Anomalous Agreement: How to find the Ideal Number of Anomaly Classes in Correlated, Multivariate Time Series Data

Ferdinand Rewicki[1,2], **Joachim Denzler**[2] **Julia Niebling**[1],

[1]Institute of Data Science, German Aerospace Center, 07745 Jena, Germany
[2]Computer Vision Group, Friedrich-Schiller University, 07743 Jena, Germany
ferdinand.rewicki@dlr.de, joachim.denzler@uni-jena.de, julia.niebling@dlr.de

Abstract

Detecting and classifying abnormal system states is critical for condition monitoring, but supervised methods often fall short due to the rarity of anomalies and the lack of labeled data. Therefore, clustering is often used to group similar abnormal behavior. However, evaluating cluster quality without ground truth is challenging, as existing measures such as the Silhouette Score (SSC) only evaluate the cohesion and separation of clusters and ignore possible prior knowledge about the data. To address this challenge, we introduce the Synchronized Anomaly Agreement Index (SAAI), which exploits the synchronicity of anomalies across multivariate time series to assess cluster quality. We demonstrate the effectiveness of SAAI by showing that maximizing SAAI improves accuracy on the task of finding the true number of anomaly classes K in correlated time series by 0.23 compared to SSC and by 0.32 compared to X-Means. We also show that clusters obtained by maximizing SAAI are easier to interpret compared to SSC.

Code — https://gitlab.com/dlr-dw/saai

1 Introduction

Detecting and classifying abnormal system states is crucial for effective monitoring and control of complex systems. Unfortunately, supervised classification approaches fall short because anomalies are by definition rare, and especially for real-world applications, no or very limited labeled data is available. Therefore, clustering is used to derive groups of similar anomalous behavior from unlabeled data (Sohn et al. 2023; Rewicki et al. 2024a). Assessing the quality of a given solution obtained by applying clustering algorithms such as K-means (MacQueen et al. 1967) to the anomalous subsequences or inferred features is challenging for a number of reasons: (a) no ground truth is available to determine the quality of a solution, (b) the true number of clusters in the data is usually unknown, (c) the solution is highly dependent on the chosen embedding in a feature space (Rewicki et al. 2024a; Raihan 2023). Furthermore, classical unsupervised cluster quality measures such as the Silhouette Score (SSC) (Rousseeuw 1987) evaluate the cohesion within and the separation between clusters but do not incorporate any prior knowledge about the data. To address this challenge in the case of multivariate time series consisting of sufficiently similar signals, we investigate the following research question: How can we exploit the similarity between signals when clustering anomalies found in these variables? The SAAI is based on the principle, that anomalies found simultaneously (i.e., synchronously) in several similar variables within a multivariate time series should belong to the same class. In this work, we deliver evidence on the effectiveness of this measure and show, that maximising SAAI is superior compared to maximizing SSC and the X-Means algorithm (Pelleg, Moore et al. 2000), a variant of K-Means that determines the ideal value for K. Our contributions are:

1. We derive the SAAI, an internal measure of the quality of anomaly clusters.
2. We justify the effectiveness of SAAI by showing that SAAI outperforms SSC and X-Means on the task of finding the true number of classes K. We show that the results obtained by using SAAI are highly correlated with those obtained by using the Adjusted Rand Index (ARI) (Hubert and Arabie 1985) and the Fowlkes Mallows Index (FMI) (Fowlkes and Mallows 1983), two external cluster quality measures that require ground truth labels.
3. We show that the clusterings obtained from maximizing SAAI are easier to interpret compared to SSC.

The rest of the paper is organized as follows: We start with discussing related work in Section 2. In Section 3 we derive the SAAI using an illustrative example and introduce the synthetic dataset. In Section 4 we present the experimental setup and results, which we discuss in Section 5. Finally, in Section 6 we conclude and give an outlook on future work.

2 Related Work

The Silhouette Score, introduced in (Rousseeuw 1987), is the standard measure for evaluating clustering results and quantifies both, cohesion and separation within clusters. It is calculated by averaging the silhouette coefficients SSC_{C_j} for each cluster C_j, defined as

$$SSC_{C_j}(C_j) = \frac{1}{|C_j|} \sum_{\mathcal{S} \in C_j} \frac{idist(\mathcal{S}) - wdist(\mathcal{S})}{\max(wdist(\mathcal{S}), idist(\mathcal{S}))}. \quad (1)$$

The measure $wdist(\mathcal{S})$ is the average distance of the object $\mathcal{S} \in C_j$ to all other elements within its own cluster C_j

(within-cluster distance), while $idist(\mathcal{S})$ is the smallest average distance to elements in another cluster $C_i \neq C_j$ (inter-cluster distance). The SSC ranges from -1 to 1, where 1 indicates well-separated clusters, 0 indicates overlapping clusters, and -1 indicates misclassification of objects. As SSC does only use information obtained from the clustering process, it is referred to as an *internal* measure.

The Fowlkes-Mallows Index (Fowlkes and Mallows 1983) is an external measure of the similarity between two clusterings C_i and C_j, i.e. two partitions of a finite set of objects. An external measure uses information obtained from outside the clustering process, e.g. ground truth class labels. The FMI is the geometric mean of the product of Precision and Recall, ensuring a balanced evaluation of the two quantities. It also defines a scalar product on the space of pairs of data points (Ben-Hur, Elisseeff, and Guyon 2002). We use FMI to measure the similarity between a clustering and the ground truth class assignments. The FMI is defined as

$$FMI = \sqrt{\frac{TP}{TP + FP}\frac{TP}{TP + FN}}\ , \tag{2}$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives. The FMI ranges from 0 to 1, where $FMI = 0$ indicates absolute disagreement and $FMI = 1$ perfect similarity of $\mathcal{C}_i$ and $\mathcal{C}_j$. In the context of comparing a predicted to a true clustering, a true positive is a pair of points that belongs to the same cluster in both, the predicted and the true solution. A false positive is a pair of points that belongs to the same cluster in the predicted, but to different clusters in the true clustering. False negative and true negative are defined accordingly.

The Rand Index (RI) (Rand 1971) is another external measure of the similarity between two clusterings and represents the ratio of correct decisions to all decisions. The RI is defined as

$$RI = \frac{TP + TN}{TP + FP + FN + TN}\ , \tag{3}$$

where TP and TN are the number of true positives and true negatives and FP and FN are the number of false positives and false negatives. Hubert and Arabie proposed the Adjusted Rand Index (ARI) in (Hubert and Arabie 1985), which corrects the original RI by accounting for the expected similarity of random cluster assignments, making the measure robust against chance agreement. This correction is achieved by subtracting the expected value of the RI and dividing by its maximum minus the expectation (Hubert and Arabie 1985). ARI ranges from -1 to 1, where an index of 1 represents perfect agreement, -1 perfect disagreement and 0 the expected agreement by chance.

A popular heuristic for finding the ideal number of clusters is the Elbow-method (Kodinariya and Makwana 2013; Purnima Bholowalia 2014; López-Rubio, Palomo, and Ortega-Zamorano 2018). This is a visual method that plots the sum of squared errors of objects to their assigned cluster centers against increasing number of clusters. The value where the curve forms an elbow is selected as the ideal value. However, this method is highly subjective and there is no guarantee that an elbow point can be identified.(Schubert 2023) Therefore, various quantitative methods for selecting the ideal number of clusters have been proposed.

(Dinh, Fujinami, and Huynh 2019) and (Shahapure and Nicholas 2020) propose methods to find the ideal number of clusters by maximizing the SSC. (Dinh, Fujinami, and Huynh 2019) test their approach on categorical data and use a Lin-similarity based measure for categorical data proposed in (Nguyen et al. 2019). (Shahapure and Nicholas 2020) evaluate their approach on continuous non-time series data. Both works find that maximizing the SSC yields the correct number of classes in their experiments.

Raihan proposes a method for finding the ideal number of clusters for time series datasets using a symbolic pattern forest algorithm in (Raihan 2023). The experiments show, that SSC fails on finding the correct number of clusters when working with raw time series. Although this observation is presented without an explanation, it underpins our observation of the shortcomings of SSC, esp. in combination with raw time series.

Pelleg, Moore et al. proposed X-Means in (Pelleg, Moore et al. 2000), a variant of K-Means that does not need the number of clusters be selected in advance but finds it by optimizing either the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC). To compare our proposed method to X-Means, we use a time series compatible version of X-Means, in which we replaced the euclidean distance with the elastic distance measure Merge-Split-Merge (MSM) (Stefan, Athitsos, and Das 2013) to compare time series of unequal length.

While maximizing SSC has been shown to work well for non-time series data to find the ideal number of clusters, its usefulness for clustering raw time series is questionable. This leaves a gap for different approaches, which we contribute to fill with this work.

In our earlier work (Rewicki et al. 2024a), we proposed a methodology for deriving anomaly types from unlabeled time series data and introduced the SAAI in an informative way. This work is intended to provide evidence of its usefulness and performance.

3 Methodology

In the following section we present the methodology of this study. We start with deriving the SAAI and give an illustrative example. Afterwards we introduce the synthetic dataset, which we use in our experiments to justify the proposed measure.

Synchronized Anomaly Agreement Index (SAAI)

To evaluate the quality of clustering results, we introduce the **Synchronized Anomaly Agreement Index (SAAI)**. The rationale behind this measure is to use prior knowledge about the signals of a multivariate time series. Assuming sufficiently correlated signals, synchronized, i.e. temporally aligned, anomalies in different channels should be assigned to the same cluster, as they are likely to represent the same anomaly type.

We begin with the basic definitions:

Definition 1 *The regular* ***time series*** $\mathcal{T}$ *of length* $N \in \mathbb{N}$ *is defined as the set of pairs* $\mathcal{T} = \{(t_n, \mathbf{x_n}) | t_n \leq t_{n+1}, 0 \leq n \leq N-1, t_{n+1} - t_n = c\}$ *with* $\mathbf{x_n} \in \mathbb{R}^D$ *being the data points with* D *behavioral attributes and* $t_n \in \mathbb{N}, n \leq N$ *being the equidistant timestamps with distance* c *to which the data refer. For* $D = 1$, $\mathcal{T}$ *is called univariate, and for* $D > 1$, $\mathcal{T}$ *is called multivariate.*

Since we discuss the matter of clustering anomalous subsequences of a time series, we define a subsequence as a connected subset of $\mathcal{T}$:

Definition 2 *The* ***subsequence*** $\mathcal{S}_{a,b} \subseteq \mathcal{T}$ *of the time series* $\mathcal{T}$*, with length* $L = b - a + 1 > 0$ *is given by* $\mathcal{S}_{a,b} := \{(t_n, \mathbf{x_n}) | 0 \leq a \leq n \leq b \leq N-1\}$*. For multivariate time series* $\mathcal{T}$, $\mathcal{S}^{(i)}_{a,b}$ *with* $i \in \mathbb{N}$ *refers to the subsequence* $\mathcal{S}_{a,b}$ *in dimension* $1 \leq i \leq D$ *For brevity, we often omit the indices and refer to arbitrary subsequences as* $\mathcal{S}$.

We continue with the definition of anomalies and synchronized anomalies:

Definition 3 *Given the time series* $\mathcal{T}$ *with* $D > 1$ *and a subsequence* $\mathcal{S}^{(i)}_{a,b}$*, the set* A *of* ***univariate, anomalous subsequences*** *is given as*

$$A := \{\mathcal{S}^{(i)}_{a,b} | i, a, b \in \mathbb{N}, i \leq D, a < b, s(\mathcal{S}^{(i)}_{a,b}) \geq \theta_s\}\,, \quad (4)$$

with $s(\cdot) : \{\mathcal{S} | \mathcal{S} \subseteq \mathcal{T}\} \rightarrow [0, 1]$ *being an anomaly score function and* $\theta_s \in [0, 1]$ *being the threshold to classify a subsequence* $\mathcal{S}^{(i)}_{a,b}$ *as anomalous.*

Definition 4 *Let* $\mathcal{S}^{(i)}_{a_i,b_i}, \mathcal{S}^{(j)}_{a_j,b_j}$ *with* $i < j$ *be two univariate subsequences in two different dimensions of the multivariate time series* $\mathcal{T}$*. We say* $\mathcal{S}^{(i)}_{a_i,b_i}$ *and* $\mathcal{S}^{(j)}_{a_j,b_j}$ *are* ***synchronized****, if they overlap by more than a threshold* θ*:*

$$\omega(a_i, b_i, a_j, b_j) := \frac{min(b_i, b_j) - max(a_i, a_j)}{max(b_i, b_j) - min(a_i, a_j)} \geq \theta \quad (5)$$

with $\theta \in [0, 1]$.

The set of all synchronized, univariate, anomalous subsequences A_S *is given as*

$$\begin{aligned} A_S := \{(\mathcal{S}^{(i)}_{a_i,b_i}, \mathcal{S}^{(j)}_{a_j,b_j}) | \mathcal{S}^{(i)}_{a_i,b_i}, \mathcal{S}^{(j)}_{a_j,b_j} \in A, \\ i < j, \omega(a_i, b_i, a_j, b_j) \geq \theta\} \end{aligned} \quad (6)$$

Definition 5 *Let* A_S *be the set of all synchronized, univariate, anomalous subsequences of time series* $\mathcal{T}$*. The subset* $A^*_S \subseteq A_S$ *of* ***synchronized anomalies that agree on their cluster*** *is given as*

$$A^*_S = \{(\mathcal{S}^{(i)}, \mathcal{S}^{(j)}) | (\mathcal{S}^{(i)}, \mathcal{S}^{(j)}) \in A_S, c(\mathcal{S}^{(i)}) = c(\mathcal{S}^{(j)})\}\,, \quad (7)$$

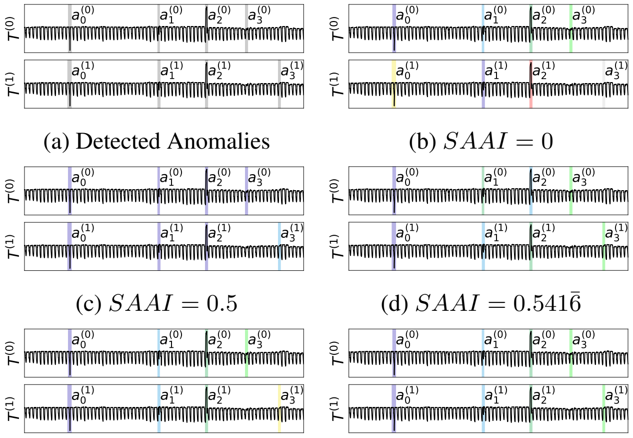
with $c(\mathcal{S})$ *denoting the cluster of subsequence* $\mathcal{S}$

Definition 6 *Given the set of synchronized, univariate anomalous subsequences* A_S *and the set of synchronized, univariate anomalous subsequences in the same cluster* A^*_S*, the number of clusters* K *and the number of pseudo-clusters containing only a single element* $n_{\mathbb{1}}$*, the* ***Synchronized Anomaly Agreement Index (SAAI)*** *is defined as:*

$$SAAI := \lambda \frac{|A^*_S|}{|A_S|} + (1 - \lambda) \frac{K - 1 - n_{\mathbb{1}}}{K}\,, \lambda \in [0, 1]\,. \quad (8)$$

Here, the first term $\frac{|A^*_S|}{|A_S|}$ *evaluates the ratio of synchronized anomalies in the same cluster to all synchronized anomalies. The second term serves as a regularization to account for small cluster sizes (*$\frac{1}{K}$*) and clusters containing only a single anomaly (*$\frac{n_{\mathbb{1}}}{K}$*). It also ensures that the SAAI is in* $[0, 1]$*. The parameter* λ *allows to adjust the influence of the regularizer on the main term.*

The algorithm to calculate the SAAI, its complexity analysis and further information on selecting λ can be found in the Appendix.



(a) Detected Anomalies

(b) $SAAI = 0$

(c) $SAAI = 0.5$

(d) $SAAI = 0.541\bar{6}$

(e) $SAAI = 0.7$

(f) $SAAI = 0.875$

Figure 1: (a) the detected anomalies $a^{(i)}_j$ and (b) - (f) different clustering solutions with increasing quality. Cluster assignment is coded by color. (b) Worst case: all but one cluster contain a single element, (c) all but one anomaly assigned to the same cluster, (d) synchronized anomalies not in the same cluster, (e) synchronized anomalies in separate clusters, pseudo-clusters exist, (f) best case: synchronized anomalies in separate clusters, no pseudo-cluster.

Example Figure 15 illustrates how the SAAI is calculated for different clusterings. In the following example we use $\lambda = 0.5$ and $\theta = 0.5$. A high resolution version of Figure 15 can be found in the appendix. Figure 1a shows the two time series $\mathcal{T} = \{T^{(1)}, T^{(2)}\}$ and the detected anomalies $A = \{a^{(i)}_j | i \in \{0, 1\}, j \in \{0, 1, 2, 3\}\}$. The SAAI is calculated over the set of synchronized anomalies: $A_S = \{(a^{(0)}_0, a^{(1)}_0), (a^{(0)}_1, a^{(1)}_1), (a^{(0)}_2, a^{(1)}_2)\}$, $|A_S| = 3$. The anomalies $a^{(0)}_3$ and $a^{(1)}_3$ are not synchronized and hence $((a^{(0)}_3, a^{(1)}_3)) \notin A_S$.

Figure 1b shows the worst-case solution with the lowest possible SAAI value of 0, where all but two unsynchronized anomalies are assigned to separate clusters. We are aware that this is an extreme edge case, although it is perfectly valid.

Figure 1c shows another extreme case, where all but one unsynchronized anomaly are assigned to the same cluster. Here, in particular, the synchronized anomalies are assigned to the same cluster, which is the goal of the main term in

Equation (8), but the information contained in this clustering is low, which is regularized by the first part of the penalty term $\frac{1}{k}$. The SAAI of this solution is 0.5.

The solution shown in Figure 1d assigned the synchronized anomalies $(a_0^{(0)}, a_0^{(1)})$ to the same cluster, the remaining synchronized anomalies are assigned to different clusters and no pseudo-cluster with only a single element is contained. The SAAI value of this solution is slightly higher with $0.541\overline{6}$ as in 1c.

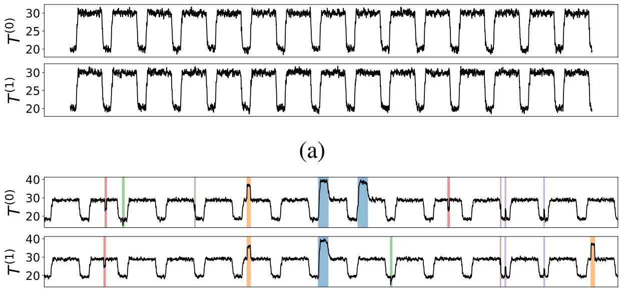
Figure 1e shows the near perfect solution where all synchronized anomalies are assigned to the same cluster, but the different anomaly types are assigned to different clusters. However, two pseudo-clusters are included in this solution, resulting in an SAAI value of 0.7.

The best case solution for this example is shown in Figure 1f, which is similar to the one shown in Figure 1e, but no pseudo clusters are included in this solution, giving a SAAI value of 0.875. The maximal value of $SAAI = 1$ could be reached if we set $\lambda = 0$.

Synthetic Dataset

For the experiments in Section 4, we created synthetic time series similar to temperature measurements from the EDEN ISS (Zabel et al. 2017) Illumination Control System (ICS) in the EDEN ISS 2020 telemetry dataset. EDEN ISS was a research greenhouse for the study of Controlled Environment Agriculture (CEA) techniques and plant growth in (semi)-closed environments, operating between 2018 and 2021 in Antarctica, near the German Neumayer III polar station.

The synthetic time series consists of a periodic signal mimicking the regular ICS temperature signal following the illumination pattern of EDEN ISS with $6h$ night phase at $20°C$, $1h$ warm-up during the simulated sunrise with a small overshoot above the desired daytime temperature of $30°C$. The warm-up is followed by $16h$ of daytime at $30°C$ and finally $1h$ of cool-down. To simulate sensor noise, we add red noise with zero mean, 0.5 standard deviation, and a correlation coefficient of 0.5 to the signal. The basic noisy signal and an example with injected anomalies are shown in Figure 2.



(a)

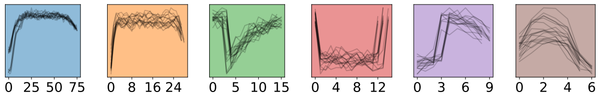
(b)

Figure 2: (a): The basic synthetic ICS signal with simulated sensor noise, (b) The synthetic ICS signal with injected anomalies and $r_{sync} = 0.8$.

To validate the SAAI, we inject 6 different anomaly types into the raw signal, which are shown in Figure 3 and described in Table 1. These anomaly types have also been observed in the real ICS temperature signals as described in (Rewicki et al. 2024a) and are considered to be distinct types of anomalous behavior. Each injected anomaly is subject to randomness with respect to start time, duration, and intensity. Another degree of freedom is the ratio of synchronized to unsynchronized anomalies. This ratio is steered via a parameter r_{sync}. Setting $r_{sync} = 1$ yields a multivariate time series with synchronized anomalies only, while setting $r_{sync} = 0$ yields no synchronized anomaly at all.

Name	**Start Time**	**Dur. (min)**	**Intensity**
Long Day Peak	04:00 - 06:20	240 - 245	$10°C - 11°C$
Short Day Peak	07:00 - 08:20	120 - 125	$8°C - 9°C$
Night Drop	01:00 - 01:40	10 - 11	$-5°C - -4°C$
Day Drop	13:00 - 15:50	60 - 65	$-5°C - -4°C$
Night Peak	01:00 - 01:40	10 - 11	$5°C - 6°C$
Cooldown Peak	22:00 - 22:30	20 - 21	$5°C - 6°C$

Table 1: Parametrization of the six anomaly types, we inject into the synthetic ICS signal.



(a) Long Peak (b) Short Peak (c) Night Drop (d) Day Drop (e) Night Peak (f) Cooldown Peak

Figure 3: The six anomaly types that are injected into the base signal.

4 Experiments and Results

Experimental Setup

We implemented our experiments using Python (version 3.11). For Clustering we used the K-Means implementation in the TSLearn package (Tavenard et al. 2020), which is compatible with subsequences of unequal length. For the same reason, we implemented the X-Means algorithm to use K-Means with an elastic distance measure. We found that we get better results when using Merge-Split-merge (MSM) (Stefan, Athitsos, and Das 2013) compared to using Dynamic Time Warping (DTW) (Vintsyuk 1968; Sakoe and Chiba 1978; Berndt and Clifford 1994) so we compare to the X-Means version with MSM as distance measure. We also noticed a major difference in our results depending on the SSC implementation. We compared that in the TSLearn package to that in scikit-learn (Pedregosa et al. 2011) with the DTW implementation from the aeon-toolkit (Middlehurst et al. 2024) and found that the SSC in scikit-learn yields much better results, even though it is not compatible with unequal length subsequences. Therefore we padded the subsequences with zero to make them equal length. A comparison of these two SSC variants can be found in the

Appendix. All experiments were run on an Intel Xeon Platinum 8260 CPU with 5GB of allocated memory

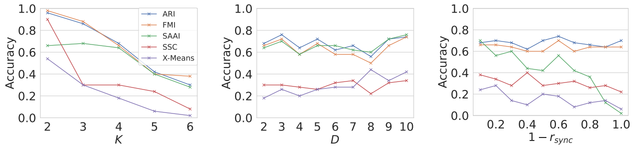
Synthetic Greenhouse Temperature Data

To evaluate SAAI on finding the ideal number of classes K within multivariate time series containing different types of anomalies, we perform the following experiments: We generate a large number of multivariate time series of the synthetic ICS temperature measurements and vary different parameters, namely the number (K) and type of injected anomaly classes, the dimension (D) of the multivariate time series, and the ratio of synchronized to unsynchronized anomalies r_{sync}. We then cluster the anomalous sequences using K-Means clustering with (DTW) as the distance measure. To remove high-frequent noise from the sequences, we apply moving average smoothing with a window size of 5. For each parameter we change, we generate 50 multivariate time series and cluster the anomalous subsequences of each time series with $2 \leq k < 20$, where k is the number of clusters for K-means. We measure how often the correct value K was found by maximizing the internal metrics SAAI and SSC. In addition, we compute the external metrics SAAI and SAAI and use them to find the true number of classes, again by maximizing their respective values. It is fair to complain that with access to the ground truth labels, finding the true number of classes K by maximizing an external metric is pointless. However, we do this for the sake of analyzing the correlation of the internal metrics SSC and SAAI with the external metrics ARI and FMI. As another competitor, we use the X-Means Algorithm (Pelleg, Moore et al. 2000) to determine the ideal value for K.

Increasing K In the first experiment we fix $D = 2$, choose $r_{sync} \in [0.5, 1]$ and increase the number of classes from $K = 2$ to $K = 6$. We run the experiment 50 times for each value of K and select a new value for r_{sync} as well as K new classes on each run uniformly at random. Figure 4a shows the accuracy in finding the true number of classes for increase K. Except for $K = 2$, SAAI shows superior performance compared to SSC. For $K > 3$, SAAI is almost as good as using the external metrics ARI and FMI. X-Means shows the worst results among the compared methods. All methods show decreasing accuracy as K increases.

Increasing D Now, we increase the dimension of the multivariate time series from $D = 2$ to $D = 10$ while fixing $K = 4$ and choosing $r_{sync} \in [0.5, 1]$. We run the experiment 50 times for each value for d and select a new value for r_{sync} as well as new $K = 4$ classes on each run uniformly at random. Figure 4b shows the accuracy in finding the true number of classes K for increasing dimension D. Again, SAAI shows superior performance in finding the true value K compared to SSC and X-Means. Compared to ARI and FMI, SAAI is almost on par with the external metrics for $D < 6$ and even slightly better for $D \geq 6$. Despite the minor variability in accuracies within the results for one method, we also see that the accuracy of finding the true value K is almost independent from the dimension D of the multivariate time series.

Decreasing r_{sync} In the third experiment, we fix $K = 4$ and $D = 2$ and decrease r_{sync} from $r_{sync} = 1$ to $r_{sync} = 0$ in steps of 0.1. We run the experiment 50 times for each value of r_{sync} and select K new classes on each run uniformly at random. Figure 4c shows the accuracy in finding the true number of classes K. As in the previous experiments, SAAI proves to be superior to SSC and X-Means in determining the true value K. Compared to ARI and FMI, the accuracy for SAAI is on par for $1 - r_{sync} < 0.2$. For $0.2 \leq 1 - r_{sync} \leq 0.8$, the accuracy for SAAI shows an decreasing trend as expected, but is still higher than for SSC and X-Means. For $1 - r_{sync} > 0.8$ the accuracy for SAAI falls below that of SSC and X-means The correlation of SAAI and r_{sync} is expected, since r_{sync} determines the proportion of synchronized anomalies in the time series.



(a) Increasing K (b) Increasing d (c) Decreasing r_{sync}

Figure 4: Results of the experiments on synthetic ICS data as described in Section 4. Except for $K = 2$ and $r_{sync} < 0.2$, maximizing SAAI is superior to maximizing SSC. X-Means beats SAAI only for $r_{sync} < 0.2$.

Lagged Variables In the experiments described above, the time series variables were all highly correlated. In Section 3 we derived SAAI for "temporally aligned anomalies in similar measurements". To get an idea of "how similar" the signals of the multivariate time series need to be, we perform the following experiment: We fix $D = 2$, select $r_{sync} \in [0.5, 1]$ and $K = 4$ new classes uniformly at random on every run. We modify the correlation between the variables of the $2D$ time series by increasing the lag l between the first and second dimensions in steps of 60 minutes, from $l = -720$ (-0.5 day) to $l = 720$ ($+0.5$ day). Again, we measure the accuracy of finding the true value K by maximizing SAAI, SSC, ARI, and FMI and by applying X-Means. The results are shown in Figure 5. The black dashed line shows the Pearson correlation coefficient ρ for the two variables of the time series. The baseline, based on random guessing, for finding the correct value of K for $2 \leq k < 20$ is $p = \frac{1}{19}$ and shown as a black dotted line.

For $-180 \leq l \leq 180$, the shaded gray in Figure 5, SAAI achieves superior results than SSC and is again almost on par with ARI and FMI. For X-Means, the accuracy of 0.14 at $l = 180$ is slightly higher than 0.12 for SAAI. The area of $-180 \leq l \leq 180$ corresponds to a correlation of $\rho > 0.43$ and marks the sweet spot for applying SAAI. For $l \geq 360$ maximizing SSC shows better or equal results compared to all other methods.

Summary The Multi Comparison Matrix (MCM) (Ismail-Fawaz et al. 2023) shown in Figure 6 summarizes the results presented before. It shows the Mean Accuracy for the task

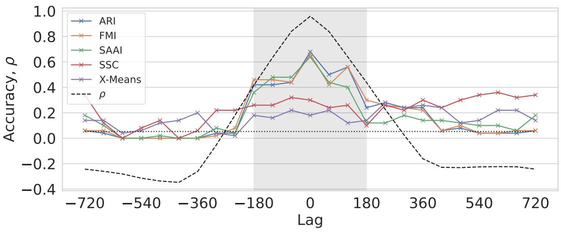


Figure 5: Accuracies for finding the correct value for K while increasing the lag l between the two variables of the time series from -720 minutes to 720 minutes. The Pearson correlation Coefficient ρ is shown as c black dahed line. The gray area between $l = -180$ and $l = 180$ marks the sweet spot for applying SAAI as well as ARI and FMI. In this range, maximizing SAAI achieves superior accuracies compared to SSC. for $l = 180$ the accuracy for X-Means is slighly higher (0.14 vs. 0.12).

of finding the correct value for K of SAAI compared to its competitors. Each cell of the MCM shows the difference in mean accuracy between SAAI and the respective competitor in the top row. The middle row contains the number of wins, ties and losses, where ”win” means, that SAAI achieved a higher accuracy than the respective competitor in one experiment. The bottom row shows the p-value of the Wilcoxon signed rank test (Wilcoxon 1945), which is a non-parametric test used to compare paired samples, without assuming normal distribution of the data. The tested null hypothesis (H_0) is, that the distribution of differences between the paired observations are symmetric around zero. For our case, we can formulate H_0 as: There is no difference in the central tendency between pairs of methods for finding the true value for K. The values in a cell are printed in bold, if the p-value is below 0.05 and hence H_0 is rejected, indicating statistical significance.

As expected, maximizing SAAI is outperformed by maximizing ARI and FMI, which had access to the ground truth labels. Interestingly does this advantage not lead to a significant improvement for FMI. Maximizing SSC and X-Means perform statistically significantly worse than maximizing SAAI.

The correlation plot shown in Figure 7 shows the average correlation coefficient ρ with respect to the experimental results between SAAI and its competitors. Maximizing SAAI shows high to very high correlation (Mukaka 2012) with maximizing ARI and FMI, but only moderate to low correlation (Mukaka 2012) with X-Means and SSC.

Real Greenhouse Temperature Data

In this experiment, we demonstrate the effectiveness of SAAI when working with real, unlabeled data. For this purpose, we use the ICS temperature measurements included in the *edeniss2020* dataset (Rewicki et al. 2024b). This time series consists of 38 temperature measurements from the EDEN ISS research greenhouse. We follow the approach in (Rewicki et al. 2024a) and find anomalous subsequences using the algorithms Maximally Divergent Intervals (MDI) (Barz et al. 2018) and Discord Aware Matrix Profile (DAMP) (Lu et al. 2022) and cluster the anomalous subsequences using K-Means clustering after removing high frequent noise using moving average smoothing with window size $w = 5$. We initialize K-means using kmeans++ initialization (Arthur and Vassilvitskii 2007). Since the anomalous sequences found by MDI and DAMP vary in length, we use DTW as distance metric. We run K-Means with increasing number of clusters k from 2 to 20 and determine the metric-specific optimal number of clusters by maximizing SSC and SAAI, respectively. The optimal clusterings for SAAI and SSC are shown in Figure 8. The detailed results can be found in Figure 13 in the Appendix.

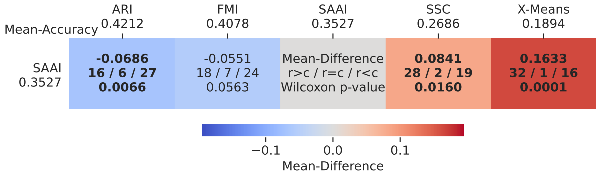


Figure 6: Multi-Comparison Matrix summarizing the results of the experiments in Section 4.

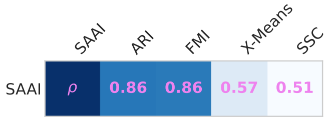
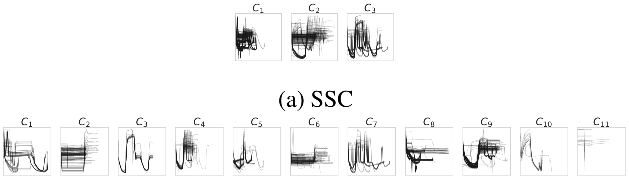


Figure 7: Correlation of the results of SAAI with the competing methods.



(a) SSC

(b) SAAI

Figure 8: Clustering solutions selected by maximizing (a) SSC, and (b) SAAI. The results obtained by maximizing SAAI yield a better clustering in terms of visual interpretability and anomaly type determination.

Maximizing SSC yields 3 clusters, while maximizing SAAI yields 11 clusters, which is much closer to the 10 anomaly types identified in (Rewicki et al. 2024a) for this time series. The cluster solution identified by maximizing SAAI is also easier to interpret. While we can hardly identify any anomaly types by visual inspection in the 3-cluster solution returned by maximizing SSC, we can identify at least six anomaly types in the 11-cluster solution found by maximizing SAAI. These anomaly types are *Peak (short)* (C_1), *Missing Night Phase* (C_2, C_6), *Peak (long)* (C_3), *Anomalous Day Phase* (C_9), *Decreasing Peaks* (C_{10}) *Near Flat Noisy or*

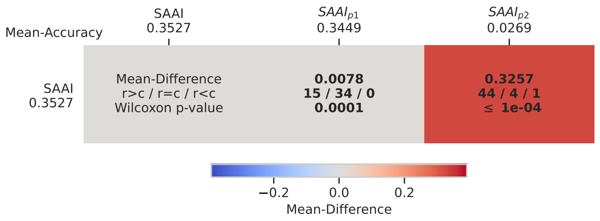


Figure 9: MCM comparing the proposed SAAI with versions using only the first or second penalty term.

Flat signal (C_{11}).

Ablation Study

In our ablation study, we evaluate the contribution of the penalty terms $\frac{1}{K}$ and $\frac{n_{\mathbb{1}}}{K}$ in Equation (8). We perform the same experiments as described in Section 4. The results are summarized in Figure 9. SAAI refers to the SAAI given in Definition 6, while SAAI$_{p1}$ and SAAI$_{p2}$ refer to the SAAI with only the first and second penalty terms, respectively:

$$\begin{aligned} SAAI_{p1} &:= \lambda \frac{|A_S^*|}{|A_S|} + (1-\lambda)\frac{K-1}{K} \ , \\ SAAI_{p2} &:= \lambda \frac{|A_S^*|}{|A_S|} + (1-\lambda)\frac{K-n_{\mathbb{1}}}{K} \ , \end{aligned} \tag{9}$$

with $\lambda \in [0, 1]$.

While the effect of penalizing pseudo-clusters through the second penalty term $\frac{n_{\mathbb{1}}}{K}$ is smaller compared to penalizing small values for K, both terms add significant improvement on the overall accuracy.

5 Discussion

Through our experiments in Section 4, we have shown that maximizing SAAI outperforms maximizing SSC, as proposed by (Shahapure and Nicholas 2020; Zhou and Gao 2014), as well as X-Means, proposed by (Pelleg, Moore et al. 2000) on the task of finding the true number of anomaly classes K in multivariate time series consisting of sufficiently similar measurements. Maximizing SAAI improves mean accuracy significantly over SAAI by 0.09 and over X-Means by alomst 0.17. The difference in mean accuracy of SAAI and FMI however is not statistically significant. The relatively low scores across all methods are subject to all runs from the *Lagged Variables* experiment being included in the evaluation. SAAI also shows a high to very high correlation with those results obtained by maximizing ARI and FMI. Our results are consistent with those of (Raihan 2023) that SSC is not suitable for finding the correct value for K by maximizing SSC when working with raw time series. Our findings contradict the proposal of (Shahapure and Nicholas 2020) and (Zhou and Gao 2014), however they did not evaluate their approaches on time series data. The *Lagged Variables* experiments give an idea of how the rather vague notion of *similar-enough* might be quantized. For correlation coefficients $\rho > 0.43$, maximizing SAAI gives an higher accuracy as maximizing SSC. However, for $l = 180$, which corresponds to $\rho = 0.43$, X-Means yields a slightly higher accuracy of 0.14 as SAAI with 0.12. Thus, as a rule of thumb, we could say that SAAI is superior in finding the correct value for K in similar measurements if their pairwise correlation coefficients ρ satisfy $\rho \geq 0.5$. When applied to real sensor data, as done in Section 4, we saw that the solution obtained by maximizing SAAI is easier to interpret compared to SSC. This is to be expected, since SAAI has already been shown to be superior in finding the ideal number of clusters. This result is also supported by those in (Rewicki et al. 2024a), where for the same ICS temperature time series, 10 clusters (9 anomaly types and one false-positive cluster) were identified. Our ablation study showed that, although the first penalty term is more influential, both are needed, especially when the ratio of synchronized anomalies to all anomalies $\frac{|A_S|}{|A|}$ is low. The influence of the second term would increase as the range of possible values k is increased, which would increase the likelihood of seeing pseudo-clusters. However, SAAI has two shortcomings. Since SAAI only considers synchronized anomalies, as defined in Definition 4, anomalies found in only one of the variables are not evaluated. Another drawback is the dependence on the similarity of the time-dependent signals. This limits the application of SAAI to these use cases, while SSC can be applied to arbitrary data.

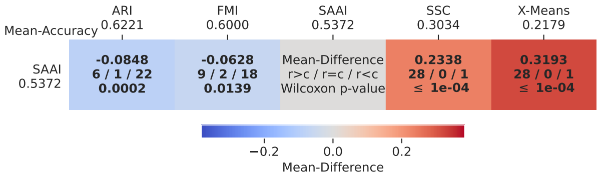


Figure 10: MCM comparing the results within the sweet spot of SAAI.

6 Conclusions

In this paper, we propose SAAI, an unsupervised measure of anomaly cluster quality that incorporates prior knowledge about the multivariate time series by exploiting the similarity between individual signals. We demonstrate the effectiveness of SAAI by showing that maximizing SAAI outperforms maximizing SSC and X-Means on the task of finding the true number of anomaly classes K. Also, SAAI shows high correlation with results obtained from maximizing the external measures ARI and FMI. When applied to real, unlabeled data, the clustering result found by maximizing SAAI is easier to interpret compared to SSC. Our ablation study shows that all parts of the SAAI formula are necessary. However, SAAI has two major shortcomings: (1) it is only applicable to univariate anomalies found in multivariate time series consisting of reasonably similar signals, and (2) SAAI does not consider anomalies found in only a single variable (i.e., unaligned). Both shortcomings will be the subject of future research, as addressing (1) would allow extension to multivariate anomalies, and including unaligned anomalies by addressing (2) will expand the range of valid use cases.

References

Arthur, D.; and Vassilvitskii, S. 2007. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, Soda '07, 1027–1035. Usa: Society for Industrial and Applied Mathematics. ISBN 9780898716245.

Barz, B.; et al. 2018. Detecting Regions of Maximal Divergence for Spatio-Temporal Anomaly Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41: 1088–1101.

Ben-Hur, A.; Elisseeff, A.; and Guyon, I. 2002. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 6–17.

Berndt, D. J.; and Clifford, J. 1994. Using Dynamic Time Warping to Find Patterns in Time Series. In *Proceedings of the 3rd Int. Conf. on Knowledge Discovery and Data Mining*, Aaaiws'94, 359–370. AAAI Press.

Dinh, D.-T.; Fujinami, T.; and Huynh, V.-N. 2019. Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient. In Chen, J.; Huynh, V. N.; Nguyen, G.-N.; and Tang, X., eds., *Knowledge and Systems Sciences*, 1–17. Singapore: Springer Singapore. ISBN 978-981-15-1209-4.

Fowlkes, E. B.; and Mallows, C. L. 1983. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383): 553–569.

Hubert, L.; and Arabie, P. 1985. Comparing partitions. *Journal of Classification*, 2(1): 193–218.

Ismail-Fawaz, A.; Dempster, A.; Tan, C. W.; Herrmann, M.; Miller, L.; Schmidt, D. F.; Berretti, S.; Weber, J.; Devanne, M.; Forestier, G.; and Webb, G. I. 2023. An Approach To Multiple Comparison Benchmark Evaluations That Is Stable Under Manipulation Of The Comparate Set. *arXiv preprint arXiv:2305.11921*.

Kodinariya, T. M.; and Makwana, P. R. 2013. Review on determining number of Cluster in K-Means Clustering.

López-Rubio, E.; Palomo, E. J.; and Ortega-Zamorano, F. 2018. Unsupervised learning by cluster quality optimization. *Information Sciences*, 436: 31–55.

Lu, Y.; et al. 2022. Matrix Profile XXIV: Scaling Time Series Anomaly Detection to Trillions of Datapoints and Ultra-fast Arriving Data Streams. In *Proceedings of the 28th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Kdd '22, 1173–1182. New York, NY, USA: Assoc. for Computing Machinery. ISBN 9781450393850.

MacQueen, J.; et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 281–297. Oakland, CA, USA.

Middlehurst, M.; Ismail-Fawaz, A.; Guillaume, A.; Holder, C.; Guijo-Rubio, D.; Bulatova, G.; Tsaprounis, L.; Mentel, L.; Walter, M.; Schäfer, P.; and Bagnall, A. 2024. aeon: a Python Toolkit for Learning from Time Series. *Journal of Machine Learning Research*, 25(289): 1–10.

Mukaka, M. M. 2012. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal : the journal of Medical Association of Malawi*, 24: 69–71.

Nguyen, T.-H. T.; Dinh, D.-T.; Sriboonchitta, S.; and Huynh, V.-N. 2019. A method for k-means-like clustering of categorical data. *J. Ambient Intell. Humaniz. Comput.*, 14: 15011–15021.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pelleg, D.; Moore, A.; et al. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *ICML'00*, 727–734. Citeseer.

Purnima Bholowalia, A. K. 2014. EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, 105(9): 17–24.

Raihan, M. N. 2023. Determining the Optimal Number of Clusters for Time Series Datasets with Symbolic Pattern Forest. arXiv:2310.00820.

Rand, W. M. 1971. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336): 846–850.

Rewicki, F.; Gawlikowski, J.; Niebling, J.; and Denzler, J. 2024a. Unraveling Anomalies in Time: Unsupervised Discovery and Isolation of Anomalous Behavior in Bio-regenerative Life Support System Telemetry. In *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track, European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 09–13, 2024, Proceedings, Part V*. Vilnius, Lithuania. Accepted for presentation.

Rewicki, F.; et al. 2024b. EDEN ISS 2020 Telemetry Dataset. Zenodo.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

Sakoe, H.; and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1): 43–49.

Schubert, E. 2023. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *SIGKDD Explor. Newsl.*, 25(1): 36–42.

Shahapure, K. R.; and Nicholas, C. 2020. Cluster Quality Analysis Using Silhouette Score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 747–748.

Sohn, K.; et al. 2023. Anomaly Clustering: Grouping Images into Coherent Clusters of Anomaly Types. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5468–5479.

Stefan, A.; Athitsos, V.; and Das, G. 2013. The Move-Split-Merge Metric for Time Series. *IEEE Transactions on Knowledge and Data Engineering*, 25(6): 1425–1438.

Tavenard, R.; Faouzi, J.; Vandewiele, G.; Divo, F.; Androz, G.; Holtz, C.; Payne, M.; Yurchak, R.; Rußwurm, M.; Kolar, K.; and Woods, E. 2020. Tslearn, A Machine Learning Toolkit for Time Series Data. *Journal of Machine Learning Research*, 21(118): 1–6.

Vintsyuk, T. K. 1968. Speech discrimination by dynamic programming. *Cybernetics*, 4(1): 52–57.

Wilcoxon, F. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6): 80–83.

Zabel, P.; Bamsey, M.; Zeidler, C.; Vrakking, V.; Schubert, D.; and Romberg, O. 2017. Future exploration greenhouse design of the EDEN ISS project. 47th International Conference on Environmental Systems.

Zhou, H. B.; and Gao, J. T. 2014. Automatic Method for Determining Cluster Number Based on Silhouette Coefficient. In *Advanced Research on Intelligent System, Mechanical Design Engineering and Information Engineering III*, volume 951 of *Advanced Materials Research*, 227–230. Trans Tech Publications Ltd.

A Appendix

SAAI Algorithm & Complexity Analysis

The algorithm for calculating the SAAI is shown in algorithm 1. To determine the set of synchronized anomalies A_S, we have to compare all anomalous subsequences in different variables $\mathcal{S}^{(i)}_{a_i,b_i}, \mathcal{S}^{(j)}_{a_j,b_j}$ that overlap in time, i.e. $i < j$ and $b_j < a_i$ or $b_i > a_j$. A sweep-line algorithm for calculating the SAAI is given in Algorithm 1. For each anomalous sequence, we create two events in lines 4-8 with a complexity of $\mathcal{O}(n)$, where $n = |A|$. The sorting of the $2n$ events in line 9 has a complexity of $\mathcal{O}(n \log n)$. The events are sorted by time and in case of ties by event, so that ”END” events are sorted before ”START” events. Each interval is added to (line 21) and removed from (line 23) the active intervals S once, resulting in $\mathcal{O}(n)$ insertions and deletions from S. Before an interval is added, it is compared to all active intervals. Since all intervals can be active at the same time, the maximum number of comparisons is $\binom{n}{2}$ in the worst case, which is in $\mathcal{O}(n^2)$. However, this worst case occurs only if n is close to the dimension of the time series D and all anomalous subsequences are synchronized. Typically, we have $D \ll n$ when clustering anomalous subsequences in multivariate time series. This gives a complexity of $\mathcal{O}(n \log n)$ for the average case where $D \ll n$ and overlaps are sparse, and $\mathcal{O}(n^2)$ for the worst case where $D \approx n$ and many overlapping intervals.

Selecting λ

The parameter λ in Equation (8) determines the weight of the main term over the regularizing term. A value of $\lambda = 1$ would evaluate only the main term and ignore the number of clusters and pseudo-clusters in the solution found. On the contrary, a value of $\lambda = 0$ would evaluate only the number of

Algorithm 1: SAAI Algorithm

Input:

A: The set of anomalies

K: Number of clusters

$n_{\mathbb{1}}$: Number of pseudo-clusters, i.e. number of clusters with only one element

θ_s: Degree of alignment between subsequences to be considered synchronous.

λ: Parameter to weight main and penalty term.

Output: saai

```
1:  $A_{\mathcal{S}}, A^*_{\mathcal{S}} \leftarrow \varnothing, \varnothing$
2:  $E, S \leftarrow [\,], [\,]$
3:  $i \leftarrow 0$
4:  for $S_{a,b} \in A$ do
5:      append $\{(”START”, a, i, b)\}$ to $E$
6:      append $\{(”END”, b, i)\}$ to $E$
7:      $i \leftarrow i + 1$
8:  end for
9:  $E \leftarrow sort(E, \text{time}, \text{type})$
10: for all events $e \in E$ do
11:     if $e.type = ”START”$ then
12:         for active intervals $s \in S$ do
13:             $v = \omega(e.time, e.end, s.time, s.end)$
14:             if $v \geq \theta$ then
15:                 $A_{\mathcal{S}} = A_{\mathcal{S}} \cup \{(A[e.id], A[s.id])\}$
16:                 if $c(A[e.id]) = c(A[s.id])$ then
17:                     $A^*_{\mathcal{S}} = A^*_{\mathcal{S}} \cup \{(A[e.id], A[s.id])\}$
18:                 end if
19:             end if
20:         end for
21:         append $e$ to $S$
22:     else
23:         delete $e$ from $S$
24:     end if
25: end for
26: $saai \leftarrow \lambda \frac{|A^*_S|}{|A_S|} + (1 - \lambda) \frac{K-1-n_{\mathbb{1}}}{K}$
27:
28: return $saai$
```

clusters and pseudo-clusters and ignore the synchronicity of anomalies. Figure 11 plots the mean accuracy for finding the true number of classes over all experiments on the synthetic greenhouse temperature data presented in Section 4 for increasing λ. As can be seen in Figure 11, weighting the main and the regularizing term equally gives the best result for this example. However, there may be situations where weighting the two terms differently makes sense, e.g. when there is no preference for a larger number of clusters. When choosing a value for λ, care should be taken, especially when giving more weight to the regularizing term, as this can be more detrimental to the performance of SAAI than giving more weight to the main term. In general, weighting both terms equally by setting $\lambda = 0.5$ is a good initial choice.

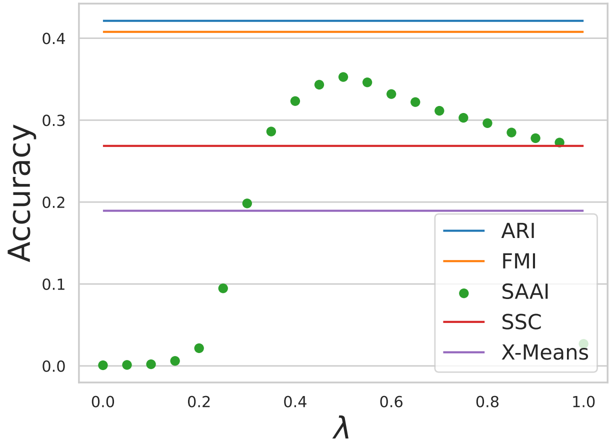
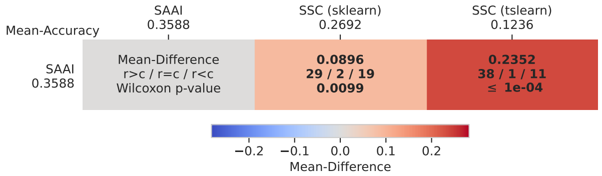


Figure 11: Accuracy of SAAI for increasing λ compared to ARI, FMI, SSC and X-Means aggregated over all experimental results on the synthetic greenhouse temperature data.

Silhouette Score Implementations

While running our experiments on synthetic data, we found that the results for the Silhouette Score depend strongly on the implementation used to compute it. As shown in Figure 12, the implementation in the *tslearn* package (Tavenard et al. 2020) gives significantly worse results than the implementation in *scikit-learn*. This is surprising since *tslearn* is a specialized package for time series analysis and supports the calculation of the Silhouette score for sequences of unequal length. However, for the sake of a fair comparison, we decided to use the *scikit-learn* implementation in our main experiments, since the average accuracy of 0.2692 is more than double that of 0.1236 using the *tslearn* implementation.



Mean-Accuracy	SAAI 0.3588	SSC (sklearn) 0.2692	SSC (tslearn) 0.1236
SAAI 0.3588	Mean-Difference r>c / r=c / r<c Wilcoxon p-value	**0.0896** **29 / 2 / 19** **0.0099**	**0.2352** **38 / 1 / 11** **≤ 1e-04**

−0.2
−0.1
0.0
0.1
0.2
Mean-Difference

Figure 12: Multi-Comparison Matrix comparing the results obtained by using different implementations of the Silhouette Score.

SAAI and SSC results for edeniss2020 (ICS) dataset

For the experiment on real ICS data from the EDEN ISS research greenhouse, we clustered the anomalous subsequences found by the MDI and DAMP algorithms with increasing number of clusters $1 < K < 20$. The results of SAAI and SSC are visualized in Figure 13. Both SSC variants have their highest score at $K = 3$, while SAAI has its maximum at $K = 11$, which seems to be a more realistic value in this case.

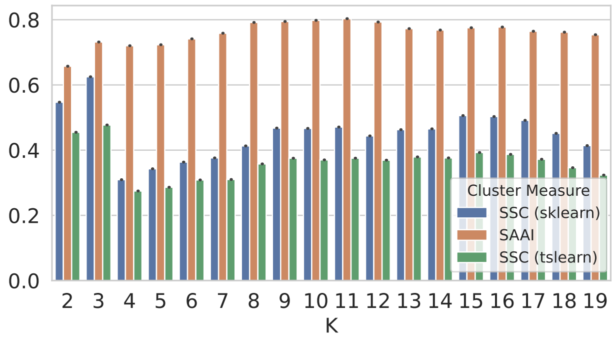


Figure 13: SAAI and SSC scores (tslearn and scikit-learn) for $1 < K < 20$

SAAI Example (high res)

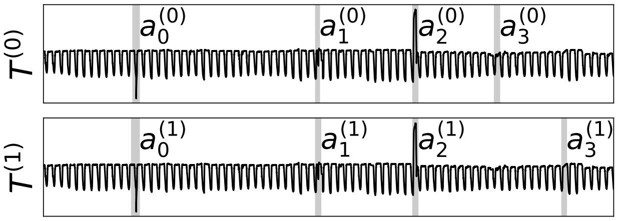
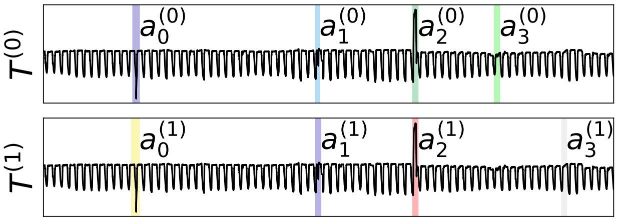
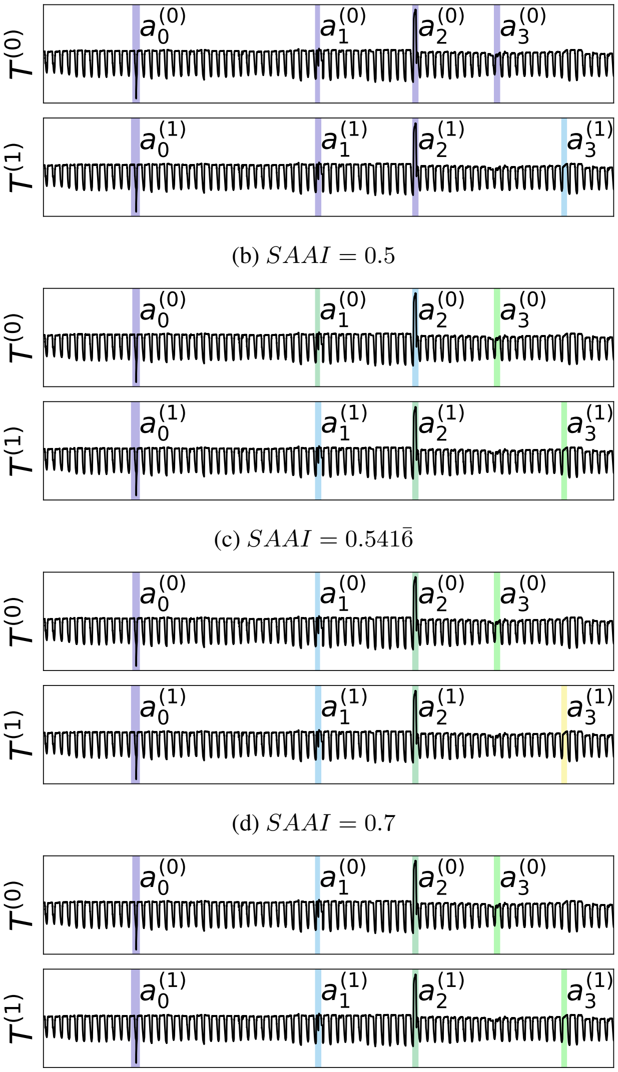


Figure 14: The detected anomalies $a_j^{(i)}$



(a) $SAAI = 0$



(b) $SAAI = 0.5$

(c) $SAAI = 0.541\bar{6}$

(d) $SAAI = 0.7$

(e) $SAAI = 0.875$

Figure 15: (a) - (e) different clustering solutions with increasing quality. Cluster assignment is coded by color. (a) Worst case: all but one cluster contain a single element, (b) all but one anomaly assigned to the same cluster, (c) synchronized anomalies not in the same cluster, (d) synchronized anomalies in separate clusters, single element clusters exist, (e) best case: synchronized anomalies in separate clusters, no single element cluster.