Statistical Evaluation of Anomaly Detectors for Sequences

Erik Scharwächter
scharwaechter@bit.uni-bonn.de
University of Bonn, Germany

Emmanuel Müller
mueller@bit.uni-bonn.de
University of Bonn, Germany

ABSTRACT

Although precision and recall are standard performance measures for anomaly detection, their statistical properties in sequential detection settings are poorly understood. In this work, we formalize a notion of precision and recall with temporal tolerance for point-based anomaly detection in sequential data. These measures are based on time-tolerant confusion matrices that may be used to compute time-tolerant variants of many other standard measures. However, care has to be taken to preserve interpretability. We perform a statistical simulation study to demonstrate that precision and recall may overestimate the performance of a detector, when computed with temporal tolerance. To alleviate this problem, we show how to obtain null distributions for the two measures to assess the statistical significance of reported results.

KEYWORDS
anomaly detection, precision, recall, confusion matrix, simulations

ACM Reference Format:
Erik Scharwächter and Emmanuel Müller. 2020. Statistical Evaluation of Anomaly Detectors for Sequences. In MileTS ’20: 6th KDD Workshop on Mining and Learning from Time Series, August 24th, 2020, San Diego, California, USA. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
While the anomaly score encodes the feature with increasing The function \( I(t) \) actually no anomaly (AnA), predicted anomaly (PA), predicted no anomaly (PnA). The confusion matrix partitions the observations so that every observations falls in exactly one category. Many performance measures can be computed from confusion matrices [11], typically by normalizing individual entries by marginal sums. The measures are interpretable because all entries and marginals have straightforward interpretations.

When introducing temporal tolerance in the confusion matrix, we have to make sure that the result is still a partition with interpretable entries and marginals. Tables 1 and 2 show the confusion matrices obtained when introducing temporal tolerance either into the ground-truth time steps or the predicted time steps, using
They show that in this case PA

They key question when analyzing evaluation measures from a

Müller [14] have generalized the formal analysis for a larger class of

anomalies follow independent Bernoulli processes. In this case, both

true positives (or any other entry of the confusion matrix) is

low some probability distribution. Only if the reported number of

the quantities in the confusion matrix as

random variables

Statistical Evaluation of Anomaly Detectors for Sequences MileTS ’20, August 24th, 2020, San Diego, California, USA

2.3 Statistical significance

They key question when analyzing evaluation measures from a

statistical point of view is whether the reported values are statisti-

cally significant. To assess statistical significance, we have to treat

the quantities in the confusion matrix as random variables that fol-

low some probability distribution. Only if the reported number of

ture positives (or any other entry of the confusion matrix) is much

larger (or smaller) than expected due to random coincidences, the

result should be considered statistically significant. Donges et al.

[7] have derived the probability distribution of the two types of

ture positives (PA-AAδ and PAδ-AA) from Tables 1 and 2, under

the assumption that the ground-truth anomalies and the predicted

anomalies follow independent Bernoulli processes. In this case, both

quantities follow simple binomial distributions. Scharwächter and

Müller [14] have generalized the formal analysis for a larger class of

problems, where the anomaly score is a strictly stationary process.

They show that in this case PAδ-AA from Table 2 also follows a

binomial distribution, where the success probability can be approxi-
mated using a result from Extreme Value Theory [5]. Unfortunately,

there is no analogous derivation for PA-AAδ from Table 1 under the

strict stationarity assumption. In this work, we do not use the

existing analytical results, but perform Monte Carlo simulations to

estimate the required probability distributions without potentially

limiting assumptions on the data generating processes.

3 SIMULATION STUDY

We now use the anomaly score (z_{t})_{t=1,...,T} from the earthquake de-

tection example in Section 1 and compute time-tolerant confusion

matrices, as well as the time-tolerant precision and recall measures,

for randomized ground-truth sequences of anomalies. We evaluate

the anomaly score against 10,000 random permutations of the
ground-truth sequence of anomalies (\epsilon_{t})_{t=1,...,T} from the exam-

ple. In doing so, we keep the number of ground-truth anomalies

constant and assume that they follow a Bernoulli process. We be-

lieve that this assumption is reasonable for ground-truth anomalies, which

typically occur rarely and are not clustered.

3.1 Monte Carlo precision and recall

First, we visualize the precision and recall values obtained from a

subset of 100 random permutations for various temporal tolerances

and thresholds in Figure 3. The visualization also shows the per-

formance measures observed on the non-permuted ground-truth

sequence of anomalies. The observed precision and recall values on the non-permuted

sequence are generally higher than the values from the randomly

permuted sequences, especially at larger thresholds. This confirms

Table 1: Relaxed confusion matrix for sequential data, with tolerance in ground-truth

| AAδ | AnAA |
|-----|------|
| PA  | \sum_{t} \left( \sum_{t'=t+\delta}^{t+\delta} \epsilon_{t'} > 0 \right) I(z_{t'} \geq \tau) |
| PnA | \sum_{t} \left( \sum_{t'=t+\delta}^{t+\delta} \epsilon_{t'} > 0 \right) \left( 1 - I(z_{t'} \geq \tau) \right) |

Table 2: Relaxed confusion matrix for sequential data, with tolerance in predictions

| AA | AnAA |
|-----|------|
| PA  | \sum_{t} \epsilon_{t} I(\sum_{t'=t+\delta}^{t+\delta} I(z_{t'} \geq \tau) > 0) |
| PnA | \sum_{t} \left( 1 - \epsilon_{t} \right) I(\sum_{t'=t+\delta}^{t+\delta} I(z_{t'} \geq \tau) > 0) |

AAδ: actual anomaly with tolerance δ, AnAA: actually no anomaly with tolerance δ, PA: predicted anomaly, PnA: predicted no anomaly; we use zero-padding at the boundaries.
and recall $R_\delta$ independent of the anomaly score. Statistically, the anomaly score allows detection of anomalies that are not reflected in the temporal tolerance, which may not reflect the actual performance of the anomaly detector. In the worst case, one might conclude that the anomaly score contains useful information on earthquake occurrences. However, when the temporal tolerance is increased, the gap between the simulated and the observed performance measures tends to shrink: the performance measures on the simulated data. We have shown that these measures are computed from two distinct time-tolerant confusion matrices. Time-tolerant confusion matrices can, in principle, be used to derive time-tolerant variants of other well-known measures. However, care has to be taken to preserve interpretability. We applied the time-tolerant

![Figure 3: Simulated and observed values for the precision $P_\delta$ and recall $R_\delta$, for $\delta \in \{0, 1, 2, 4\}$ and various thresholds.](image)

![Figure 4: Cumulative distribution functions of the two types of true positives required for precision and recall.](image)

### 3.2 Null distributions

The simulations clearly show that assessment of the statistical significance of the observed performance measures is imperative. For this purpose, we fix the temporal tolerance to $\delta = 2$ and set the threshold $\tau$ to the .9-quantile of the anomaly score. We observe $PA_\delta - AA = 80$ (recall $R_\delta = .49$) and $PA - AA_\delta = 145$ (precision $P_\delta = .56$) on the non-permuted ground-truth anomaly sequence from our example. To assess the statistical significance of the reported numbers, we now have a closer look at the null distributions for the performance measures obtained in the Monte Carlo simulations.

Figure 4 (simulated) shows the cumulative distribution functions for the numbers of true positives obtained from 10,000 simulations for the specific choice of $\delta$ and $\tau$ mentioned above. Given the simulated distributions, we can easily compute Monte Carlo p-values [6] for the numbers of true positives. The p-value is the probability that we obtain a value for the true positive at least as high as the observed one. Since the performance measures were smaller than the reported values in all of the simulated runs, we have $p < .0001$ for both precision and recall, which is highly significant.

The analytical null distributions derived in the literature [7, 14] are all binomial. To complete our analysis, we now check whether our simulations also yield binomial distributions. Figure 4 (binomial) shows the cumulative distribution functions of binomial random variables when the binomial success probabilities are estimated from our Monte Carlo simulations. The plots suggest that the true positive $PA_\delta - AA$ for the recall follows a binomial distribution, whereas the true positive $PA - AA_\delta$ for the precision seems to be overdispersed with respect to the binomial distribution (it has a larger variance). We have repeated the experiment with different thresholds and temporal tolerances and observed the same behavior across all experiments. The exact form of the overdispersed distribution should be investigated more deeply in future work.

### 4 CONCLUSION

We have presented time-tolerant variants of the precision and recall measures routinely used to evaluate anomaly detectors for sequential data. We have shown that these measures are computed from two distinct time-tolerant confusion matrices. Time-tolerant confusion matrices can, in principle, be used to derive time-tolerant variants of other well-known measures. However, care has to be taken to preserve interpretability. We applied the time-tolerant
precision and recall measures on an example anomaly detection problem, and analyzed their statistical behaviors in a simulation study. Our experiments suggest that reported values for precision and recall can overestimate the performance of an anomaly detector even with moderate temporal tolerances. We have demonstrated how to obtain Monte Carlo \( p \)-values to assess the statistical significance of reported performance measures, using randomly permuted ground-truth sequences. We believe that establishing the statistical significance of reported precision and recall values should become a community standard. Future work should improve the analytical understanding of the null distributions required for this task.

REFERENCES

[1] Roy J. Adams and Benjamin M. Marlin. 2017. Learning time series detection models from temporally imprecise labels. In AISTATS.
[2] Roy J. Adams and Benjamin M. Marlin. 2018. Learning time series segmentation models from temporally imprecise labels. In UAI.
[3] Robert A Bridges, Jessie D Jamieson, and Joel W Reed. 2017. Setting the threshold for high throughput detectors. In IEEE Big Data.
[4] Lianhua Chi, Bo Han, and Yun Wang. 2016. Open Problem: Accurately Measuring Event Impacts on Time Series. In KDD MiLeTS Workshop.
[5] Stuart Coles. 2001. An Introduction to Statistical Modeling of Extreme Values. Springer-Verlag London, Ltd, London, UK.
[6] A. C. Davison and D. V. Hinkley. 1997. Bootstrap Methods and Their Application. Cambridge University Press, Cambridge, UK.
[7] J. F. Donges, C.-F. Schleusner, J. F. Siegmund, and R. V. Donner. 2016. Event coincidence analysis for quantifying statistical interrelationships between event time series: On the role of flood events as triggers of epidemic outbreaks. European Physics Journal Special Topics 487 (2016), 471–487.
[8] Paul S. Earle, Daniel C. Bowden, and Michelle Guy. 2011. Twitter earthquake detection: Earthquake monitoring in a social world. Annals of Geophysics 54, 6 (2011), 708–715.
[9] Chen Luo, Jian-Guang Lou, Qiongwei Lin, Qiang Fu, Rui Ding, Dongmei Zhang, and Zhe Wang. 2014. Correlating events with time series for incident diagnosis. In KDD.
[10] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. 2015. Long Short Term Memory Networks for Anomaly Detection in Time Series. In ESANN.
[11] David M. W. Powers. 2007. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report. Flinders University of South Australia, Adelaide, Australia.
[12] R. Quian Quiroga, T. Kreuz, and P. Grassberger. 2002. Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. Physical Review E 66, 4 (2002).
[13] Hanhong Ren, Bixiong Xu, Yijing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. 2019. Time-Series Anomaly Detection Service at Microsoft. In KDD.
[14] Erik Scharwächter and Emmanuel Müller. 2020. Does Terrorism Trigger Online Hate Speech? On the Association of Events and Time Series. Annals of Applied Statistics (2020).
[15] Erik Scharwächter and Emmanuel Müller. 2020. Two-Sample Testing for Event Impacts in Time Series. In Proceedings of the SIAM International Conference on Data Mining (SIAM SDM).
[16] Jonatan F. Siegmund, Marc Wiedermann, Jonathan F. Donges, and Reik V. Donner. 2016. Impact of temperature and precipitation extremes on the flowering dates of four German wildlife shrub species. Biogeosciences (2016).
[17] A. Siffer, P.-A. Fouque, A. Termier, and C. Lagoutet. 2017. Anomaly detection in streams with extreme value theory. In KDD.
[18] Ya Su, Rong Liu, Youqian Zhao, Wei Sun, Chenhao Niu, and Dan Pei. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In KDD.
[19] Nesime Tatbul, Tie Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. 2018. Precision and recall for time series. In NeurIPS.
[20] M. A.M.M. van Dortmont, S. van den Elzen, and J. J. van Wijk. 2019. ChronoCorrelator: Enriching events with time series. In EuroVis.
[21] Marc Wiedermann, Alexander Radebach, Jonathan F. Donges, Jürgen Kurths, and Reik V. Donner. 2016. A climate network-based index to discriminate different types of El Niño and La Niña. Geophysical Research Letters 43, 13 (2016), 7176–7185.
[22] Mitchell Withers, Richard Aster, Christopher Young, Judy Beiriger, Mark Harris, Susan Moore, and Julian Trujillo. 1998. A comparison of select trigger algorithms for automated global seismic phase and event detection. Bulletin of the Seismological Society of America 88, 1 (1998), 95–106.
[23] Haowen Xu, Yang Feng, Jie Chen, Zhaogang Wang, Honglin Qiao, Wenziao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zihan Li, Ying Liu, Youqian Zhao, and Dan Pei. 2018. Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications. In WWW.