents unique challenges that amplify these metric selection concerns. Standard binary-classification metrics, such as the Receiver Operating Characteristic (ROC) curve and the F_1-Score, are widely used to quantify detector performance under severe class imbalance (Bhattacharya et al. 2024; Kim et al. 2022). However, as we see in Figure 1 the contiguous nature of anomalies in time series complicates the direct application of the pointwise F_1 score: a single long event spans many time points, and a full detection of a large anomaly over-rewards the model (F1 score = 0.84), while a partial detection can penalize it (F1 score = 0.20). Heuristic "point adjustment" (PA) protocols remedy the partial detection problem by marking an entire ground-truth segment as correct if any point is flagged, but they bias results toward the inflation of true positives and can hide differences between models (Bhattacharya et al. 2024).

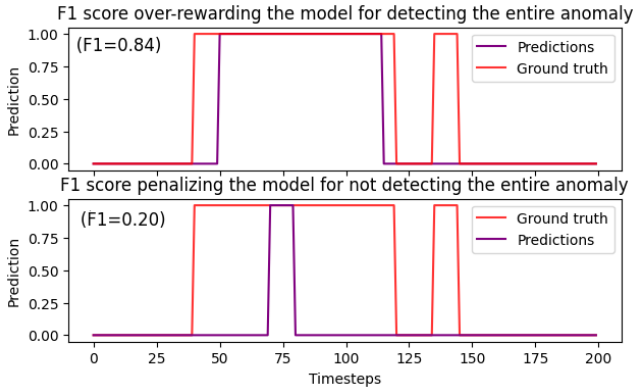

Figure 1: Comparison of two anomaly detection behaviors within the anomaly zone. The red line shows the true anomalous region, and the purple line shows the model's predicted anomaly. In the top graph, the model achieves an F1 score of 0.84, whereas in the bottom graph the score drops to 0.20.

Recent work has highlighted these metric limitations, where even random anomaly scores can achieve high PA-adjusted F_1 (F_1PA), leading to misleading model rankings (Kim et al. 2022). Alternatives, such as k-percent adjustment (F_1KPA), partially address bias but lack theoretical guarantees, and existing metrics can fail to penalize false positives appropriately (Bhattacharya et al. 2024).

In this paper, we introduce the *Zonal Aware Anomaly Score* (ZAAS) for time series, along with a matching evaluation protocol that avoids the biases of prior adjustments. ZAAS treats each detected anomaly "zone" holistically, balancing rewards for true positives against penalties for false

positives for fair evaluation and practical awareness. Concretely, our contributions are:

- **Zonal-aware detection.** We define "zones" of anomalies around each detected point and establish a scoring rule that (a) treats the detection of part of a real anomaly as valid for the entire anomaly, and (b) imposes penalties on isolated, erroneous detections to ensure a more discriminating evaluation of event-level performance.
- **Practical evaluation.** Through experiments on synthetic and public datasets, we show ZAAS remains below chance for purely random scores and yields consistent monotonic ordering across detectors, unlike F_1 and F_1PA.

Methodology and Proposal

In high-frequency time series anomaly detection, the primary objective is to flag each anomaly exactly when it occurs, triggering timely interventions in applications such as network intrusion detection. Once an anomaly zone has been entered and an alert raised, subsequent flags within the same zone provide little additional value, since the incident is already known. Consequently, the ideal detector should register at least one positive detection within each true anomaly zone, and the evaluation metric must capture this requirement. However, practical constraints such as limited model memory, sliding window lengths, and computational budgets often force deployment on windows shorter than the full duration of an anomaly zone. Under these circumstances, both the standard pointwise F_1 and the PA point adjusted F_1 become biased by the length of the anomaly and by repeated detections within the same zone (Kim et al. 2022). To this end, in the following section we formalize the core components of our scoring function, which together enable the definition of the Zonal Aware Anomaly Score (ZAAS).

We model a time series of length N divided into windows $w_1, \ldots, w_N$, each with true label $y_j \in \{0, 1\}$ and binary prediction $\hat{y}_j \in \{0, 1\}$.

$$y_j = \begin{cases} 1, & \text{if } w_j \text{ overlaps an anomaly zone,} \\ 0, & \text{otherwise,} \end{cases}$$

and a binary prediction

$$\hat{y}_j = \begin{cases} 1, & \text{model predicts anomaly in } w_j, \\ 0, & \text{otherwise.} \end{cases}$$

Let M be the number of true anomaly zones and K the number of predicted zones :

$Z_i = \{j : y_j = 1 \text{ in the } i\text{-th anomaly interval}\}, i = 1, \ldots, M$

True anomaly windows form M disjoint *zones* $Z_1, \ldots, Z_M$, each of length

$$L_i = |Z_i| = \sum_{j=1}^{N} \mathbf{1}\{j \in Z_i\}, \quad \sum_{i=1}^{M} L_i = \sum_{j=1}^{N} y_j. \quad (1)$$

We group the predictions into K predicted zones

$\hat{Z}_k = \{j : \hat{y}_j = 1 \text{ in the } k\text{-th contiguous run}\}, k = 1, \ldots, K.$

Standard Window-Level F_1

Define

$$\mathrm{TP} = \sum_{j=1}^{N} \mathbf{1}\{\hat{y}_j = 1 \wedge y_j = 1\} \quad (2)$$

$$\mathrm{FP} = \sum_{j=1}^{N} \mathbf{1}\{\hat{y}_j = 1 \wedge y_j = 0\} \quad (3)$$

$$\mathrm{FN} = \sum_{j=1}^{N} \mathbf{1}\{\hat{y}_j = 0 \wedge y_j = 1\} \quad (4)$$

Precision (P), recall (R), and F_1 are given by equation (5) below.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad F_1 = \frac{2PR}{P + R}. \quad (5)$$

One drawback is that anomaly zones of different lengths L_i dominate window level counts and penalize the model for false negatives in longer zones even if it detects anomalies windows within those zones. In the standard F1 calculation, adjacent windows w_j, in the anomaly zone, that the model labels as normal ($\hat{y}_j = 0$) are counted as false negatives, so the score ignores the fact that any successful detection is still valuable in practice. As a result, the model's ability to flag at least one anomalous window in a long anomaly zone isn't properly rewarded when L_i is larger compared to window to detection.

Point-Adjusted F_1

Point-adjusted F_1 is a widely used heuristic that credits an entire contiguous anomaly segment/zone as detected if any single point within it is correctly flagged (Xu et al. 2018).

That succinct line defines a per-zone "detection flag" δ_i as follows:

$$\delta_i = \max_{j \in Z_i} \hat{y}_j \quad \in \{0, 1\} \quad (6)$$

Where Z_i is the set of all window indices j belonging to the i-th true anomaly zone, and $\hat{y}_j \in 0, 1$ is the model's binary prediction for window j with $1 =$ anomaly and $0 =$ normal).

Given that each $\hat{y}_j$ receives 0 or 1 based on the maximum over $j \in Z_i$, is equivalent to evaluate if the model predicts any window in zone i as anomalous or not. Therefore, true and false positives, and true and false negatives are adapted by integrating the result of δ_i given in equation (6) as follows.

$$\mathrm{TP}' = \sum_{i=1}^{M} L_i \, \delta_i \quad (7)$$

$$\mathrm{FN}' = \sum_{i=1}^{M} L_i \, (1 - \delta_i), \quad (8)$$

$$\mathrm{TP}' = \sum_{i=1}^{M} L_i \, \delta_i \quad (9)$$

$$\mathrm{FN}' = \sum_{i=1}^{M} L_i \,(1 - \delta_i), \tag{10}$$

Then the adjusted F1 uses these new quantities as follows:

$$P' = \frac{\mathrm{TP}'}{\mathrm{TP}' + \mathrm{FP}'}, \quad R' = \frac{\mathrm{TP}'}{\mathrm{TP}' + \mathrm{FN}'}, \quad F_1' = \frac{2P'R'}{P' + R'}.$$

This removes FNs inside detected zones but still weights by zone length. Although PA reduces penalties for partial detections, it overestimates performance in long intervals and is still biased by interval length.

Zonal Aware Anomaly Score

To eliminate length bias, we treat each zone equally, by following the next three steps.

1. Per-Predicted-Zone Precision For each predicted zone $\hat{Z}_k$, we define the indicator function P_k given in equation (11), which equals to 1 if the k-th predicted interval overlaps at least one true anomaly zone, and 0 otherwise (a false alarm).

$$P_k = \mathbf{1}_{\left\{\hat{Z}_k \cap (Z_1 \cup \cdots \cup Z_M) \neq \emptyset\right\}} \in \{0, 1\}. \tag{11}$$

Classic precision at the zone level counts true positives ($\sum_k P_k$) over all predictions (K):

$$P = \frac{\sum_{k=1}^{K} P_k}{\sum_{k=1}^{K} P_k + \left(K - \sum_{k=1}^{K} P_k\right)} = \frac{1}{K} \sum_{k=1}^{K} P_k. \tag{12}$$

We denote this *macro-averaged precision* by

$$P_{\mathrm{macro}} = \frac{1}{K} \sum_{k=1}^{K} P_k. \tag{13}$$

2. Per-True-Zone Recall

Symmetrically, for each true anomaly zone we define the function R_i by the equation (14) below. It equals to 1 if the model predicted at least one positive within zone i, and 0 otherwise.

$$R_i = \mathbf{1}\{Z_i \cap \{j : \hat{y}_j = 1\} \neq \emptyset\} \in \{0, 1\}. \tag{14}$$

where, $Z_i = \{j : y_j = 1 \text{ in zone } i\}, \quad i = 1, \ldots, M,$

Classic recall at the zone level is

$$R = \frac{\sum_{i=1}^{M} R_i}{\sum_{i=1}^{M} R_i + \left(M - \sum_{i=1}^{M} R_i\right)} = \frac{1}{M} \sum_{i=1}^{M} R_i, \tag{15}$$

so we write the *macro-averaged recall*

$$R_{\mathrm{macro}} = \frac{1}{M} \sum_{i=1}^{M} R_i. \tag{16}$$

3. Zonal Aware Anomaly Score

Finally, we combine these macro-averaged measures in the harmonic mean to obtain the event-level F_1:

$$F_{1,\mathrm{zone}} = \frac{2\, P_{\mathrm{macro}}\, R_{\mathrm{macro}}}{P_{\mathrm{macro}} + R_{\mathrm{macro}}}. \tag{17}$$

This metric treats each anomaly event and each predicted event equally, eliminating evaluation bias from interval duration.

Experimental Results

Benchmark datasets

We utilized three distinct datasets in our experiments, Pooled Server Metrics (PSM) (Abdulaal, Liu, and Lancewicki 2021), Soil Moisture and Ocean Salinity (SMAP) (Hundman et al. 2018), and a generated TOY dataset.

The PSM public dataset was gathered internally from various application server nodes at eBay, anonymized, and made available with this study. It comprises 26 features, excluding localization meta-attributes, which were omitted due to anonymization requirements. These features detail server machine metrics, such as CPU utilization and memory usage. The dataset includes a training set spanning 13 weeks, followed by an eight-week testing period. Anomalies are present in both the training and testing datasets, with labels provided only for the testing set. These labels were manually created by engineers and application experts and may include both planned and unplanned anomalies.

The SMAP dataset originates from real-world data collected by a NASA spacecraft. It comprises anomaly data extracted from an Incident Surprise Anomaly (ISA) report, specifically for a spacecraft monitoring system. This type of dataset is crucial for understanding and analyzing unexpected events or anomalies that occur during space missions.

The toy dataset serves as an additional dataset that we have created specifically for the purpose of evaluating and comparing the performance of various models.

Benchmark models

In our evaluation, we tested two models using the TimeADDM method (Hu et al. 2024). The first model employs an approach with 50 diffusion steps. In contrast, the second model adopts a more complex strategy by combining the predictions obtained from multiple diffusion steps configurations, specifically 50, 100, 500, and 1000 steps.

Results

The results are summarized in Tables 1 and 2, which report the performance of the models under different evaluation metrics. In both tables, the bold values highlight for each metric (original, PA and ZAAS) and dataset, the best F1 score obtained across the models.

Table 1 presents the scores calculated by the standard metrics and the adjusted metrics on the PSM and SMAP datasets. The adjusted metrics indicate that the models achieve very good performance. In contrast, the standard metrics reveal that the models have poor performance. Additionally, the ZAAS metrics indicate that the models exhibit poor performance, but not to the same extent as the standard metrics. This discrepancy arises because the ZAAS metrics do not penalize the model if a true sequence of anomalies is not detected multiple times, providing a more lenient evaluation.

Figure 2 illustrates a toy example of ground truth and predictions. The purple color represents the predictions of the first model, the yellow color represents the predictions of the

Dataset	PSM	PSM	SMAP	SMAP
Model	**1**	**2**	**1**	**2**
Original				
Precision	0.376	0.435	0.43	0.54
Recall	0.048	0.015	0.162	0.044
F1	**0.086**	0.029	**0.24**	0.081
PA				
$\text{Precision}_{\text{PA}}$	0.919	0.977	0.804	0.956
$\text{Recall}_{\text{PA}}$	0.913	0.833	0.887	0.824
F1_{PA}	**0.916**	0.899	0.844	**0.885**
ZAAS(Our)				
Precision	0.349	0.377	0.339	0.495
Recall	0.208	0.139	0.236	0.167
F1	**0.26**	0.203	**0.278**	0.249

Table 1: Metrics comparison on PSM and SMAP datasets

Dataset	TOY	TOY
Model	**1**	**2**
Original		
Precision	0.607	0.579
Recall	0.144	0.093
F1	**0.233**	0.161
PA		
$\text{Precision}_{\text{PA}}$	0.879	0.826
$\text{Recall}_{\text{PA}}$	0.677	0.322
F1_{PA}	**0.765**	0.463
ZAAS(Our)		
Precision	0.70	0.80
Recall	0.20	0.80
F1	0.31	**0.80**

Table 2: Metrics comparison on toy datasets

second model, and the red color indicates the ground truth. As shown in Table 2, unlike the other metrics, the ZAAS metric clearly favors the second model, which detects each anomaly zone even if the predicted sequence is short. Furthermore, we observe that the standard metrics classify the second model as poor (with an F1 score of 0.16), despite the fact that it correctly identifies 4 out of the 5 true sequence anomalies. Moreover, we note that using point adjustment leads to the first model being classified as good ($\text{F1} = 0.765$), despite correctly identifying only 1 out of the 5 true sequence anomalies.

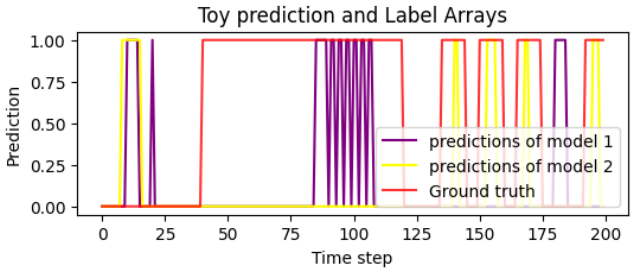

Figure 2: Toy model predictions and labels

Conclusions

In this paper, we have addressed the critical challenge of evaluating time-series anomaly detection. Traditional evaluation metrics, such as window-level F1 scores and heuristic "point-adjusted" variants, often introduce biases that can mislead the assessment of detector performance. Specifically, these metrics can penalize the model for not entirely predicting long anomaly intervals or inflate true positives by extending single detections across entire zones.

To remedy these issues, we introduced the Zone Normalized F1 metric, which treats each true and predicted anomaly interval as a unit. By macro-averaging precision and recall over intervals, our proposed metric eliminates length bias and provides a fairer comparison of detectors. We formalized this metric and illustrated its behavior using both toy and real-world examples, demonstrating its effectiveness in complementing existing evaluation protocols. Future work will focus on exploring potential extensions and adaptations to address additional challenges in time-series anomaly detection.

Acknowledgements

This work has been supported by the French government under the " France 2030 " program, as part of the SystemX Technological Research Institute within the JNI5 project.

References

Abdulaal, A.; Liu, Z.; and Lancewicki, T. 2021. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2485–2494.

Bhattacharya, D.; Mukherjee, S.; Kamanchi, C.; Ekambaram, V.; Jati, A.; and Dayama, P. 2024. Towards Unbiased Evaluation of Time-series Anomaly Detector. *arXiv preprint arXiv:2409.13053*.

Hu, R.; Yuan, X.; Qiao, Y.; Zhang, B.; and Zhao, P. 2024. Unsupervised anomaly detection for multivariate time series using diffusion model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 9606–9610. IEEE.

Hundman, K.; Constantinou, V.; Laporte, C.; Colwell, I.; and Soderstrom, T. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 387–395.

Kim, S.; Choi, K.; Choi, H.-S.; Lee, B.; and Yoon, S. 2022. Towards a Rigorous Evaluation of Time-series Anomaly Detection. *AAAI Conference on Artificial Intelligence*.

Xu, H.; Feng, Y.; Chen, J.; Wang, Z.; Qiao, H.; Chen, W.; Zhao, N.; Li, Z.; Bu, J.; Li, Z.; et al. 2018. Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*, 187–196.