Towards Unbiased Evaluation of Time-series Anomaly Detector

Debarpan Bhattacharya*
Indian Institute of Science
Bangalore, India
debarpanb@iisc.ac.in

Sumanta Mukherjee
IBM Research
Bangalore, India
sumanm03@in.ibm.com

Chandramouli Kamanchi
IBM Research
Bangalore, India
Chandramouli.Kamanchi@ibm.com

Vijay Ekambaram
IBM Research
Bangalore, India
vijaye12@in.ibm.com

Arindam Jati
IBM Research
Bangalore, India
Arindam.Jati@ibm.com

Pankaj Dayama
IBM Research
Bangalore, India
pankajdayama@in.ibm.com

Abstract

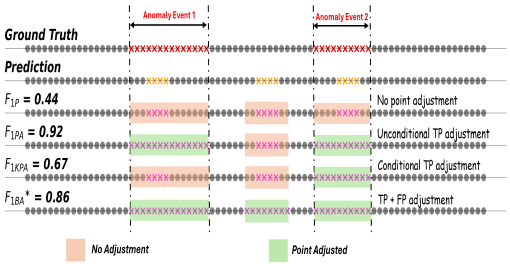
Time series anomaly detection (TSAD) is an evolving area of research motivated by its critical applications, such as detecting seismic activity, sensor failures in industrial plants, predicting crashes in the stock market, and so on. Anomalies are rare events, making the F1-score the most commonly adopted metric for anomaly detection. However, in time series the challenge of using standard F1-score is the dissociation between 'time points' and 'time events'. To accommodate this, anomaly predictions are adjusted, called point adjustment (PA), before the F_1-score evaluation. However, these adjustments are heuristics-based, and biased towards true positive detection, resulting in over-estimated detector performance. However, the current time-series foundation model literature continues to use PA for model evaluation. Such obtained model perspectives are not a true indication of the performance. This work proposes an alternative adjustment protocol called "Balanced point adjustment" (BA). It addresses the limitations of existing point adjustments and provides fairness guarantees backed by axiomatic definitions of TSAD evaluation.

1 Introduction

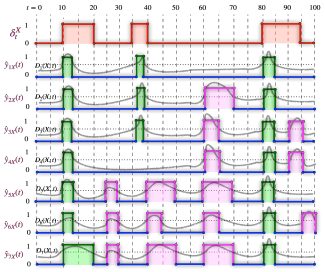
Anomaly detection plays a crucial role in identifying system failures or performance deviations. Anomalies are rare data patterns, often detected by comparing the likelihood of an instance with respect to the background distribution. As a result, anomaly and outlier detection are closely related, with applications spanning tabular, image, and audio data. The detection process involves classifying observations as either anomalous or normal, making it similar to binary classification. Consequently, binary classification metrics, such as the receiver operating characteristics (ROC)(1) and *F1 score*(2), are commonly used for Time Series Anomaly Detection (TSAD). ROC captures detector behavior across thresholds, while the F_1 score is crucial for assessing a detector's performance, especially in the setting of imbalanced data with low anomaly ratios (3).

*This work was carried out during his stay at IBM Research, Bangalore, India for IBM Research Global Internship Program 2024.

NeurIPS 2024 Workshop on Time Series in the Age of Large Models.



(a) A comparative view of different point adjustment methods. F_{1KPA} has $K = 40\%$. F_{1BA} is proposed in this paper. F_{1BA} is the only metric that penalizes false positives. The orange highlights detection which is left as it is, and green highlights describe instances that are adjusted before F_1 score computation.



Different F_1 adjusted metrics

TP	FP	FN	F_{1P}	F_{1PA}	F_{1PAK}	F_{1BA}
3	0	0	0.5	1.00	0.84	1.00
3	1	0	0.40	0.86	0.80	0.76
3	2	0	0.40	0.86	0.80	0.66
2	2	1	0.28	0.76	0.28	0.58
2	3	1	0.24	0.62	0.24	0.46
2	4	1	0.26	0.67	0.26	0.45
2	3	1	0.42	0.62	0.42	0.46

(b) Comparison of F_{1p}, F_{1PA}, F_{1KPA}, and our proposed F_{1BA}. In the table, the green color shows ideal metric values for perfect detection (no false detections), while red highlights the failure to indicate correct predictions. The proposed F_{1BA} consistently makes meaningful transitions, unlike other metrics.

An anomaly detection system typically has two components: the scorer and the detector (4). The scorer processes signals into a score, and the detector determines a threshold for detecting anomalies. The ROC curve informs threshold selection, while the *F1 score* evaluates the detector's overall performance.

Applying the standard F_1-score to time series presents challenges due to the contiguous nature of anomalies. Point adjustment (PA), introduced by (5), modifies predictions before evaluating the F_1 score to account for this. However, PA often biases results in favor of true positives, leading to inflated performance estimates (6). The state-of-the-art time series foundation models still use F_{1PA} for model evaluation (7; 8; 9). A detailed discussion on the related works is provided in Section A.1.

We introduce a new protocol, 'Balanced Point Adjustment' (BA), to address these biases. BA penalizes false positives and balances the adjustments made for true positives, providing a more accurate and fair evaluation of TSAD models. This method, as shown in Figure 1a, ensures a more reliable assessment of anomaly detectors, supported by controlled experiments.

2 Methods

2.1 Notations

Time-series: Let's consider univariate time-series sample space $\mathcal{X}$. Time series X of length T can be sampled from $\mathcal{X}$: $X \in \mathcal{X}$, so that $X = (x_1, x_2, \ldots, x_T)$. **Anomaly labels**: For a given time series X, anomaly labels is a time series $\delta^X = (\delta_1^X, \delta_2^X, \cdots, \delta_T^X)$ with $\delta_t^X \in \{0, 1\}\ \forall t$. **Anomaly segment**: An anomaly segment is a contiguous subsequence of timesteps corresponding to an anomaly event. We define an i^{th} anomaly segment occurring in X as: $\mathcal{A}_i^X := (a_i, w_i, s_i)$ where a_i, w_i and s_i denote the starting timestamp, time-width, and severity of the i^{th} anomaly segment respectively. Note that,

$$\delta_t^X = \begin{cases} 1 & t \in \cup_i \{a_i, a_{i+1} \cdots a_{i+w_i-1}\} = \cup_i S_a^i \\ 0 & \text{otherwise} \end{cases}$$

where S_a^i denotes the set of time-steps corresponding to an anomaly event, $\{a_i, a_{i+1} \cdots a_{i+w_i-1}\}$. **Anomaly detector**: An anomaly detector $\mathcal{D}(t, X)$ labels a timestep t in X as an anomaly point if $D(t, X) > \gamma$ with $t \in [0, T-1]$ and γ being a threshold.

Metric for time-series anomaly detection: Let the metric to quantify the anomaly detection performance of an anomaly detector $\mathcal{D}(\cdot, \cdot)$ is $\mathcal{M}(y, \delta)$, where $y(t) = D(t, X)$, $\delta \in \{0, 1\}^T$, $y \in \mathbb{R}^T$.

2.2 Illustrative analysis

We provide analysis of the different metrics concerning diverse scenarios with respect to True Positives (TP), False Positives (FP), False Negatives (FN) numbers in Figure 1b. The table in the right panel of Figure 1b studies different metric values for diverse scenarios of TP, FP, and FN events. Firstly, F_{1p} and F_{1KPA} fail to indicate the perfect detection $\hat{y}_{1X}(t)$ by detector D_1. Further, with varying number of TP, FP, and FN events, the expected transitions (increase/decrease) of different metrics are also studied. We observe, that the proposed metric F_{1BA} makes consistent transitions with excellent coverage.

2.3 Axiomatic criterion for TSAD metrics

In TSAD, a detector requires a reliable metric for accurate comparison with other detectors. We formalize the essential requirements of a TSAD metric: (a) The metric should be resistant to random noise and uncorrelated data (6), (b) it should reward better detectors with higher scores, and (c) it should grant the best score exclusively to the best detection performance. To enable use of standard mathematical tools, we analyze the detector scores $y = D(X, t)$ directly, rather than binary predictions derived from thresholding.

A TSAD metric $\mathcal{M}(y, \delta)$ is said to be a *good* metric if it satisfies the following axioms:

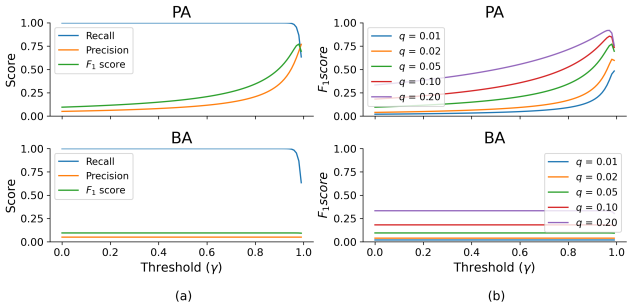


Figure 2: (a) The behavior of BA metrics P_{BA}, R_{BA}, F_{1BA} compared to PA metrics for scores from uniform noise with varying thresholds γ, using anomaly width of 100 and ratio $q = 0.2$. F_{1PA} rises above 0.75 for random anomaly scores, (b) The right panel illustrates the behavior of F_{1PA} and F_{1BA} with varying γ for different anomaly ratios (q). F_{1PA} increases with higher thresholds, while F_{1BA} remains unaffected by threshold choice.

- ***C-1 (robust)***: For any random detection signal y_{random} ($\perp\!\!\!\perp\ \delta$), the metric value should always be less than its chance level value ($\mathcal{M}_{chance} = 0.5$ for F_1 score based metrics),
$$\mathcal{M}(y_{random}, \delta) \leq \mathcal{M}_{chance} \tag{1}$$
- ***C-1a (threshold agnostic)***: If $y_{random} \sim \mathcal{U}[0, 1]$ uniformly distributed noise, then $\mathcal{M}(y_{random}, \delta)$ should remain unaffected by threshold variation.
$$\mathcal{M}(y_{random}, \delta) \to \mathcal{M}_u, \forall\, \gamma \tag{2}$$
Note that $\mathcal{M}_u$ is a constant ($\mathcal{M}_u \leq \mathcal{M}_{chance}$).
- ***C-2 (ordered)***: For two detectors $D_1(\cdot, \cdot)$ and $D_2(\cdot, \cdot)$ with discriminability order $D_1 \geq D_2$, the following holds,
$$\mathcal{M}(y_1, \delta) \geq \mathcal{M}(y_2, \delta) \tag{3}$$
where $y_1 = D_1(t, X)$ and $y_2 = D_2(t, X)$.
- ***C-3 (exclusive)***: The metric should reach maximum value for a perfect detection signal (y_{prf}) with no FP and FN events ($\mathcal{M}(y_{prf}, \delta) \to \mathcal{M}_{max}$), and should fail to converge if a single FP/FN occurs.
$$\mathcal{M}(y_{prf} + \{1 \text{ FP/FN}\}) \to \mathcal{M}_{max} - \epsilon, \epsilon > 0 \tag{4}$$

2.4 Metric analysis

We formally show the following:

- The point adjusted (PA) F_1 score leads to overestimation of precision and can give very high F_1 score for random anomaly scores, uncorrlated with anomaly labels, as shown in the Figure 2. Hence, F_{1PA} violates many of the above axioms.
- In contrary, F_{1BA} stays low for random noisy scores (less than 0.5, which is chance level value of F_1 score) as shown in Figure 2, and obeys all the above axioms.

The theoritical proof of the above is detailed in the appendix A.2.

3 Experiments

We provide a detailed empirical evaluation of F_{1BA} against other F1 scores through controlled experiments. We have implemented a controlled experimental setup that allows generating anomaly label sequences and detector predictions with controlled number of anomaly events, TP, FP, and detector score quality. The experimental setup and data preperation details are provided in appendix A.3. We study the scaling relation of F_1 scores with the 4 most important anomaly detector characteristics- *precision*, *recall*, *detection coverage*, and *separation score* (anomaly vs non-anomaly score).

3.1 F1 scores vs separation score with varying recall

The behaviour of different metrics against separation score is shown in Figure 3 for three different recall ranges, separately. Separation score indicates the discriminability between detector anomaly scores in anomalous vs non-anomalous regions as discussed in A.3. Firstly, all the 4 metrics show an increasing trend with separation. The increase in the separation score suggests better anomaly identification with a constant threshold, which results in an increasing F1 value. However, we note the following three observations: (a) The F_{1P} is always very low, even for a high recall and high separation, (b) for low recall of $< 25\%$ (and precision maintained between $25\%-75\%$), F_{1PA} and F_{1KPA} achieve very high score close to 0.8, which indicates biased behaviour of the metrics, while F_{1BA} is least affected as it penalizes false positives. (c) for high recall, F_{1BA} and F_{1PA} converges, indicating applicability of F_{1PA} at only high recall setting.

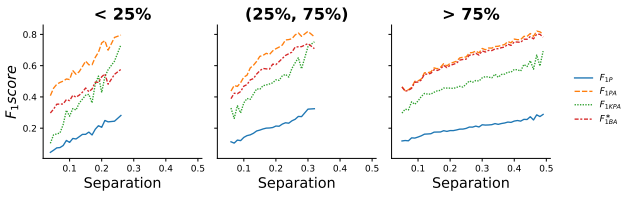
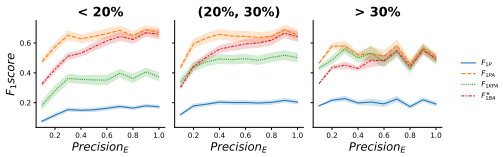


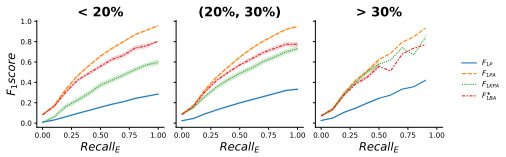
Figure 3: Metric behavior against the score separation (A.3). Plots are made for varying recall A.3 in 3 different bins of $< 25\%, (25\% - 75\%)$ and $> 75\%$. The bins are chosen so that similar data point cardinality is maintained. Precision is maintained within $(25\% - 75\%)$.

3.2 F1 scores vs precision with varying coverage

The behaviour of different F1 scores against precision is shown in Figure 4a for three different ranges of coverage, separately with medium recall value $(25\% - 75\%)$. We make the following remarks: (a) Among all the metrics, the F_{1BA} shows the highest sensitivity (high slope) to precision which shows the stricter compliance to the *ordering* axiom (equation 3), (b) The F_{1PA} is an overestimate at the low precision and the F_{KPA} is an underestimate at high precision. F_{1BA} behaves like F_{1KPA} at low precision and like F_{1PA} at high precision. Note that all the metrics show drop in score for high coverage and high precision as we maintain medium recall while generating the samples.



(a) Metric behavior plotted against the precision (A.3). Plots are made for varying coverage score A.3 in 3 different bins of $< 20\%, (20\% - 30\%)$ and $> 30\%$. Recall is maintained within $(25\% - 75\%)$.



(b) Metric behavior plotted against the recall (A.3). Plots are made for varying coverage score A.3 in 3 different bins of $< 20\%, (20\% - 30\%)$ and $> 30\%$. Precision is maintained within $(25\% - 75\%)$.

3.3 F1 scores vs recall with varying coverage

The behaviour of different metrics against recall is shown in Figure 4b for three different ranges of coverage, separately with medium precision value $(25\% - 75\%)$. We note the following: (a) Because of the maintained precision range, the metric scores should not cross 0.75 even for the highest recall, as the F_1 score is a harmonic mean of precision and recall. However, F_{1PA} approaches 1.0 for all coverage values. Similar to F1 scores vs separation score case (Figure 3), this arises from inappropriate penalization of false positives, overestimating the precision, (b) F_{1KPA} and F_{1P} continue to behave as underestimate.

4 Conclusion & Future Work

We proposed a new F_1 score (F_{1BA}) and motivated it with an illustrative examples and axiomatic criterions. Additionally, we developed a simulation setup for controlled metric comparison to demonstrate the efficacy of our proposed score. We believe this contribution will aid both the applied and research communities by facilitating a more systematic and unbiased approach to anomaly model selection. We continue to pursue this work to show the effectiveness of the balanced adjustment across domains.

Acknowledgments and Disclosure of Funding

The work was done at IBM Research, Bangalore during IBM Research Internship Program 2024.

References

[1] James A Hanley and Barbara J McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[2] V Rijsbergen and C Keith Joost, "Information retrieval butterworths london," *Google Scholar Digital Library*, 1979.

[3] Santonu Sarkar, Shanay Mehta, Nicole Fernandes, Jyotirmoy Sarkar, and Snehanshu Saha, "Can tree based approaches surpass deep learning in anomaly detection? a benchmarking study," *arXiv preprint arXiv:2402.07281*, 2024.

[4] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[5] Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, et al., "Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 187–196.

[6] Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon, "Towards a rigorous evaluation of time-series anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 7194–7201.

[7] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al., "One fits all: Power general time series analysis by pretrained lm," *Advances in neural information processing systems*, vol. 36, pp. 43322–43355, 2023.

[8] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski, "Moment: A family of open time-series foundation models," in *Forty-first International Conference on Machine Learning*, 2024.

[9] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," *arXiv preprint arXiv:2210.02186*, 2022.

[10] Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich, "Precision and recall for time series," *Advances in neural information processing systems*, vol. 31, 2018.

[11] Sondre Sørbø and Massimiliano Ruocco, "Navigating the metric maze: A taxonomy of evaluation metrics for anomaly detection in time series," *Data Mining and Knowledge Discovery*, vol. 38, no. 3, pp. 1027–1068, 2024.

[12] Jiehui Xu, "Anomaly transformer: Time series anomaly detection with association discrepancy," *arXiv preprint arXiv:2110.02642*, 2021.

[13] Zekai Chen, Dingshuo Chen, Xiao Zhang, Zixuan Yuan, and Xiuzhen Cheng, "Learning graph structures with transformer for multivariate time-series anomaly detection in iot," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9179–9189, 2021.

[14] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga, "Usad: Unsupervised anomaly detection on multivariate time series," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 3395–3404.

[15] Hao Zhou, Ke Yu, Xuan Zhang, Guanlin Wu, and Anis Yazidi, "Contrastive autoencoder for anomaly detection in multivariate time series," *Information Sciences*, vol. 610, pp. 266–280, 2022.

[16] Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun, "Tfad: A decomposition time series anomaly detection architecture with time-frequency analysis," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 2497–2507.

[17] Bo Wu, Chao Fang, Zhenjie Yao, Yanhui Tu, and Yixin Chen, "Decompose auto-transformer time series anomaly detection for network management," *Electronics*, vol. 12, no. 2, pp. 354, 2023.

[18] Kukjin Choi, Jihun Yi, Jisoo Mok, and Sungroh Yoon, "Self-supervised time-series anomaly detection using learnable data augmentation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[19] Astha Garg, Wenyu Zhang, Jules Samaran, Ramasamy Savitha, and Chuan-Sheng Foo, "An evaluation of anomaly detection and diagnosis in multivariate time series," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2508–2517, 2021.

[20] Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu C Aggarwal, and Mahsa Salehi, "Carla: Self-supervised contrastive representation learning for time series anomaly detection," *Pattern Recognition*, p. 110874, 2024.

[21] Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi, "Deep learning for time series anomaly detection: A survey," *ACM Computing Surveys*, 2022.

[22] Wenkai Li, Cheng Feng, Ting Chen, and Jun Zhu, "Robust learning of deep time series anomaly detection models with contaminated training data," *arXiv preprint arXiv:2208.01841*, 2022.

[23] Yueyue Yao, Jianghong Ma, Shanshan Feng, and Yunming Ye, "Svd-ae: An asymmetric autoencoder with svd regularization for multivariate time series anomaly detection," *Neural Networks*, vol. 170, pp. 535–547, 2024.

[24] A.W. van der Vaart, *Asymptotic Statistics*, Asymptotic Statistics. Cambridge University Press, 2000.

A Appendix

A.1 Related Works

Time series anomalies often span multiple time points, and partial detection is typically considered valid (10). To accommodate this, the *F1 score* for time series anomaly detection (TSAD) is computed after point adjustment. Point adjustment extends a detection across the anomaly span, which is then used in the *F1 score* calculation (5). However, this introduces bias by overestimating model accuracy, as even random scores can produce favorable results (11; 6). (6) proposed a stricter adjustment based on a $k\%$ threshold, though its improvements over the original point adjustment remain unclear.

There is ambiguity regarding the appropriate choice of the metric for TSAD due to the shortcomings of pointwise F_1 (F_{1p}), biased point-adjusted F_1 (F_{1PA}), and heuristic-based corrections like F_{1KPA}.

- Many studies still use F_{1PA} despite its limitations (12; 13; 14; 15; 16; 17).
- F_{1PA} has also been used extensively in the recent large time-series foundation model evaluation, viz, GPT4TS (7), Moment (8), TimesNet (9).
- Some studies have adopted F_{1KPA} for TSAD evaluation (18; 19).
- A few revert to using pointwise F_1 (F_{1p})(20; 21; 22; 23), while others combine all three metrics(18).

A.2 Metric analysis

A.2.1 Pre-requisites

To allow analysis of the proposed metric and existing ones, we first need their closed-form expressions. We find out the limiting expressions of F_{1PA} and F_{1BA} first.

Definition 1 (**Point adjustment**)**.** The point adjustment procedure involves the following steps:

- **Thresholding**:

$$\hat{y}_X(t) = \begin{cases} 1 & \text{if } D(X,t) > \gamma \\ 0 & \text{if } D(X,t) \leq \gamma \end{cases} \tag{5}$$

- **Point-adjustment (PA)**:

$$\hat{y}_{X,PA}(t) = \begin{cases} 1 & \text{if } \hat{y}_X(t) = 1 \\ 1 & \text{if } t \in S_a^i \text{ and } \exists\, t' \in S_a^i \text{ s.t. } \hat{y}_X(t') = 1 \\ 0 & \text{elsewhere} \end{cases} \tag{6}$$

- $\mathbf{F_{1PA}}$ **score**:

$$F_{1PA} = F_1(\text{true} = \delta, \text{prediction} = \hat{y}_{X,PA}(t)) \tag{7}$$

In contrary, our proposed Balanced Adjustment (BA) procedure can be detailed as:

Definition 2 (**Balanced Adjustment (BA)**)**.** The BA involves:

- **Thresholding**: Based on Eqn. 5.
- **BA**:
  We define islands of width w_N at time u around FPs as:

$$S_N(u) := \{k : u - \frac{w_N}{2} - 1 \leq k \leq u + \frac{w_N}{2}, \\ \hat{y}_X(u) = 1,\, \delta(u) = 0\} \tag{8}$$

  Clearly, $|S_N(u)| = w_N, \forall u$ and it is defined only at FP time steps. Using the islands $S_N(u)$, we define adjusted prediction:

$$\hat{y}_{X,BA}(t) = \begin{cases} 1 & \text{if } \hat{y}_X(t) = 1 \\ 1 & \text{if } t \in S_a^i \text{ and } \exists\, t' \in S_a^i \text{ s.t. } \hat{y}_X(t') = 1 \\ 1 & \text{if } \exists\, u,\, t \in S_N(u) \\ 0 & \text{elsewhere} \end{cases} \tag{9}$$

- $\mathbf{F_{1BA}}$ **score**:

$$F_{1BA} = F_1(\text{true} = \delta, \text{prediction} = \hat{y}_{X,BA}(t)) \tag{10}$$

Now, we obtain expressions of F_{1PA} and F_{1BA} based on definitions.

Theorem 1 ($\mathbf{F_{1PA}}$ in random noise). *The point-adjusted (PA) F1 score (F_{1PA}) of any random time-series anomaly detector working on a sufficiently large time series of length T having a single anomaly event ($S_A := S_a$) is:*

$$F_{1PA} = \frac{2q\left(1 - N(\gamma)^{|S_a|}\right)}{(1 - N(\gamma)) + q\left(1 + N(\gamma) - N(\gamma)^{|S_a|}\right)} \tag{11}$$

where $q = \frac{|S_a|}{T}$ is the anomaly ratio, $N(\cdot)$ is the noise cdf.

Proof. For sufficiently long time-series ($T \to \infty$),

$$\begin{aligned} R_{PA} &= Pr\left(\hat{y}_{X,PA}(t) = 1 \,|\, \delta(t) = 1\right) \\ &= 1 - \prod_{t \in S_a} Pr\left(\hat{y}_X(t) \leq \gamma\right) \\ &\quad (\because \hat{y}_X(t_i) \perp\!\!\!\perp \hat{y}_X(t_j), i \neq j \text{ as } \hat{y}_X \sim N(\text{noise})) \\ &= \left(1 - N(\gamma)^{|S_a|}\right) \end{aligned} \tag{12}$$

$$\begin{aligned} P_{PA} &= Pr\left(\delta(t) = 1 \,|\, \hat{y}_{X,PA}(t) = 1\right) \\ &= \frac{Pr\left(\hat{y}_{X,PA}(t) = 1 \,|\, \delta(t) = 1\right) Pr\left(\delta(t) = 1\right)}{Pr\left(\hat{y}_{X,PA}(t) = 1\right)} \\ &= \frac{q\left(1 - N(\gamma)^{|S_a|}\right)}{(1 - N(\gamma)) + q(N(\gamma) - N(\gamma)^{|S_a|})} \end{aligned} \tag{13}$$

The expression of F_{1PA} follows. □

Theorem 2 ($\mathbf{F_{1BA}}$ in random noise). *The balanced point-adjusted (BA) F1 score (F_{1BA}) of any random time-series anomaly detector working on a sufficiently large time series of length T having a single anomaly event ($S_A := S_a$) is:*

$$F_{1BA} = \frac{2q\left(1 - N(\gamma)^{|S_a|}\right)}{(1 - N(\gamma)^{w_N}) + q\left(1 + N(\gamma)^{w_N} - N(\gamma)^{|S_a|}\right)} \tag{14}$$

Proof. Assume that the minimum separation between false positive predictions in $\hat{y}_X(t)$ is more than the island width (w_N). Now, R_{PA} remains unaltered as in PA.

$$\begin{aligned} P_{BA} &= Pr\left(\delta(t) = 1 \,|\, \hat{y}_{X,BA}(t) = 1\right) \\ &= \frac{q(1 - N(\gamma)^{|S_a|})}{(1 - N(\gamma)^{w_N}) + q(N(\gamma)^{w_N} - N(\gamma)^{|S_a|})} \end{aligned} \tag{15}$$

The expression of F_{1BA} follows. Interestingly, expressions hold same structure as in Eqn. 11 with additional exponentials of w_N. □

The Figure 2 studies the F_{1PA} and F_{1BA} for uniformly randomly drawn noise as anomaly score. It shows that F_{1PA} increases by threshold selection and can give a very high score as well. However, F_{1BA} score is unaffected by threshold choice and remains below 0.5.

A.2.2 *C-1 (robust)*

Lemma 2.1. For $w_N = |S_a|$, the F_{1BA} is always less than equal to the chance level $F1$ score of 0.5 for any randomly generated anomaly score, as long as the anomaly ratio (q) is less than equal to $\frac{1}{3}$.

$$F_{1BA}(\delta, \hat{y}_{X,BA}(t)) \leq F1_{chance} = 0.5,\ X \sim N(\cdot)\, \forall N(\cdot) \tag{16}$$

Proof. Assuming the anomaly event is sufficiently sustained (not momentary) so that the anomaly width ($|S_a|$) is not very small,

$$\begin{aligned} F_{1BA} &= \frac{2q(1-N(\gamma)^{|S_a|})}{(1-N(\gamma)^{|S_a|})+q} \,(\text{using } |S_a| = w_N) \\ &= \frac{2q}{1+q} (\because |S_a| >> 1 \implies N(\gamma)^{|S_a|} \to 0) \end{aligned} \tag{17}$$

Now, $\frac{2q}{1+q} \leq 0.5 \implies q \leq \frac{1}{3} = 33.33\%$ □

A.2.3 *C-1a (threshold agnostic)*

Lemma 2.2. If an anomaly scorer generates random score from uniform distribution, the F_{1BA} not only stays less than the chance value, but also behaves as threshold agnostic (stays constant across all possible thresholds γ) for almost the entire range of thresholds as long as the anomaly width is not very small.

$$\begin{aligned} F_{1BA}(\delta, \hat{y}_{X,BA}(t)) \leq F1_{chance} = 0.5,\ X \sim \mathcal{U}[0,1] \\ \frac{\partial F_{1BA}(\delta, \hat{y}_{X,BA}(t))}{\partial \gamma} \to 0, \forall \gamma \in (\gamma_{min}, \gamma_{max}) \end{aligned} \tag{18}$$

Proof. Because lemma 2.1 hold for all $N(\cdot)$, it holds for $\mathcal{U}[0,1]$ too. Now, using $N(\gamma) = \gamma$ for $N(\cdot) := \mathcal{U}[0,1]$,

$$\frac{\partial F_{1BA}}{\partial \gamma} = -\frac{2q^2|S_a|\gamma^{|S_a|-1}}{(1-\gamma^{|S_a|}+q)^2} \tag{19}$$

Note that, F_{1BA} is monotonic w.r.t. γ: $\frac{\partial F_{1BA}}{\partial \gamma} \leq 0$. Further,

$$\frac{\partial F_{1BA}}{\partial \gamma} = -2.\left(\frac{q}{(1+q-\gamma^{|S_a|})}\right)^2 .|S_a|\gamma^{|S_a|-1}$$

with

$$\frac{q}{(1+q-\gamma^{|S_a|})} < 1 \text{ as } |q| < 1,\ |S_a| >> 1 \tag{20}$$

$$|S_a|\gamma^{|S_a|-1} \to 0 \text{ as } |S_a| >> 1 \tag{21}$$

It can be formally shown using limit:

$$\lim_{x \to \infty} x.b^{x-1} = 0,\ 0 < b < 1 \tag{22}$$

Hence, $\frac{\partial F_{1BA}}{\partial \gamma} \to 0$, implying F_{1BA} remains constant across γ. □

A.2.4 *C-2 (ordered)*

Lemma 2.3. Let, for any detector $D(t, X)$ has the following parameters:

$$\begin{aligned} \alpha = Pr(\hat{y}_X(t) \leq \gamma | \delta(t) = 1) \\ \beta = Pr(\hat{y}_X(t) \leq \gamma | \delta(t) = 0) \end{aligned} \tag{23}$$

Then, for two different detectors D_1 and D_2 with strength order ($D_1 > D_2$) given by the conditions: $\alpha_{D_1} \leq \alpha_{D_2}$, $\beta_{D_1} \leq \beta_{D_2}$, the following holds,

$$\begin{aligned} F_{1BA}\left(\hat{y}_{X,BA}(t), \delta(t)\right) \leq F_{1BA}\left(\hat{z}_{X,BA}(t), \delta(t)\right) \\ \hat{y}_X(t) = D_1(t, X),\ \hat{z}_X(t) = D_2(t, X) \end{aligned} \tag{24}$$

Proof. For Detector D_i ($i \in \{1, 2\}$):

$$F_{1BA}^{D_i} = \frac{2q(1-\alpha_{Di}^{|S_a|})}{1-\beta_{Di}^{|S_a|}+q(1+\beta_{Di}^{|S_a|}-\alpha_{Di}^{|S_a|})} \tag{25}$$

So, $F_{1BA}(\alpha_1, \beta) \geq F_{1BA}(\alpha, \beta) \forall \alpha > \alpha_1$ and $F_{1BA}(\alpha, \beta) \geq F_{1BA}(\alpha, \beta_2) \forall \beta_2 > \beta$ gives $F_{1BA}(\alpha_1, \beta_1) \geq F_{1BA}(\alpha_2, \beta_2)$. □

A.2.5 *C-3 (exclusive)*

Lemma 2.4. For a perfect anomaly signal y_{prf} that detects the anomaly event correctly and gives no false alarm, and $y_1 = y_{prf} + f_k$ being the score which triggers k false alarms,

$$F_{1BA}(\delta, \hat{y}_{prf,BA}(t)) = 1 \tag{26}$$

$$F_{1BA}(\delta, \hat{y}_{1,BA}(t)) = 1 - \epsilon,\ \epsilon > 0 \tag{27}$$

Proof. By definition, in absence of any FP/FN in $\hat{y}_{prf,BA}(t)$, $F_{1BA}(\delta, \hat{y}_{prf,BA}(t)) = 1$. Now, for a tiny FP event of width $w_{fp} << |S_a|$, $P_{PA} = \frac{|S_a|}{|S_a|+w_{fp}} \to 1$, $P_{BA} = \frac{|S_a|}{|S_a|+w_{fp}+w_N} \to 1 - \epsilon,\ \epsilon > 0$. As $R_{PA} = R_{BA} = 1$, $F_{1BA} \to 1 - \epsilon,\ \epsilon > 0$. Note, $F_{1PA} \to 1$. □

A.3 Data preparation

We develop a simulation tool for data preparation. The simulator generates the controlled anomaly sequences in two steps. First, it produces the ground truth label (δ^X), with constraints on the total number of anomalies, the width of an anomaly event, and the separation between two successive events. The second step yields a controlled generation of the detector scores $y = D(t, X)$. The second step consists of three sub-steps, (1) latent detection, (2) score simulation, and (3) threshold-driven anomaly marking.

Many measurements are computed from simulated labels, along with *F1 scores*. Let N and M represent the number of anomaly events in the ground truth and the predictions respectively. m^s_{tp} denotes the number of true events that are detected, and n^s_{tp} is the number of true positives in prediction. It must be noted that $m^s_{tp} \leq n^s_{tp}$, as a single anomaly event can overlap with multiple detections. c_i is the anomaly detection strength of i^{th} event.

We derive the essential attributes from the simulated samples that are important for our study of F_1 metric scores.

Separation Score It is defined as the distribution distance (Hellinger distance (24)) between the regular and anomalous scores. $H(P, Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}\left(\sqrt{p_i} - \sqrt{q_i}\right)^2}$, where P and Q are discrete probability distribution over regular and anomalous region scores. P and Q are evaluated only over the overlapped score values, using non-parametric kernel density estimates. A higher value signifies a better anomaly detector.

Precision We define the precision metric at an event label as Precision$_E$ = $\frac{m^s_{tp}}{M}$. The precision uses predicted labels only.

Recall Similar to precision, recall is computed at an event level with the ground-truth label, Recall$_E$ = $\frac{n^s_{tp}}{N}$.

Coverage This metric measures the average fraction of the detected true events, $\frac{1}{n^s_{tp}}\sum_{i=1}^{N} c_i$.

We have conducted $15,000$ distinct controlled simulated experiments with varying TP, FP, and total anomaly count. We uniformly sample coverage, and separation score, using parameter grid search. F_{1KPA} computation uses $K = 20\%$.