SoftED: Metrics for Soft Evaluation of Time Series Event Detection

Rebecca Salles, Janio Lima, Rafaelli Coutinho, Esther Pacitti, Florent Masseglia, Reza Akbarinia, Chao Chen, Jonathan M. Garibaldi, Fábio André Machado Porto, Eduardo S. Ogasawara

SoftED: Metrics for Soft Evaluation of Time Series Event Detection

Rebecca Salles
CEFET/RJ
rebeccapsalles@acm.org

Janio Lima
CEFET/RJ
janio.lima@eic.cefet-rj.br

Rafaelli Coutinho
CEFET/RJ
rafaelli.coutinho@cefet-rj.br

Esther Pacitti
University of Montpellier & INRIA
esther.Pacitti@lirmm.fr

Florent Masseglia
University of Montpellier & INRIA
florent.masseglia@inria.fr

Reza Akbarinia
University of Montpellier & INRIA
reza.akbarinia@inria.fr

Chao Chen
University of Nottingham
chao.chen@nottingham.ac.uk

Jonathan Garibaldi
University of Nottingham
jon.garibaldi@nottingham.ac.uk

Fabio Porto
LNCC
fporto@lncc.br

Eduardo Ogasawara
CEFET/RJ
eogasawara@ieee.org

Abstract

Time series event detection methods are evaluated mainly by standard classification metrics that focus solely on detection accuracy. However, inaccuracy in detecting an event can often result from its preceding or delayed effects reflected in neighboring detections. These detections are valuable to trigger necessary actions or help mitigate unwelcome consequences. In this context, current metrics are insufficient and inadequate for the context of event detection. There is a demand for metrics that incorporate both the concept of time and temporal tolerance for neighboring detections. This paper introduces SoftED metrics, a new set of metrics designed for soft evaluating event detection methods. They enable the evaluation of both detection accuracy and the degree to which their detections represent events. They improved event detection evaluation by associating events and their representative detections, incorporating temporal tolerance in over 36% of experiments compared to the usual classification metrics. SoftED metrics were validated by domain specialists that indicated their contribution to detection evaluation and method selection.

***K*eywords** Time Series · Event Detection · Evaluation Metrics · Soft Computing

1 Introduction

In time series analysis, it is often possible to observe a significant change in behavior at a certain point or time interval. Such behavior change generally characterizes the occurrence of an event [29]. An event can represent a phenomenon with a defined meaning in a domain. Event detection is the process of identifying events in time series. With this process, we may be interested in learning/identifying past events [45, 20, 2, 3, 54, 71, 43, 62], identifying events in real-time (online detection) [69, 1, 5, 41], or even predicting future events before they happen (event prediction) [67, 50, 39, 26, 70]. It is recognized as a basic function in surveillance and monitoring systems and has gained much attention in research for application domains involving large datasets from critical systems [45].

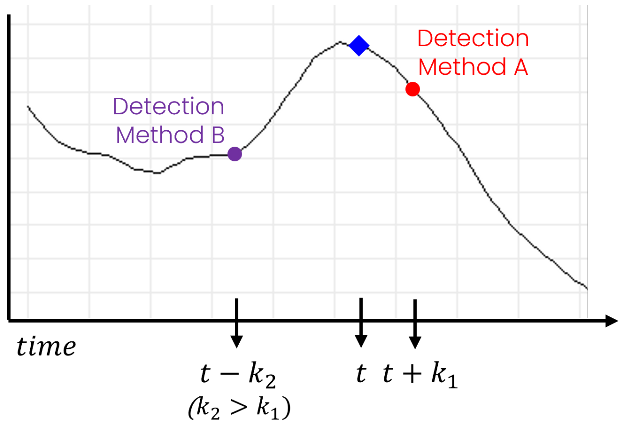


Figure 1: Example regarding the problem of evaluating the detection of an event at time t. Method A detects an event at time $t + k_1$, while Method B detects an event at time $t - k_2$ ($k_2 > k_1$).

To address the task of time series event detection, several methods have been developed and are surveyed in the literature [31, 12, 28, 15, 16, 10, 6, 53]. Each detection method specializes in time series that present different characteristics or make assumptions about the data distribution. Therefore, the assessment of their detection performance is important to infer their adequacy to a particular application [24]. In this case, detection performance refers to how accurate an event detection method is at identifying events in a time series. Detection performance is generally measured by classification metrics [30].

Currently, standard classification metrics (around since the 1950s), including Recall, Precision, and F1, are usually adopted [35, 59]. Although Accuracy is a specific metric [30], the expression detection accuracy is henceforth used to refer to the ability of a method to detect events correctly. Classification metrics focus mainly on an analysis of detection accuracy. On the other hand, inaccuracy in event detection does not always indicate a bad result, especially when detections are sufficiently close to events.

1.1 Motivating example and problem definition

This section gives an example of the problem of evaluating inaccurate event detections and defines the problem of the paper regarding the demand for adequate detection performance metrics. Consider, for example, a time series X containing an event at time t, represented in Figure 1. Given detection methods A and B applied to X, a user must select one of them as the most adequate for the underlying application. Method A detects an event at time $t + k_1$, while Method B detects an event at time $t - k_2$ ($k_2 > k_1$). As none of the methods could correctly detect the event at time t, based on the usual detection accuracy evaluation, the user would deem both as inaccurate and disposable.

However, inaccuracy in detecting an event can often result from its preceding or lingering effects. Take the adoption of a new policy in a business. While a domain specialist may consider the moment of policy enforcement as a company event, its effects on profit may only be detectable a few months later. On the other hand, preparations for policy adoption may be detectable in the antecedent months. Moreover, when accurate detections are not achievable, which is common, detection applications demand events to be identified as soon as possible [35], or early enough to allow necessary actions to be taken, mitigating possible critical system failures or helping mitigate urban problems resulting from extreme weather events, for example. In this context, the results of Methods A and B would be valuable to the user. Note that while Method B seems to anticipate the event, its detection is made after the event's occurrence. On the other hand, the detection of Method A came temporally closer to the event, possibly more representative of its effects.

In this context, evaluating event detection is particularly challenging, and the detection accuracy metrics usually adopted are insufficient and inadequate for the task [56]. Standard classification metrics do not consider the concept of time, which is fundamental in the context of time series analysis, and do not reward early detection [1], for example, or any relevant neighboring detections. For the remainder of this paper, neighboring or close detections refer to detections whose temporal distance to events is within a desired threshold. Current metrics only reward true positives (exact matches in event detection). All other results are "harshly" and equally discredited.

In this case, there is a demand to soften the usual concept of detection accuracy and evaluate the methods while considering neighboring detections. However, state-of-the-art metrics designed for scoring anomaly detection [35] are still limited [56], while also being biased towards results preceding events, such as the ones produced by Method B.

To the best of our knowledge, there are no metrics available in the literature that consider both the concept of time and tolerance for detections that are sufficiently close to time series events. This paper focuses on addressing this demand.

1.2 Contribution

This paper introduces the SoftED metrics, a new set of metrics for evaluating event detection methods regarding both their detection accuracy and ability to produce neighboring detections to events. Inspired by soft (approximate) computing, SoftED metrics are designed for soft evaluation, assessing the degree to which a detection represents a particular event. Hence, they incorporate both the concept of time and temporal tolerance for inaccuracy in event detection evaluation, a scenario that domain specialists and users often face, with no standard or adequate evaluation metrics until now. SoftED metrics soften the standard classification metrics, which are considered in this paper as hard metrics, to support the decision-making regarding the most appropriate method for a given application.

Computational experiments were conducted to analyze the contribution of the developed metrics against the usual hard and state-of-the-art metrics [35]. Results indicate that the developed SoftED metrics improved event detection evaluation by associating events and their representative (neighboring) detections, incorporating temporal tolerance in over 36% of the conducted experiments compared to usual hard metrics. More importantly, surveyed domain specialists validated the contribution of SoftED metrics to the problem of detection method evaluation.

The remainder of this paper is organized as follows. Section 2 provides concepts on time series event detection and reviews the literature on detection performance metrics for detection evaluation. Section 3 formalizes the developed SoftED metrics. Section 4 presents a quantitative and qualitative experimental evaluation of the developed metrics and their empirical results. Finally, conclusions are made in Section 5.

2 Literature review

This section provides relevant concepts on time series events and their detection and reviews the literature on detection performance metrics and related works. Events are pervasive in real-world time series, especially in the presence of nonstationarity [29, 51]. Commonly, the occurrence of an event can be detected by observing anomalies or change points. Most event detection methods in the literature specialize in identifying a specific type of event. There exist methods that can detect multiple events in time series, generally involving the detection of both anomalies and change points [36, 4, 70]. Nonetheless, these methods are still scarce. This paper approaches methods for detecting anomalies, change points or both.

2.1 Time series events

Events correspond to a phenomenon, generally pre-defined in a particular domain, with an inconstant or irregular occurrence relevant to an application. In the context of time series, events represent significant changes in expected behavior at a certain time or interval [29]. In general, punctual events of a given time series $X = <x_1, x_2, x_3, \cdots, x_n>$, can be identified in a simplified way by $e(X, k, \sigma)$ using the Eq. 1, where k represents the length of nearby observations. If an observation x_t escapes the expected behavior based on previous $\{x_{t-k}, \ldots, x_{t-1}\}$ or later $\{x_{t+1}, \ldots, x_{t+k}\}$ observations (above a threshold σ), it can be considered an event[1].

$$\begin{aligned} e(X, k, \sigma) = \{t, &|x_t - E(x_t|\{x_{t-k}, \ldots, x_{t-1}\})| > \sigma \lor \\ &|x_t - E(x_t|\{x_{t+1}, \ldots, x_{t+k}\})| > \sigma\} \end{aligned} \tag{1}$$

Anomalies Most commonly, events detected in time series refer to anomalies. Anomalies appear not to be generated by the same process as most of the observations in the time series [12]. Thus, anomalies can be modeled as isolated observations of the remaining nearby data. In this case, an event identified in x_t can be considered an anomaly if it escapes expected behavior both before and after time point t according to $a(X, k, \sigma)$ in Eq. 2. Generally, anomalies are identified by deviations from the time series inherent trend. However, anomalies may also present themselves as data volatility variations.

$$\begin{aligned} a(X, k, \sigma) = \{t, &|x_t - E(x_t|\{x_{t-k}, \ldots, x_{t-1}\})| > \sigma \land \\ &|x_t - E(x_t|\{x_{t+1}, \ldots, x_{t+k}\})| > \sigma\} \end{aligned} \tag{2}$$

Change points Change points in a time series are the points or intervals in time that represent a transition between different states in a process that generates the time series [57]. In this case, a change point event identified in x_t follows

[1] Due to limited space, the general formalization of event intervals lie outside the scope of this paper.

the expected behavior observed before or after the time point t, but not both at the same time according to $cp(X, k, \sigma)$ in Eq. 3.

$$cp(X, k, \sigma) = \{t, |x_t - E(x_t|\{x_{t-k}, \ldots, x_{t-1}\})| > \sigma \ \veebar \\ |x_t - E(x_t|\{x_{t+1}, \ldots, x_{t+k}\})| > \sigma\} \tag{3}$$

2.2 Event detection

Event detection is the process of identifying the occurrence of such events based on data analysis. It is recognized as a basic function in surveillance and monitoring systems. Moreover, it becomes even more relevant for applications based on time series and sensor data analysis [45]. Event detection methods found in the literature are usually based on model deviation analysis, classification-based analysis, clustering-based analysis, domain-based analysis, or statistical techniques [12, 30, 45, 3]. Regardless of the adopted detection strategy, an important aspect of any event detection method is how the events are reported. Typically, the outputs produced by event detection methods are either scores or labels. Scoring detection methods assign an anomaly score to each instance in the data depending on the degree to which that instance is considered an anomaly. On the other hand, labeling detection methods assign a label (normal or anomalous) to each data instance. Such methods are the most commonly found in the literature [12].

2.3 Detection performance metrics

Detection methods might variate their performance under different time series [23]. Therefore, there is a demand for comparing the results provided by them. Such a process aims to guide the choice of suitable methods for detecting events of a time series in a particular application. For comparing event detection methods, standard classification metrics, such as F1, Precision, and Recall, are usually adopted [30, 35, 59].

As usual, the standard classification metrics depend on measures of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [30]. In event detection, the TP refers to the number of events correctly detected (labeled) by the method. Analogously, TN is the number of observations that are correctly not detected. On the other hand, the measure FP is the number of detections that did not match any event, that is, false alarms. Analogously, FN is the number of undetected events. Among the standard classification metrics, Precision and Recall are widely adopted. Precision reflects the percentage of detections corresponding to time series events (exactness), whereas Recall reflects the percentage of correctly detected events (completeness). Precision and Recall are combined in the F_β metrics [30]. The F1 metric is also widely used to help gauge the quality of event detection balancing Precision and Recall [59].

Event detection is particularly challenging to evaluate, as discussed in Section 1.1. In this context, the Numenta Anomaly Benchmark (NAB) provided a common scoring algorithm for evaluating and comparing the efficacy of anomaly detection methods [35]. The NAB score metric is computed based on anomaly windows of observations centered around each event in a time series. Given an anomaly window, NAB uses the sigmoidal scoring function to compute the weights of each anomaly detection. It rewards earlier detections within a given window and penalizes FPs. Also, NAB allows the definition of application profiles: standard, reward low FPs, and reward low FNs. The standard profile gives relative weights to TPs, FPs, and FNs based on the window size.

Nonetheless, the NAB scoring system presents challenges for its usage in real-world applications. For example, the anomaly window size is automatically defined as 10% of the time series size, divided by the number of events it contains, values that are generally not known in advance, especially in streaming environments. Furthermore, Singh and Olinsky [56] pointed out poor definitions and arbitrary constants in the scoring equations. Finally, score values increase with the number of events and detections. Every user can tweak the weights in application profiles, making it difficult to interpret and benchmark results obtained by other users or setups.

In addition to NAB [35], this section presents other works related to the problem of analyzing and comparing event detection performance [12]. Recent works focus on the development of benchmarks to evaluate univariate time series anomaly detection methods [34, 8, 66]. Jacob et al. [34] provide a comprehensive benchmark for explainable anomaly detection over high-dimensional time series, while the benchmark developed by Boniol et al. [8] allows the user to assess the advantages and limitations of both anomaly detection methods and detection accuracy metrics.

Standard classification metrics are generally used for evaluating the ability of an algorithm to distinguish normal from abnormal data samples [12, 59]. Aminikhanghahi and Cook [4] review traditional metrics for change point detection evaluation, such as Sensitivity, G-mean, F-Measure, ROC, PR-Curve, and MSE. Detection evaluation measures have also been investigated in the areas of sequence data anomaly detection [13], time series mining and representation [19], and sensor-based human activity learning [17, 64].

Metrics found in the literature are mainly designed to evaluate the detection of punctual anomalies. However, many real-world event occurrences extend over an interval (range-based). Motivated by this, Tatbul et al. [59] and Paparrizos et al. [44] extend the well-known Precision and Recall metrics, and the AUC-based metrics, respectively, to measure the accuracy of detection algorithms over range-based anomalies. Other recent metrics developed for the task of detecting range-based time series anomalies are also included in the benchmark of Boniol et al. [8]. In addition, Wenig et al. [66] published a benchmarking toolkit for algorithms designed for detecting anomalous subsequences in time series [7, 9].

To the best of our knowledge, few works opt to evaluate event detection algorithms based on other than traditional metrics. For example, Wang, Vuran, and Goddard [63] calculate the delay until an individual node and the delivery delay in a transmission network detect an event. The work presents a framework for capturing delays in detecting events in large-scale WSN networks with a time-space simulation. Conversely, Tatbul et al. [59] also observe the neighborhood of event detections, not to calculate detection delays, but to evaluate positional tendency in anomaly ranges. Ultimately, our previous work uses the delay measure to evaluate the ability of algorithms to detect real events in time series [22]. Escobar et al. [22] contribute by studying the time distance between event detections and identified events. It furthers a qualitative analysis of the tendency of algorithms to detect before or after the occurrence of an event.

Under these circumstances, there is still a demand for event detection performance metrics that incorporate both the concept of time and tolerance for detections that are sufficiently close to time series events. Therefore, this paper contributes by introducing new metrics for evaluating methods regarding their detection accuracy while also considering neighboring detections, incorporating temporal tolerance for inaccuracy in event detection.

3 SoftED

This paper adopts a distance-based approach to develop novel metrics designed to evaluate the performance of methods for detecting events in time series. The inspiration for the proposed solution is found in soft (or approximate) computing. Soft computing is a collection of methodologies that exploit tolerance for inaccuracy, uncertainty, and partial truth to achieve tractability, robustness, and low solution cost [60]. In this context, the main proposed idea is to soften the hard metrics (standard classification metrics) to incorporate temporal tolerance or inaccuracy in event detection. Such metrics seek to support the decision-making of the most appropriate method for a given application with a basis not only on the usual analysis of the detection accuracy but also on the analysis of the ability of a method to produce detections that are close enough to actual time series events. Henceforth, the proposed approach is named Soft classification metrics for Event Detection, or SoftED. This section formalizes the SoftED metrics.

Figure 2 gives a general idea of the proposed approach, illustrating the key difference between the standard hard evaluation and the proposed soft evaluation. Blue rhombuses represent actual time series events. Circles correspond to detections produced by a particular detection method. The hard evaluation concerns a binary value regarding whether detection is a perfect match to the actual event. In this case, circles are green when they perfectly match the events and red when they do not. Conversely, soft evaluation assesses the degree to which detection relates to a particular event.

3.1 Defining an event membership function

In order to soften the standard hard metrics, we incorporate a distance-based temporal tolerance for events. It is done by defining the relevance of a particular detection to an event. This section formalizes the proposed approach. Table 1 defines the main variables used in the formalization of SoftED. Given a time series X containing a set of m events, $E = \{e_1, e_2, \ldots, e_m\}$, where e_j, $j = 1, \ldots, m$, is the j–th event in E occurring at time point t_{e_j}. A particular detection method applied to X produces a set of n detections, $D = \{d_1, d_2, \ldots, d_n\}$, where d_i, $i = 1, \ldots, n$, is the i–th detection in D indicating the time point t_{d_i} as a detection occurrence.

The degree to which a detection d_i is relevant to a particular event e_j is given by an event membership function $\mu_{e_j}(t)$ as defined in Equation 4 and illustrated in Figure 3a. This solution was inspired by Fuzzy sets, where we innovate by fuzzifying the time dimension rather than the time series observations [68]. The definition of $\mu_{e_j}(t)$ considers the acceptable tolerance for inaccuracy in event detection for a particular domain application. The acceptable time range in which an event detection is relevant for allowing an adequate response reaction to a domain event is given by the constant k.

$$\mu_{e_j}(t_{d_i}) = max\left(min\left(\frac{t_{d_i} - (t_{e_j} - k)}{k}, \frac{(t_{e_j} + k) - t_{d_i}}{k}\right), 0\right) \tag{4}$$

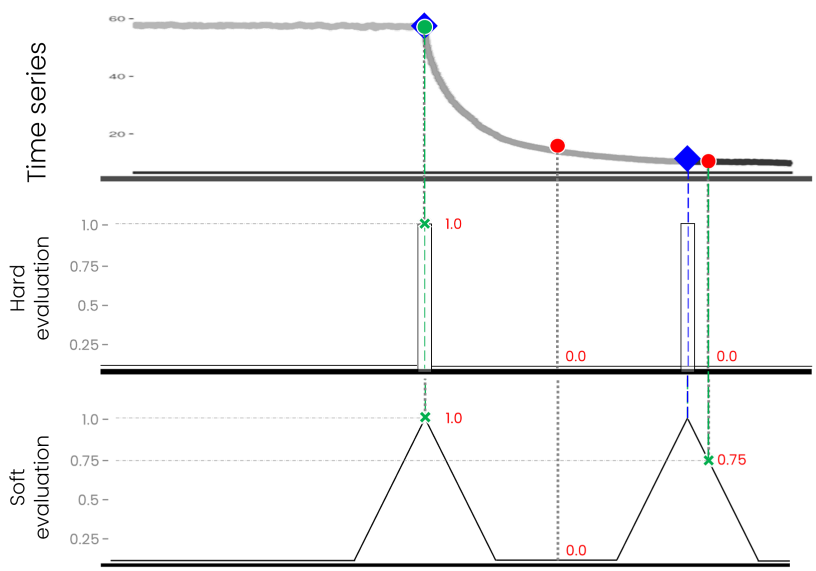


Figure 2: The general idea behind the proposed approach comparing the standard "hard" evaluation and the "soft" evaluation of the event detection.

Table 1: Definition of main variables for the formalization of SoftED

Var.	Value	Description
E	$\{e_1, e_2, \ldots, e_m\}$	set of time series events
m	$\|E\|$	number of events
j	$1, \ldots, m$	event index
e_j	–	the j–th event in E
t_{e_j}	time point	time point where the e_j occurs
D	$\{d_1, d_2, \ldots, d_n\}$	set of detections
n	$\|D\|$	number of detections
i	$1, \ldots, n$	detection index
d_i	–	the i–th detection in D
t_{d_i}	time point	time point where d_i occurs
k	time duration	constant of tolerance for event detections

Figure 3b represents the evaluation of $\mu_{e_j}(t)$ for two detections, d_1 and d_2, produced by a particular detection method. In this context, $\mu_{e_j}(t_{d_i})$ gives the extent to which a detection d_i represents event e_j, or, in other words, its temporal closeness to a hard true positive (TP) regarding e_j. In that case, detection d_1 is closer to a TP, and d_2 lies outside the tolerance range given by k and could be considered a false positive.

3.2 Maintaining integrity with hard metrics

We are interested that the SoftED metrics still preserve concepts applicable to traditional (hard) metrics. In particular, the SoftED metrics are designed to express the same properties as their hard correspondents. Moreover, they are designed to maintain the reference to the perfect detection performance (score of 1 as in hard metrics) and indicate how close a detection method came to it. In order to achieve this goal, this approach defines constraints necessary for maintaining integrity concerning the standard hard metrics:

1. A given detection d_i must have only one associated score.
2. The total score associated with a given event e_j must not surpass 1.

The first constraint comes from the idea that the detection d_i should not be rewarded more than once. It avoids the possibility of the total score for d_i surpassing the perfect reference score of 1. Take, for example, the first scenario, presented in Figure 3c, in which we have one detection and many close events. The detection d_1 is evaluated for events

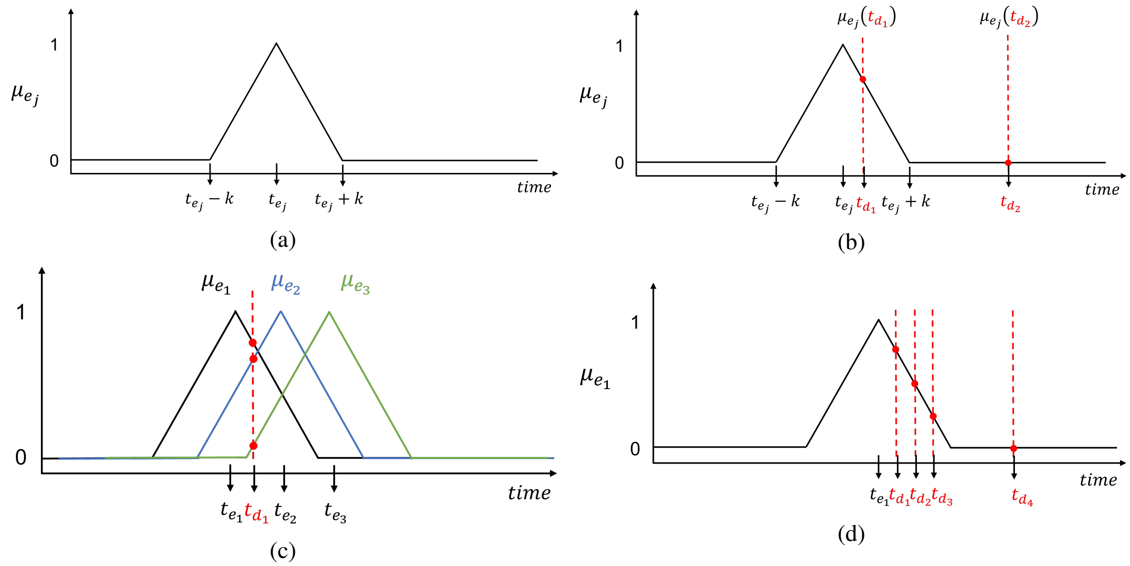


Figure 3: Auxiliary plots for comprehension of SoftED. (a) represents an event membership function $\mu_{e_j}(t)$. (b) represents $\mu_{e_j}(t)$ for detections d_1 and d_2. (c) depicts the example scenario containing one detection to many events, motivating the first constraint of SoftED. (d) depicts the example scenario containing many detections to a single event, motivating the second constraint of SoftED.

e_1, e_2, and e_3 resulting in three different membership evaluation of $\mu_{e_1}(t_{d_1})$, $\mu_{e_2}(t_{d_1})$, and $\mu_{e_3}(t_{d_1})$, respectively. Nevertheless, to maintain integrity with hard metrics, a given detection d_1 must not have more than one score. Otherwise, d_1 would be rewarded three times, and its total score could surpass the score of a perfect match, which would be 1.

To address this issue, we devise a strategy for attributing each detection d_i to a particular event e_j. The adopted attribution strategy is based on the temporal distance between d_i and e_j. It facilitates interpretation and avoids the need for solving an optimization problem for each detection. In that case, we attribute d_i to the event e_j that maximizes the membership evaluation $\mu_{e_j}(t_{d_i})$. This attribution is given by E_{d_i} defined in Equation 5. According to Figure 3c, d_1 is attributed to event e_1 given the maximum membership evaluation of $\mu_{e_1}(t_{d_1})$. In case there is a tie for the maximum membership evaluation of two or more events, E_{d_i} represents the set of events to which d_i is attributed. As a consequence, we can also derive the set of detections attributed to each event e_j, D_{e_j}, defined by Equation 6. The addition of a detection d_i to the set D_{e_j} is further conditioned by the tolerance range, that is, a membership evaluation greater than 0 ($\mu_{e_j}(t_{d_i}) > 0$).

$$E_{d_i} = \arg\max_{e_j} \left(\mu_{e_j}(t_{d_i})\right) \tag{5}$$

$$D_{e_j} = \{d_i \mid E_{d_i} \supset e_j \wedge \mu_{e_j}(t_{d_i}) > 0\} \tag{6}$$

The second constraint defined by this approach comes from the idea that a particular detection method should not be rewarded more than once for detecting the same event e_j. It assures that the total score of detections for event e_j does not surpass the perfect reference score of 1. Take, for example, the second scenario, presented in Figure 3d, in which we have many detections attributed to the same event. The event e_1 is present in the sets E_{d_1}, E_{d_2}, E_{d_3} and E_{d_4}. Moreover, D_{e_1} contains d_1, d_2, and d_3. But in order to maintain integrity with hard metrics, the total score for D_{e_1} ($\mu_{e_1}(t_{d_1}) + \mu_{e_1}(t_{d_2}) + \mu_{e_1}(t_{d_3})$) must not surpass the score of a perfect match.

To address this issue we devise an analogous distance-based strategy for attributing a representative detection d_i to each event e_j. In that case, we attribute to event e_j the detection d_i, contained in D_{e_j}, that maximizes the membership evaluation $\mu_{e_j}(t_{d_i})$. This attribution is given by $\hat{d}_{e_j}$ defined in Equation 7. According to Figure 3d, e_1 is best represented by detection d_1 given the maximum membership evaluation of $\mu_{e_1}(t_{d_1})$. As a consequence, we can compute the associated score for each event e_j as $es(e_j)$, defined by Equation 8.

$$\hat{d}_{e_j} = \arg\max_{d_i} \left(\{\mu_{e_j}(t_{d_i}) \mid d_i \in D_{e_j}\}\right) \tag{7}$$

$$es(e_j) = \mu_{e_j}(t_{\hat{d}_{e_j}}) \tag{8}$$

Finally, each detection d_i produced by a particular detection method is scored by $ds(d_i)$ defined in Equation 9. Representative detections ($d_i = \hat{d}_{e_j}$) are scored based on $es(e_j)$. All other detections are scored 0. This definition ensures the total score for detections of a particular method does not surpass the number of real events m contained in the time series X. It holds true the Equation 10 and maintains the reference to the score of a perfect detection Recall according to usual hard metrics. Furthermore, it penalizes false positives and multiple detections for the same event e_j.

$$ds(d_i) = \begin{cases} es(e_j), & \text{if } \exists e_j \in E \mid d_i = \hat{d}_{e_j} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$$\sum_{i=1}^{n} ds(d_i) \leq m \tag{10}$$

3.3 Computing the SoftED metrics

The scores computed for each detection d_i, $ds(d_i)$, are used to create soft versions of the hard metrics TP, FP, TN, and FN, as formalized in Table 2. In particular, while the value of TP gives the number of detections that perfectly matched an event (score of 1), the sum of $ds(d_i)$ scores indicate the degree to which the detections of a method approximate the m events contained in time series X given the temporal tolerance of k observations. Hence, the soft version of the TP metric, TP_s is given by $\sum_{i=1}^{n} ds(d_i)$. Conversely, the soft version of FN, FN_s, indicates the degree to which a detection method could not approximate the m events in an acceptable time range. The FN_s can then be defined as the difference between TP_s and the perfect Recall score ($m - \text{TP}_s$).

On the other hand, while the value of FP gives the number of detections that did not match an event (score of 0), its soft version, FP_s, indicates how far the detections of a method came to the events contained in time series X given the temporal tolerance of k observations. In that sense, FP_s is the complement of TP_s and can be defined by $\sum_{i=1}^{n} (1 - ds(d_i))$. Finally, the soft version of TN, TN_s, indicates the degree to which a detection method could avoid nonevent observations of X ($|t| - m$). The TN_s is given by the difference between FP_s and the perfect specificity score $((|t| - m) - \text{FP}_s)$.

Table 2: Formalization of SoftED Metrics

$\text{TP}_s =$	$\sum_{i=1}^{n} ds(d_i)$	$\text{FN}_s =$	$m - \text{TP}_s$
$\text{FP}_s =$	$\sum_{i=1}^{n} (1 - ds(d_i))$	$\text{TN}_s =$	$(\lvert t\rvert - m) - \text{FP}_s$

Due to the imposed constraints described in Section 3.2, the defined SoftED metrics TP_s, FP_s, TN_s, and FN_s, hold the same properties and the same scale as traditional hard metrics. Consequently, using their same characteristic formulas, they can derive soft versions of traditional scoring methods, such as Sensitivity, Specificity, Precision, Recall, and F1. Moreover, SoftED scoring methods still provide the same interpretation while including temporal tolerance for inaccuracy, which is pervasive in time series event detection applications. An implementation of SoftED metrics in R is made publicly available at Github[2].

4 Experimental evaluation

SoftED metrics were submitted to an experimental evaluation to analyze their contribution against the traditional hard metrics and the NAB score, both being the current state-of-the-art detection scoring methods [35]. The proposed metrics are evaluated based on two complementary analyses: (i) a quantitative analysis of the effects of the incorporated temporal tolerance in event detection evaluation; and (ii) a qualitative analysis of its contribution under different scenarios. For that, a large set of computational experiments were performed with the application of several different methods for event detection in real-world and synthetic time series datasets containing ground truth event data. Detection results were evaluated based on SoftED, hard, and NAB metrics. First, this section describes the adopted time series datasets and experimental settings. Finally, the quantitative and qualitative results are presented.

[2]SoftED implementation, datasets and experiment codes: `https://github.com/cefet-rj-dal/softed`

4.1 Datasets

This section presents the datasets selected for evaluating the SoftED metrics. The selected datasets are widely available in the literature and are composed of simulated and real-world time series regarding several different domain applications such as water quality monitoring (GECCO) [48][3], network service traffic (Yahoo) [65], social media (NAB) [1], oil well exploration (3W) [61], and public health (NMR) [4], among others. The selected datasets present over 6 hundred representative time series containing different types of events. In particular, GECCO, Yahoo, and NAB contain mostly anomalies, while 3W and NMR contain mostly change points. Moreover, the datasets present different types of nonstationarity and statistical properties to provide a more thorough discussion of the effects of the incorporated temporal tolerance on event detection evaluation on diverse datasets.

4.2 Experimental settings

For evaluating SoftED metrics, a set of up to 12 different event detection methods were applied to all time series in the adopted datasets, totalizing $4,026$ event detection experiments. Each experiment comprised an offline detection application, where the methods had access to the entire time series given as input. The applied detection methods are implemented and publicly available in the *Harbinger* framework [52]. It integrates and enables the benchmarking of different state-of-the-art event detection methods. These methods encompass searching for anomalies and change points through different techniques, including statistical, volatility, proximity, and machine learning methods. The adopted methods are described in detail by Escobar et al. [22], namely: the Forward and Backward Inertial Anomaly Detector (FBIAD) [38], K-Nearest Neighbors (KNN-CAD) [25], anomalize (based on time series decomposition [21, 55]) [28, 18], and GARCH [11], for anomaly detection; the Exponentially Weighted Moving Average (EWMA) [47], seminal method of detecting change points (SCP) [29], and ChangeFinder (CF) [57], for change point detection; and the machine learning methods based on the use Feed-Forward Neural Network (NNET) [49], Convolutional Neural Networks (CNN) [27, 37], Support Vector Machine (SVM) [14, 46], Extreme Learning Machine (ELM) [32, 58], and K-MEANS [42], for general purpose event detection.

In each experiment, detection methods were evaluated using the hard metrics Precision, Recall, and F1. Among them, the F1 was the main metric used for comparison. The NAB score was also computed with the standard application profile for each detection method result. The NAB scoring algorithm is implemented and publicly available in the R-package *otsad* [33]. Anomaly window sizes were automatically set. During the computation of the NAB score, confusion matrix metrics are built. Based on these metrics, the F1 metric of the NAB scoring approach was also computed. Finally, the SoftED metrics were computed for soft evaluation of the applied event detection methods. In particular, as the constant of temporal tolerance, k, is domain-dependent, this experimental evaluation was set to 15, defining a tolerance window of 30 observations enough to hold the central limit theorem unless stated otherwise. Nevertheless, k also experimented with values in $\{30, 45, 60\}$ for sensitivity analysis. The datasets and codes used in this experimental evaluation were made available for reproducibility.

4.3 Quantitative analysis

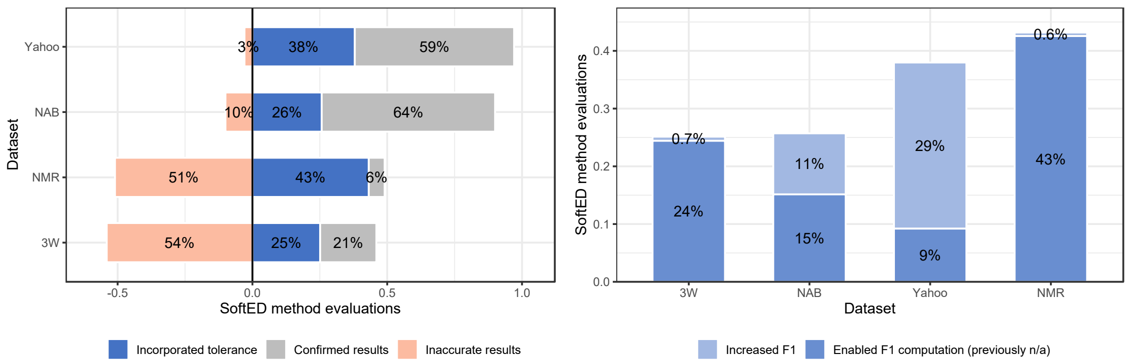
This section presents a quantitative analysis of the SoftED metrics. The main goal of this analysis is to assess the effects of temporal tolerance incorporated by SoftED in event detection evaluation. Four experiments were conducted to compare SoftED metrics against hard metrics and the NAB score.

Experiment 1 The first experiment assesses the number of times SoftED metrics considered more TPs while evaluating detection methods to answer whether SoftED can incorporate temporal tolerance to event detection evaluation. For that, we are interested in comparing SoftED F1 and hard F1 metrics. Time series detections where SoftED F1 was higher than its corresponding hard metric, represent the incorporation of temporal tolerance. In contrast, detection results that maintained an unchanged F1 score had their evaluation confirmed, representing either perfect Recall scenarios in which no tolerance is needed or scenarios with a low rate of neighboring detections in which there are few opportunities for tolerance. Finally, there are inaccurate results, with detections that did not allow temporal tolerance, presenting zero Precision/Recall, and no detections were sufficiently close to events given the defined tolerance level ($k = 15$). In the latter case, the F1 metric cannot be computed.

Figure 4a compares SoftED F1 and hard F1 metrics for each adopted dataset. In blue are presented the percentage of time series detections where SoftED F1 was able to incorporate temporal tolerance. SoftED metrics incorporated temporal tolerance in over 43% (NMR) and at least 25% (3W) of detection method evaluations in all datasets. In total,

[3]The GECCO dataset is provided by the R-package EventDetectR [40].

[4]The NMR dataset was produced by Fiocruz and comprised data on neonatal mortality in Brazilian health facilities from 2005 to 2017. It is publicly available at `https://doi.org/10.7303/syn23651701`.



(a) Incorporated temporal tolerance

(b) Detailed temporal tolerance opportunities

Figure 4: Incorporated temporal tolerance from SoftED F1 metric evaluation of event detection methods compared to hard F1 metric.

36% of the overall conducted time series detections were more tolerantly evaluated (in blue). Furthermore, 45% of all detection results had their evaluation confirmed (in gray), maintaining an unchanged F1 score, reaching a maximum of 64% for the NAB dataset and a minimum of 6% for the NMR dataset. Finally, other 19% of the overall results corresponded to inaccurate detections that did not allow temporal tolerance (in red). The percentages of inaccurate results (F1 n/a) for each dataset are also given in Figure 4a in red.

Figure 4b shows the cases of incorporated temporal tolerance in detail. The datasets NAB and Yahoo presented an increase in F1 in 11% and 29% of the cases, respectively (lighter blue). The other respective 15% and 9% are cases in which methods got no TPs, presenting zero Precision/Recall and non-applicable F1 based on hard metrics (darker blue). Nonetheless, SoftED could score sufficiently close detections, enabling the evaluation of such methods. This is also the case for almost all evaluations of the 3W and NMR datasets that had incorporated temporal tolerance. In fact, in total 17% of the overall conducted detection evaluations could not have been made without SoftED metrics incorporating temporal tolerance.

It is possible to observe by Figure 4 that datasets that contained more anomalies (NAB and Yahoo) got more accurate detection results. It occurs as most adopted methods are designed for anomaly detection. Also, the number of anomaly events in their time series gives several opportunities for incorporating temporal tolerance and increasing F1. On the other hand, 3W and NMR datasets, containing only one or two change points per series, got a higher rate of inaccurate detections. These results indicate that change points pose a particular challenge for detection evaluation. SoftED metrics contribute by incorporating temporal tolerance whenever possible and scoring methods that could be disregarded.

Experiment 2 The second experiment focuses on whether the temporal tolerance incorporated by SoftED can affect the selection of different detection methods. For that, we measured the number of times the use of SoftED metrics as criteria changed the ranking of the best-evaluated detection methods. Figure 5 presents the changes in the top-ranked methods for each time series based on the SoftED F1 metric compared to the hard F1. For all datasets, there were changes in the best-evaluated detection method (Top 1) in over 74% (NMR) or at least 6% (Yahoo) of the cases (in blue), affecting the recommendation of the most suitable detection method for their time series. While the most accurate results maintained their top position (in dark gray), over all adopted time series, 31% of detection methods that could have been dismissed became the most prone to selection.

Furthermore, SoftED metrics also caused changes in the second (Top 2) and third-best (Top 3) evaluated methods. Percentages for each dataset are depicted in Figure 5, however, over all adopted time series, 24% of the methods in the Top 2 climbed to that position, while 16% dropped to that position when other methods assumed the Top 1 (in light gray). For methods in the Top 3, 23% climbed to the position, and in 24% of the cases, they were pushed down by methods that climbed to the first two rank positions. Due to the higher rates of perfect Recall results in the NAB and Yahoo datasets (Figure 4), most of the methods applied maintained their ranking positions at the top. In contrast, 3W and NMR datasets presented more changes in ranking based on the SoftED metrics, affecting the selection of suitable methods, especially for change point detection.

Experiment 3 Once analyzed the incorporated temporal tolerance and its effects in the ranking of detection methods, the third experiment encompasses a sensitivity analysis to answer the question of how SoftED is affected by different

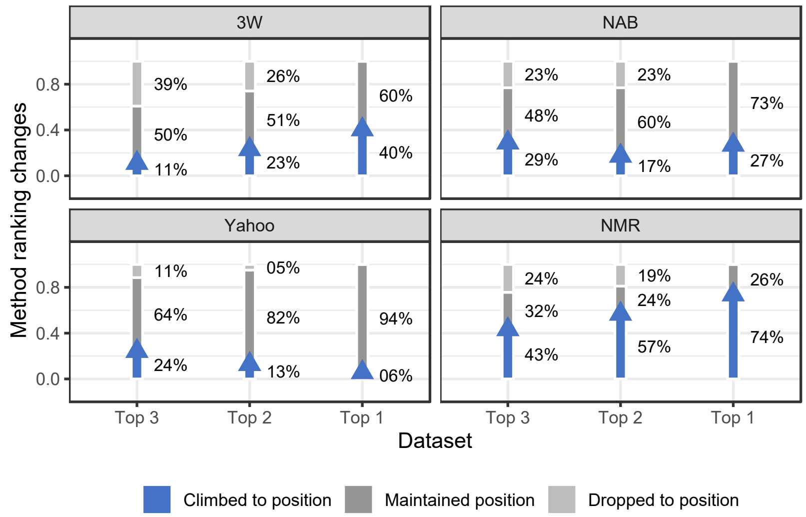


Figure 5: Changes in the ranking of top evaluated event detection methods based on the SoftED F1 metric compared to hard F1 metric

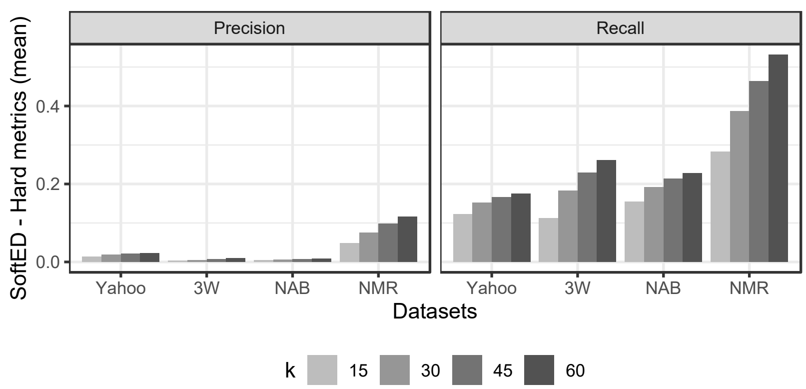


Figure 6: Average difference between SoftED and hard Precision and Recall metrics given different levels of temporal tolerance

levels of temporal tolerance. The temporal tolerance level of SoftED metrics is given by the k constant set to 30, 45, and 60, besides the minimum value of 15 as in the previous experiments. Figure 6 presents the average difference between SoftED and hard Precision and Recall metrics given the different levels of temporal tolerance for each dataset. Overall, as temporal tolerance increases, more TPs were considered, and metrics increased in value, which means the detection methods were more tolerantly evaluated. In particular, higher levels of temporal tolerance lead to a decrease in the number of FNs, which most directly affected Recall values.

Experiment 4 The last quantitative experiment aims to answer the following question: whether there is a difference between the temporal tolerance incorporated by SoftED and the tolerance incorporated by NAB score anomaly windows. For that, we measured the number of times the NAB F1 metrics, derived from the NAB scoring algorithm, considered more TPs than hard F1 metrics while evaluating detection methods. This measure is compared against the tolerance incorporated by SoftED metrics presented in Experiment 1 (Figure 4a). For datasets 3W, NAB, and Yahoo, NAB increased the incorporated tolerance at 42%, 40%, and 39%, respectively. Whereas, for the NMR dataset, the percentage of incorporated tolerance decreased by 6%.

NAB metrics were more tolerant than SoftED in method evaluations over most datasets, which does not mean better. The tolerance level incorporated by NAB depends directly on the anomaly window size, which is automatically set by the algorithm. Table 3 presents the interval and the average of the anomaly window sizes set for the time series of each dataset. While the tolerance level given by SoftED was consistently set by $k = 15$, giving a tolerance window of 30 observations, the NAB anomaly windows were mostly wider, reaching a maximum of $12,626$ observations or $1,357$ on average for the 3W dataset. Wider anomaly windows allow a greater number of hard FPs to be considered TPs, which causes F1 metrics to increase in value. It is similar to what was discussed in Experiment 3, explaining the increase in tolerance opportunities. The inverse is also true, as exemplified by the NMR dataset, for which anomaly windows did not surpass 14 observations, decreasing the number of tolerance opportunities compared to SoftED.

Table 3: Summary of anomaly window sizes automatically set by the NAB scoring algorithm for each dataset

Dataset	**Anomaly window sizes**	
	Interval	**Mean**
3W	[52, 12626]	1357
NAB	[0, 902]	286
Yahoo	[0, 168]	38
NMR	[0, 14]	13

Overall, wider anomaly windows caused the NAB score to increase its computation time by three orders of magnitude on average compared to hard metrics (1 versus 1×10^{-3} seconds). In contrast, SoftED increased metrics computation time by only one order of magnitude higher than the hard metrics (1×10^{-2} versus 1×10^{-3} seconds). Moreover, it is important to note from Table 3 that the anomaly window size computation proposed by the NAB algorithm allows the definition of zero-sized windows, which do not give any tolerance to inaccuracy as in hard metrics or narrow windows which are not enough to hold the central limit theorem guaranteed in SoftED results. On the other hand, the NAB window size automatic definition is not domain-dependent. Consequently, domain specialists may find windows too wide or too narrow for their detection application, making the incorporated tolerance and metric results non-applicable or at least difficult to interpret. In this context, SoftED contributes by allowing domain specialists to define the desired temporal tolerance level for their detection method results.

4.4 Qualitative analysis

This section presents a qualitative analysis of SoftED metrics and the scenarios in which they bring the most contribution compared to hard metrics and the NAB score. To this end, we have surveyed 13 specialists from three domains: oil exploration, public health, and weather monitoring. We have interviewed 3 specialists from Petrobras (Brazil oil company), 5 specialists from the Oswaldo Cruz Foundation (Fiocruz), linked to the Brazilian Ministry of Health, the most prominent institution of science and technology applied to health in Latin America, and 5 weather forecast specialists from the Rio Operations Center (COR) of the City Hall of Rio de Janeiro. All interviewed specialists work on the problem of time series analysis and event detection daily. Furthermore, we have also surveyed other 57 student volunteers from the Federal Center for Technological Education of Rio de Janeiro (CEFET/RJ) and the National Laboratory for Scientific Computing (LNCC), totalizing 70 participants.

The survey addressed the problem of selecting the most suitable method in six experiments, each representing a particular event detection scenario. Two event detection methods (A and B) were applied to a representative time series of the GECCO, 3W, or NMR datasets for each experiment. The plots of the detection results were presented to participants as in Figure 7, where blue dots represent events, red dots represent detections, and green dots represent detections that match events. Moreover, we presented the participants with Table 4, containing detection evaluation metrics computed for methods A and B for each experiment scenario, namely the F1 metric, in its hard and SoftED versions and the NAB score. Values that maximize each metric and could be used for recommending a particular method are underlined.

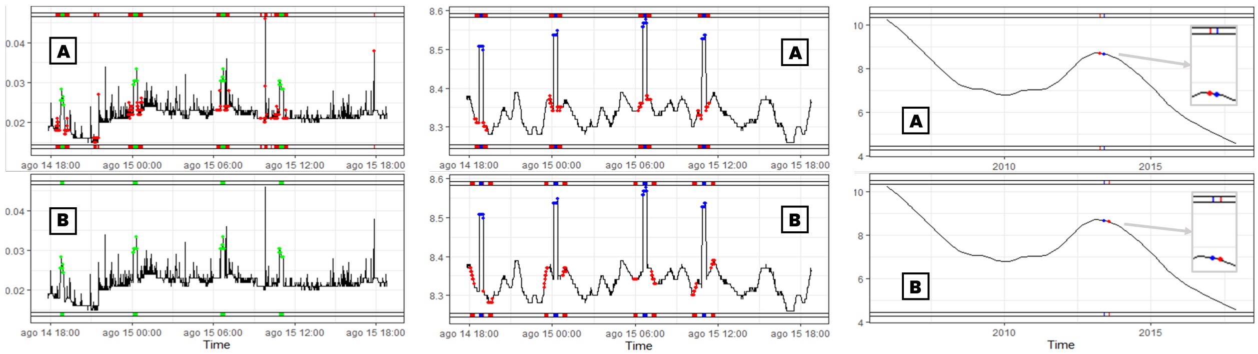
Given the results of both methods, we asked the participants to analyze the plots in Figure 7 and answer, for each experiment, the first question of the survey:

Question 1 - Which event detection method performed better?

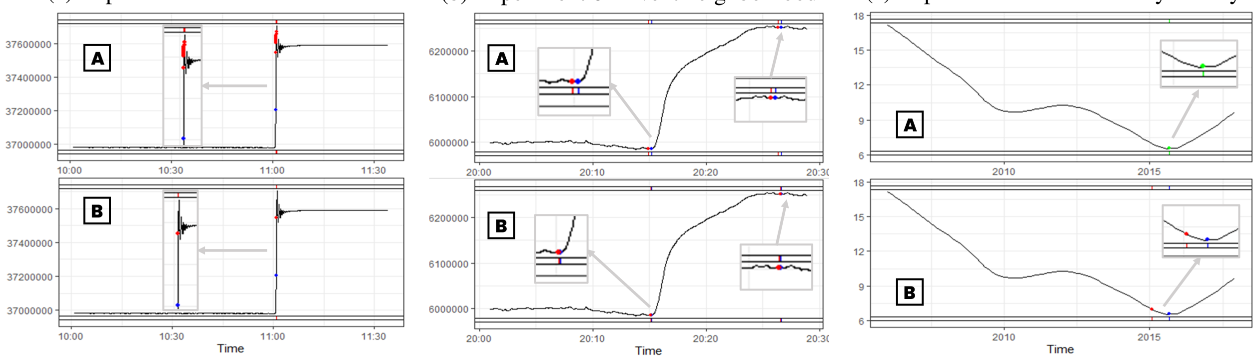
Question 1 was closed with three disjoint options: *Method A*, *Method B*, or *None*. The main goal of Question 1 was to get the specialists' intuitive and personal opinions of the most suitable detection method for selection on that particular application scenario.

Next, we asked the participants to analyze the metrics in Table 4 and answer, for each experiment, the second question of the survey:

Question 2 - Which metric corroborates with your opinion?



(a) Experiment 5 - Perfect Recall (b) Experiment 6 - Event neighborhood (c) Experiment 7 - Detection symmetry



(d) Experiment 8 - Number of detections (e) Experiment 9 - Detection distances (f) Experiment 10 - Detection bias

Figure 7: Detection results of experiments for qualitative analysis, each representing a different scenario of comparison of two given event detection methods (A and B). Blue dots refer to time series events, red dots refer to method detections, and green dots refer to detections that coincide with events. (a) and (b) show time series of the GECCO dataset (variables Trueb and pH). (c) and (f) show time series of the NMR dataset (health facilities code 2080052 and 2295407). (d) and (e) show time series of the 3W dataset (event type 2, variable P-PDG and event type 6, variable P-MON-CKP).

Table 4: Event detection metrics for methods A and B for each experiment scenario. Values that could be used for recommending a particular method are underlined.

Experiment	**Method**	**Metric**		
		F1		**NAB score**
		Hard	**SoftED**	
5	A	0.43	0.43	16.05
	B	1	1	35.87
6	A	n/a	0.07	- 36.53
	B	n/a	0.01	- 50.17
7	A	n/a	0.87	0.94
	B	n/a	0.87	0.77
8	A	n/a	0.12	0.85
	B	n/a	0.6	0.85
9	A	n/a	0.07	1.89
	B	n/a	0.87	1.76
10	A	1	1	0.88
	B	n/a	0.53	1

Question 2 was also closed with three joint options: *F1*, *NAB score*, or *Other*. The main goal of Question 2 was to assess the metrics (and corresponding evaluation approach) that would further the selection of the most suitable detection method in that particular application scenario, according to specialist opinion.

Table 5 presents the domain specialists' responses to the survey questions for each experiment scenario. Their winning responses are underlined. Furthermore, student volunteers winning responses are given to study how specialist opinion compares with common sense. All participants were also allowed to comment and elaborate on their responses for each experiment in an open question. The remainder of this section further discusses the results of each survey experiment.

Table 5: Domain specialists' responses to the survey questions for each experiment scenario. The winning responses are underlined. Volunteers winning responses are also given for comparison with the non-specialist common sense.

Experiment	Specialists responses						Volunteer winning responses	
	Question 1			Question 2			Question 1	Question 2
	Method A	Method B	None	F1	NAB Score	Other		
5	1 (8%)	<u>12 (92%)</u>	0 (0%)	<u>12 (92%)</u>	6 (46%)	1 (8%)	Method B (84%)	F1 (96%)
6	<u>11 (84%)</u>	1 (8%)	1 (8%)	<u>12 (92%)</u>	7 (54%)	1 (8%)	Method A (88%)	F1 (96%)
7	<u>12 (92%)</u>	0 (0%)	1 (8%)	4 (31%)	<u>12 (92%)</u>	0 (0%)	Method A (84%)	NAB Score (82%)
8	0 (0%)	<u>12 (92%)</u>	1 (8%)	<u>13 (100%)</u>	1 (8%)	0 (0%)	Method B (86%)	F1 (98%)
9	3 (23%)	<u>10 (77%)</u>	0 (0%)	<u>11 (85%)</u>	5 (38%)	0 (0%)	Method B (74%)	F1 (86%)
10	<u>10 (77%)</u>	3 (23%)	0 (0%)	<u>10 (77%)</u>	6 (46%)	0 (0%)	Method A (82%)	F1 (89%)

Experiment 5 The first survey experiment refers to a scenario of perfect Recall, where all events contained in a time series of the GECCO dataset were detected by both Method A and Method B. However, Method A presents more detections (in red). From Table 5, we observe that almost all specialists (12/13) agreed that Method B performed better. According to them, Method B managed to minimize FPs, presenting a higher Precision rate, indicated by the F1 metric, which was also the winning response for Question 2 with 12/13 votes. For this experiment, both hard and SoftED F1 give the same evaluation of Method B so that both approaches can be used for recommendation.

Nonetheless, 6 specialists (46%) also selected the NAB score, which also corroborates with the recommendation of Method B. Other specialists said they preferred not to select the NAB score, as they were unfamiliar with the metric and wanted to avoid drawing any conclusions with this experiment. Overall, all computed metrics corroborated with specialists' opinions recommending Method B as the most suitable for the application. This result indicates that the evaluation of decidedly good detection performances based on SoftED and other state-of-the-art metrics available in the literature is still valid.

Experiment 6 The second survey experiment is based on another time series taken from the GECCO dataset. It addresses the scenario in which Method A and Method B presented detections that, despite not coinciding with the events contained in the series, are in the surroundings or the neighborhood of the events. Furthermore, Method A and Method B detections differ in the distance to events. In this case, most specialists (11/13) agreed to select Method A as giving the best detection performance, for their detections are temporally closer to the events. Both metrics, F1 and NAB score, corroborated with specialists' opinions recommending Method A, while F1 was the winning response to Question 2. At this point, it is important to note that the hard approach to F1 computation can no longer give an evaluation for the methods, as both results had no Precision or Recall. Hence, the winning response for Question 2 regards the F1 metric produced by the SoftED approach as the one that furthers the selection of the best detection performance according to specialists.

Experiment 7 The third survey experiment is based on a time series from the NMR dataset containing monthly neonatal mortality rates for a healthcare facility in Brazil over the years. In this scenario, Method A and Method B produced only one detection close to the event contained in the series. The detections of Method A and Method B are symmetric. They have the same distance from the event and differ only in whether they come before or after it, respectively. In this experiment, almost all specialists (12/13) responded that Method A gave the best detection performance, as it seems to anticipate the event, allowing time to take prior needed actions. Furthermore, as there was a tie regarding the F1 metrics, the NAB score was the winning response, corroborating with specialists' opinions.

However, a public health specialist from Fiocruz disagreed and responded that none of the methods performed better, which is corroborated by the F1 metrics. For example, consider implementing a public health policy in which a human milk bank is supposed to decrease neonatal mortality rates. Although it makes sense to detect the first effects of preparing for the implementation of the policy, it may not be reasonable to give greater weight to anticipated detections rather than the detection of the effects after the implementation. They defend:

> It is important to deepen the understanding of the context of the event and the reach of its effects (before and after).

Experiment 8 The fourth survey experiment is based on a time series taken from the 3W dataset produced by Petrobras. In this scenario, Method A and Method B presented detections close to the event contained in the series, differing only in the number of detections made. The closest detections for both methods have the same distance from the event. For this experiment, except for one specialist that responded *None* to Question 1, all specialists agree that Method B performed better. As it minimizes the overall FPs, it increases Precision which conditions F1, the winning response of Question 2, selected by 100% of the specialists. The NAB score indicates a tie between both methods, therefore, not penalizing the excess FPs, and the hard F1 does not provide any method evaluation. In this case, the SoftED F1 metric is the only one that corroborates with specialists' opinions.

Experiment 9 The fifth survey experiment is based on another time series taken from the 3W dataset. This scenario addresses the problem of evaluating methods based on their detection proximity to events. In this experiment, both Method A and Method B presented a detection close and antecedent to the two events contained in the series. Method A and Method B differ only concerning the distance of their detections to the events. Most specialists (10/13) agreed that Method B performed better, as they say:

> It seems reasonable to give greater weight to detections closer to the actual events.

Again, the only metric that corroborated with the specialists' opinion was the F1 from SoftED. Furthermore, specialists mentioned that SoftED F1 was approximately 12 times greater for Method B than for Method A, while the difference in the NAB score did not seem high enough to give the same confidence in results from Method A.

Experiment 10 Finally, we used another time series of neonatal mortality rates from the NMR dataset for the sixth and final survey experiment. This experiment addresses the problem of detection bias in detection evaluation. Method A and Method B produced a detection related to the event contained in the series. However, Method A and Method B detections differ regarding their distance to the event. Method A managed to correctly detect the time series event, while the detection of Method B came close before the event. Most specialists (10/13) agreed that Method A performed better since it produced, for all intents and purposes, a TP, presenting perfect Recall and perfect Precision. On the other hand, the evaluation of Method B depended solely on the incorporation of temporal tolerance.

As metrics disagree with the recommendation, the F1 metric again corroborates with the specialists' opinion, being the winning response for Question 2. In particular, the SoftED F1 metric is the only approach that recommends Method A. The hard F1 metric cannot be computed for Method B, being incomparable. The difference in metric values of SoftED is also greater than for the NAB score, increasing confidence in the recommendation.

4.5 Summary of results and discussion

Given different detection evaluation scenarios, the majority of the surveyed domain specialists agreed that the most desired detection method for selection was the one that minimizes FPs and FNs, giving higher Precision and Recall rates, while also producing detections that are temporally closer to the events. In this context, the F1 was the metric most corroborated with specialists' opinions for 5 of the 6 experiments. In particular, the SoftED F1 metric was the only one that furthered the selection of the most desired detection method according to specialists in four experiments, only tying with hard F1 in Experiment 5. To elaborate, a domain specialist from Fiocruz argued that:

> For health policies, for example, the SoftED approach seems to make more sense since the hard and NAB approaches do not seem adequate for events that produce prior and subsequent effects that may have a gradual and even non-monotonous evolution.

Volunteer winning responses in Table 5 also indicate that common sense does not differ from specialist opinion, which means the contribution of SoftED metrics is noticeable even to a wider and non-specialist research public.

There was still one experiment (Experiment 7) where the NAB score was the metric that most corroborated with specialists. They claimed that for their usual detection application, it is interesting to have FPs (warnings) before the event, so there is a time window for measures to be taken to prevent any of its unwelcome effects. Also, the longer the time window set by FPs preceding the event, the better, as there is more valuable time to take preventive actions. All specialists consistently presented this argument that either disagreed regarding Question 1 or gave the NAB score as a response for Question 2 over all experiments. This argument demands a deeper discussion.

At this point, it is important to mention that to avoid bias in the responses, the discussion regarding our motivating example of Section 1.1 was not presented prior to the interviews. Hence, detections that preceded the events were misconceived as event predictions [70]. However, this problem was not in the scope of our experiments. Also, as discussed in Section 1.1, detections preceding events can be made passed their occurrence. Evaluating methods that anticipate events is not about how temporally distant a preceding detection is from the events. However, it is actually about the time lag necessary for a method to detect the event accurately. The misconception regarding detections that preceded the events was addressed in detail by the end of the survey interviews.

After discussion and deliberation, all disagreeing specialists changed their opinion and sided with the majority that evaluated methods regarding contexts (i) and (ii). Also, all specialists rethought their responses for Experiment 7. Finally, all domain specialists agreed that:

> SoftED metrics contribute to the problem of detection method evaluation and selection in different domains.

They allow the assessment of the adequacy of a method for a time series event detection application regarding the quality of its detections and its ability to approximate the events compared to other methods. Furthermore, the specialists see the evaluation regarding detection lags, that is, the analysis of its ability to anticipate (or not) the events as complementary.

5 Conclusions

This paper introduced the SoftED metrics, new softened versions of the standard classification metrics designed to incorporate temporal tolerance in evaluating the performance of methods for detecting events in time series applications. SoftED metrics support the comparative analysis of methods based on their ability to accurately produce detections of interest to the user, given their desired tolerance level, both accurately and neighboring the events.

The SoftED metrics were quantitatively and qualitatively evaluated and compared against the current state-of-the-art in detection scoring methods. They incorporated temporal tolerance in event detection, enabling evaluations that could not have been made without them while also confirming accurate results. Consequently, SoftED metrics changed evaluation rankings, causing detection methods that could be disregarded to become the top best evaluated and most prone to selection. Moreover, surveyed domain specialists noted the contribution of SoftED metrics to the problem of detection method evaluation in different domains. In particular, SoftED metrics were the only metrics able to improve the selection of the most desired detection method, according to specialists in most experimental scenarios.

Specialists suggest that SoftED metrics are particularly adequate for evaluating detections of domain events that produce prior and subsequent effects of gradual or non-monotonous evolution. At the same time, they can also be used to benchmark different initial conditions, parameters, and threshold values, whose definition is one of the main challenges for event detection algorithms [39]. Furthermore, analyzing a method's ability to anticipate (or not) the events is complementary after the evaluation enabled by SoftED metrics.

Acknowledgments

The authors thank CNPq, CAPES (finance code 001), FAPERJ, and CEFET/RJ for partially funding this research.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147, 2017.

[2] M. Ahmed, A. Mahmood, and M. Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2016.

[3] E. Alevizos, A. Skarlatidis, A. Artikis, and G. Paliouras. Probabilistic complex event recognition: A survey. *ACM Computing Surveys*, 50(5), 2017.

[4] S. Aminikhanghahi and D. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017.

[5] R. Ariyaluran Habeeb, F. Nasaruddin, A. Gani, I. Targio Hashem, E. Ahmed, and M. Imran. Real-time big data processing for anomaly detection: A Survey. *International Journal of Information Management*, 45:289–307, 2019.

[6] A. Blázquez-García, A. Conde, U. Mori, and J. Lozano. A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys*, 54(3), 2021.

[7] P. Boniol and T. Palpanas. Series2graph: Graph-based subsequence anomaly detection for time series. *Proceedings of the VLDB Endowment*, 13(11):1821–1834, 2020.

[8] P. Boniol, J. Paparrizos, Y. Kang, T. Palpanas, R. Tsay, A. Elmore, and M. Franklin. Theseus: Navigating the Labyrinth of Time-Series Anomaly Detection. *Proceedings of the VLDB Endowment*, 15(12):3702–3705, 2022.

[9] P. Boniol, J. Paparrizos, T. Palpanas, and M. Franklin. Sand in action: Subsequence anomaly detection for streams. *Proceedings of the VLDB Endowment*, 14(12):2867–2870, 2021.

[10] M. Braei and S. Wagner. Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art, apr 2020.

[11] R. Carmona. *Statistical Analysis of Financial Data in R*. Springer Science & Business Media, dec 2013.

[12] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 2009.

[13] V. Chandola, V. Mithal, and V. Kumar. Comparative evaluation of anomaly detection techniques for sequence data. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 743–748, 2008.

[14] V. Chauhan, K. Dahiya, and A. Sharma. Problem formulations and solvers in linear SVM: a review. *Artificial Intelligence Review*, 52(2):803–855, 2019.

[15] D. Choudhary, A. Kejariwal, and F. Orsini. On the Runtime-Efficacy Trade-off of Anomaly Detection Techniques for Real-Time Streaming Data, oct 2017.

[16] A. Cook, G. Misirli, and Z. Fan. Anomaly Detection for IoT Time-Series Data: A Survey. *IEEE Internet of Things Journal*, 7(7):6481–6494, 2020.

[17] D. Cook and N. Krishnan. *Activity learning: Discovering, recognizing, and predicting human behavior from sensor data*. 2015.

[18] M. Dancho and D. Vaughan. anomalize: Tidy Anomaly Detection. Technical report, https://cran.r-project.org/web/packages/anomalize/vignettes/anomalize_quick_start_guide.html, 2020.

[19] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: Experimental comparison of representations and distance measures. In *Proceedings of the VLDB Endowment*, volume 1, pages 1542–1552, 2008.

[20] X. Ding, Y. Li, A. Belatreche, and L. Maguire. An experimental evaluation of novelty detection methods. *Neurocomputing*, 135:313–327, 2014.

[21] G. Dudek. Neural networks for pattern-based short-term load forecasting: A comparative study. *Neurocomputing*, 205:64–74, 2016.

[22] L. Escobar, R. Salles, J. Lima, C. Gea, L. Baroni, A. Ziviani, P. Pires, F. Delicato, R. Coutinho, L. Assis, and E. Ogasawara. Evaluating Temporal Bias in Time Series Event Detection Methods. *Journal of Information and Data Management*, 12(3), oct 2021.

[23] H. Fanaee-T and J. Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2-3):113–127, 2014.

[24] R. Fontugne, P. Borgnat, P. Abry, and K. Fukuda. MAWILab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking. In *Proceedings of the 6th International Conference on Emerging Networking Experiments and Technologies, Co-NEXT'10*, 2010.

[25] A. Gammerman and V. Vovk. Hedging predictions in machine learning. *Computer Journal*, 50(2):151–163, 2007.

[26] F. Gmati, S. Chakhar, W. Chaari, and M. Xu. A taxonomy of event prediction methods. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11606 LNAI:12–26, 2019.

[27] T. Guo, J. Dong, H. Li, and Y. Gao. Simple convolutional neural network on image classification. In *2017 IEEE 2nd International Conference on Big Data Analysis, ICBDA 2017*, pages 721–724, 2017.

[28] M. Gupta, J. Gao, C. Aggarwal, and J. Han. Outlier Detection for Temporal Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2250–2267, 2014.

[29] V. Guralnik and J. Srivastava. Event detection from time series data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 33–42, New York, NY, USA, 1999. Association for Computing Machinery.

[30] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. 2012.

[31] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[32] S. Ismaeel, A. Miri, and D. Chourishi. Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis. In *2015 IEEE Canada International Humanitarian Technology Conference, IHTC 2015*, 2015.

[33] A. Iturria, J. Carrasco, F. Herrera, S. Charramendieta, and K. Intxausti. otsad: Online Time Series Anomaly Detectors. Technical report, https://cran.r-project.org/web/packages/otsad/, 2019.

[34] V. Jacob, F. Song, A. Stiegler, B. Rad, Y. Diao, and N. Tatbul. Exathlon: A benchmark for explainable anomaly detection over time series. *Proceedings of the VLDB Endowment*, 14(11):2613–2626, 2021.

[35] A. Lavin and S. Ahmad. Evaluating Real-Time Anomaly Detection Algorithms – The Numenta Anomaly Benchmark. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 38–44, dec 2015.

[36] V. Lawhern, W. Hairston, and K. Robbins. DETECT: A MATLAB Toolbox for Event Detection and Identification in Time Series, with Applications to Artifact Detection in EEG Signals. *PLoS ONE*, 8(4), 2013.

[37] B. Lim and S. Zohren. Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 2021.

[38] J. Lima, R. Salles, F. Porto, R. Coutinho, P. Alpis, L. Escobar, E. Pacitti, and E. Ogasawara. Forward and Backward Inertial Anomaly Detector: A Novel Time Series Event Detection Method. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2022-July, 2022.

[39] S. Molaei and M. Keyvanpour. An analytical review for event prediction system on time series. In *2015 2nd International Conference on Pattern Recognition and Image Analysis, IPRIA 2015*, 2015.

[40] S. Moritz and F. Rehbach. EventDetectR. Technical report, https://cran.r-project.org/web/packages/EventDetectR/index.html, 2020.

[41] M. Munir, M. Chattha, A. Dengel, and S. Ahmed. A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data. In *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, pages 561–566, 2019.

[42] A. Muniyandi, R. Rajeswari, and R. Rajaram. Network anomaly detection by cascading k-Means clustering and C4.5 decision tree algorithm. In *Procedia Engineering*, volume 30, pages 174–182, 2012.

[43] G. Pang, C. Shen, L. Cao, and A. Hengel. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2), 2021.

[44] J. Paparrizos, P. Boniol, T. Palpanas, R. Tsay, A. Elmore, and M. Franklin. Volume Under the Surface: A New Accuracy Evaluation Measure for Time-Series Anomaly Detection. *Proceedings of the VLDB Endowment*, 15(11):2774–2787, 2022.

[45] M. Pimentel, D. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[46] Rahul and B. Choudhary. An Advanced Genetic Algorithm with Improved Support Vector Machine for Multi-Class Classification of Real Power Quality Events. *Electric Power Systems Research*, 191, 2021.

[47] H. Raza, G. Prasad, and Y. Li. EWMA model based shift-detection methods for detecting covariate shifts in non-stationary environments. *Pattern Recognition*, 48(3):659–669, 2015.

[48] F. Rehbach, S. Moritz, S. Chandrasekaran, M. Rebolledo, M. Friese, and T. Bartz-Beielstein. GECCO Industrial Challenge 2018 Dataset. Technical report, https://zenodo.org/record/3884398#.Y9qShK3MJD8, 2018.

[49] F. Riese and S. Keller. Supervised, semi-supervised, and unsupervised learning for hyperspectral regression. *Advances in Computer Vision and Pattern Recognition*, pages 187–232, 2020.

[50] F. Salfner, M. Lenk, and M. Malek. A survey of online failure prediction methods. *ACM Computing Surveys*, 42(3), 2010.

[51] R. Salles, K. Belloze, F. Porto, P. Gonzalez, and E. Ogasawara. Nonstationary time series transformation methods: An experimental review. *Knowledge-Based Systems*, 164:274–291, 2019.

[52] R. Salles, L. Escobar, L. Baroni, R. Zorrilla, A. Ziviani, V. Kreischer, F. Delicato, P. F. Pires, L. Maia, R. Coutinho, L. Assis, and E. Ogasawara. Harbinger: Um framework para integração e análise de métodos de detecção de eventos em séries temporais. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 73–84. SBC, sep 2020.

[53] S. Schmidl, P. Wenig, and T. Papenbrock. Anomaly Detection in Time Series: A Comprehensive Evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797, 2022.

[54] G. Sebestyen, A. Hangan, Z. Czako, and G. Kovacs. A taxonomy and platform for anomaly detection. In *2018 IEEE International Conference on Automation, Quality and Testing, Robotics, AQTR 2018 - THETA 21st Edition, Proceedings*, pages 1–6, 2018.

[55] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer, apr 2017.

[56] N. Singh and C. Olinsky. Demystifying Numenta anomaly benchmark. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2017-May, pages 1570–1577, 2017.

[57] J.-I. Takeuchi and K. Yamanishi. A unifying framework for detecting outliers and change points from time series. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):482–492, 2006.

[58] J. Tang, C. Deng, and G.-B. Huang. Extreme Learning Machine for Multilayer Perceptron. *IEEE Transactions on Neural Networks and Learning Systems*, 27(4):809–821, 2016.

[59] N. Tatbul, T. Lee, S. Zdonik, M. Alam, and J. Gottschlich. Precision and recall for time series. In *Advances in Neural Information Processing Systems*, volume 2018-December, pages 1920–1930, 2018.

[60] A. Tettamanzi and M. Tomassini. *Soft Computing: Integrating Evolutionary, Neural, and Fuzzy Systems*. Springer Science & Business Media, apr 2013.

[61] R. Vargas, C. Munaro, P. Ciarelli, A. Medeiros, B. Amaral, D. Barrionuevo, J. Araújo, J. Ribeiro, and L. Magalhães. A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering*, 181, 2019.

[62] H. Wang, M. Bah, and M. Hammad. Progress in Outlier Detection Techniques: A Survey. *IEEE Access*, 7:107964–108000, 2019.

[63] Y. Wang, M. Vuran, and S. Goddard. Analysis of event detection delay in wireless sensor networks. In *Proceedings - IEEE INFOCOM*, pages 1296–1304, 2011.

[64] J. Ward, P. Lukowicz, and H. Gellersen. Performance metrics for activity recognition. *ACM Transactions on Intelligent Systems and Technology*, 2(1), 2011.

[65] Y. Webscope. Labeled anomaly detection dataset. Technical report, https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70, mar 2015.

[66] P. Wenig Phillip.Wenig@Hpi.De, S. Schmidl Sebastian.Schmidl@Hpi.De, and T. Papenbrock. TimeEval: A Benchmarking Toolkit for Time Series Anomaly Detection Algorithms. *Proceedings of the VLDB Endowment*, 15(12):3678–3681, 2022.

[67] X.-B. Yan, L. Tao, Y.-J. Li, and G.-B. Cui. Research on event prediction in time-series data. In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, volume 5, pages 2874–2878, 2004.

[68] L. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[69] Y. Zhang, N. Meratnia, and P. Havinga. Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys and Tutorials*, 12(2):159–170, 2010.

[70] L. Zhao. Event Prediction in the Big Data Era: A Systematic Survey. *ACM Computing Surveys*, 54(5), 2021.

[71] Y. Zhou, R. Arghandeh, H. Zou, and C. Spanos. Nonparametric Event Detection in Multiple Time Series for Power Distribution Networks. *IEEE Transactions on Industrial Electronics*, 66(2):1619–1628, 2019.