Maximally Divergent Intervals for Anomaly Detection

Erik Rodner[1,3], **Björn Barz**[1], **Yanira Guanche**[1], **Milan Flach**[2], **Miguel Mahecha**[2,3], **Paul Bodesheim**[2], **Markus Reichstein**[2,3], **Joachim Denzler**[1,3] ERIK.RODNER@UNI-JENA.DE

[1]Computer Vision Group, Friedrich Schiller University of Jena, Germany, http://www.inf-cv.uni-jena.de

[2] Max Planck Institute for Biogeochemistry, Department Biogeochemical Integration, Jena, Germany

[3] Michael Stifel Center for Data-driven and Simulation Science, Jena, Germany

Abstract

We present new methods for batch anomaly detection in multivariate time series. Our methods are based on maximizing the Kullback-Leibler divergence between the data distribution within and outside an interval of the time series. An empirical analysis shows the benefits of our algorithms compared to methods that treat each time step independently from each other without optimizing with respect to all possible intervals.

1. Introduction

Scientific data increases with respect to both volume and dimensionality. Manually analyzing large-scale multivariate data is therefore intractable and automatic methods are needed to structure data and point researchers to the most interesting parts of scientific measurements. We focus on detecting anomalies in time series, which is an essential task, *e.g.*, in climate and ecosystem studies (Zscheischler et al., 2014), oceanic research (Mínguez et al., 2012), or in industrial processes (Darkow et al., 2014).

The survey article of (Chandola et al., 2009) categorizes anomaly event detection methods into six main groups. *Classification-based* methods utilize common classification techniques, such as neural networks, Bayesian networks (Dieh et al., 2002), or Support Vector Machines (Ma & Perkins, 2003) and learn their models by sliding windows strategies. Anomalies can also be detected by considering the distance to the k^{th} *nearest neighbor* (Byers & Raftery, 1998; Bodesheim et al., 2015) or the relative density (Chiu & Fu, 2003). *Clustering* techniques (Smith et al., 2002) group similar data into clusters leading to anomalies being far from the cluster centroids (Smith et al., 2002). An intuitive strategy is based on *statistical modeling*, where anomalies are assumed to be points that do not fit to a previously estimated statistical model (Anscombe, 1960; Mínguez et al., 2012). Other approaches are based on *information theory* (Ando, 2007; Bodesheim et al., 2012) or *spectral analysis* (Shyu et al., 2003), where subspaces of the normal data are detected.

All of these methods determine an anomaly score for each point in the time series individually. In contrast, we propose a method that directly considers the detection of contiguous intervals, an important property for scientific data analysis. Time intervals whose distribution is considerably different from the rest of the time series can be considered as anomalies. This is done by maximizing a divergence criterion between the distributions. Depending on the assumptions on these distributions, we derive different methods that allow for batch detection of anomalies.

The most related paper to ours is the method proposed by (Liu et al., 2013), which also uses a divergence criterion to detect changes in the data. However, their method makes use of the more general f-divergence and directly estimates the ratio of the two densities. Since we are optimizing over all possible intervals instead of only a window of fixed size, we rely on efficient update formulas, which are not available for the methods proposed in (Liu et al., 2013). The paper of (Görnitz et al., 2015) combines a hidden Markov model with a latent one-class SVM for detecting time series containing an anomaly. Their method requires some kind of supervision to learn a state model and also does not directly focus on anomaly localization in contrast to our approach.

In the following sections, we give a brief description of the problem, the general framework we propose for batch anomaly detection as well as specific algorithms. Subsequent experiments show the properties and benefits of the algorithms especially with respect to single point anomaly detection. Furthermore, we propose average precision and an intersection-over-union criterion as a suitable evaluation methodology for anomaly detection in time series.

Presented at ICML2016 Anomaly Detection Workshop, New York, NY, USA, 2016.

2. Maximally divergent intervals (MDI)

Definitions and problem description The basic idea behind our approach is that an anomaly interval within a data distribution is significantly different from the rest of the time series. Therefore our main objective is to be able to find intervals in a time series $(\boldsymbol{x}_t)_{t=1}^n$ with $\boldsymbol{x}_t \in \mathbb{R}^D$ being a multivariate observation at time t.

Let $I = \{t \mid t_1 \leq t < t_2\}$ be an interval, where data points are assumed to be sampled from p_I. The remaining set of data points is denoted by $\Omega = \{1, \ldots, n\} \setminus I$ with the data distributed by p_Ω. In order to find those intervals: i) we need to define a parameterized model for the distributions p_I and p_Ω that can be estimated from the data, and ii) be able to calculate the "difference" between p_I and p_Ω.

For the latter, we propose to use the Kullback-Leibler (KL) divergence to measure the difference between distributions. Furthermore, we model the data distributions either by kernel density estimation (KDE) or multivariate Gaussian distributions. These two models allow us to compute the KL divergence in an efficient manner.

Maximizing the Kullback-Leibler divergence In the following, we assume the data points in either Ω and I to be sampled independently from each other. This is for sure a severe assumption that does not hold for relevant time series, however, we will demonstrate later on how the dependencies between subsequent data points can be handled with a simple pre-processing step. The Kullback-Leibler divergence of two distributions p_Ω and p_I is defined as:

$$\mathrm{KL}(p_I, p_\Omega) = \int p_I(\boldsymbol{x}) \log \frac{p_I(\boldsymbol{x})}{p_\Omega(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \ . \tag{1}$$

The KL divergence is zero for identical distributions and large for "significantly different" data distributions. We approximate it using an empirical expectation over the set of anomalous points leading to:

$$\mathrm{KL}_{I,\Omega} = \frac{1}{|I|} \sum_{t \in I} (\log p_I(\boldsymbol{x}_t) - \log p_\Omega(\boldsymbol{x}_t)) \tag{2}$$

This resulting criterion is very intuitive since it is calculating the differences of log-likelihoods for p_I and p_Ω. To find the interval belonging to an anomaly, we maximize the KL divergence with respect to the interval I:

$$\hat{I} = \mathrm{argmax}_{I \in \mathcal{I}} \ \mathrm{KL}_{I,\Omega} \ . \tag{3}$$

The set $\mathcal{I}$ contains suitable intervals and is important to integrate prior expectations about anomaly intervals, such as a range of possible interval sizes. Naive brute-force optimization of the empirical KL divergence requires $\mathcal{O}(|\mathcal{I}| \cdot T)$ operations, where T is the time needed to evaluate the KL divergence and $\mathcal{I}$ is usually $\mathcal{O}(n \cdot n')$ with n' being the maximum size of an anomaly interval. A property of the KL divergence is its asymmetry, $\mathrm{KL}_{I,\Omega} \neq \mathrm{KL}_{\Omega,I}$. Other work (Liu et al., 2013) often relied on a symmetric version of it. We use $\mathrm{KL}_{I,\Omega}$ for reasons theoretically explained later on and validated in our experiments. To obtain m anomalies, a non-maximum-suppression method (Neubeck & Van Gool, 2006) is used to select the m non-overlapping intervals with highest divergence.

MDI with kernel density estimation (MDI KDE) A very flexible way to model and estimate distributions is kernel density estimation (KDE). For a given kernel function K, the estimate for p_I is defined by:

$$p_I(\mathbf{x}) = \tfrac{1}{|I|} \textstyle\sum_{t_1 \leq t < t_2} K(\mathbf{x}, \mathbf{x}_t) \tag{4}$$

for an arbitrary multivariate observation $\mathbf{x}$. We use the same model for p_Ω. As a kernel function, we use the Gaussian kernel normalized such that p_I is a proper density.

Straightforward computation of the KL divergence for p_I and p_Ω estimated by kernel density estimation requires $O(n^2)$ operations (distance calculations). Together with our brute-force optimization, this yields an $\mathcal{O}(n^3 \cdot n')$ algorithm, which is only practical for small time series. However, the idea of cumulative sums is used to achieve a significant speed-up (Viola & Jones, 2004). Let $\boldsymbol{K} = (K_{t,t'}) \in \mathbb{R}^{n \times n}$ be the kernel matrix of the time series. We compute the cumulative sums in this symmetric matrix along a single axis: $C_{t,t'} = \sum_{t'' \leq t'} K_{t,t''}$ which can be pre-computed in $\mathcal{O}(n^2)$ time. Since we only need to compute our kernel density estimate for points of the time series, we can evaluate the estimates in constant time:

$$p_I(\boldsymbol{x}_t) = \frac{1}{|I|} \left(C_{t,t_2-1} - C_{t,t_1-1}\right) \ , \tag{5}$$

$$p_\Omega(\boldsymbol{x}_t) = \frac{1}{n - |I|} \left(C_{t,n} - C_{t,t_2-1} + C_{t,t_1-1}\right) \ . \tag{6}$$

After computing the kernel matrix and cumulative sums in $\mathcal{O}(n^2)$ time, we get an asymptotic time of $\mathcal{O}(n')$ for evaluating the KL divergence and a total time of $\mathcal{O}(\max(n^2, n'^2 \cdot n))$ for finding the maximally divergent interval.

MDI for normally-distributed data (MDI Gaussian) Another possibility to model the data distributions is a Gaussian model for p_I and p_Ω:

$$p_I(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_I, \boldsymbol{S}_I), \ p_\Omega(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_\Omega, \boldsymbol{S}_\Omega) \tag{7}$$

Estimating the mean vectors and covariance matrices can be also achieved with integral series. The exact KL divergence even has a closed form solution (Duchi, 2007):

$$\begin{aligned} \mathrm{KL}_{I,\Omega} = \frac{1}{2}\big(&\mathrm{trace}(\boldsymbol{S}_\Omega^{-1}\boldsymbol{S}_I) \\ &+ (\boldsymbol{\mu}_I - \boldsymbol{\mu}_\Omega)^T \boldsymbol{S}_\Omega^{-1} (\boldsymbol{\mu}_I - \boldsymbol{\mu}_\Omega) \\ &- D + \log(|\boldsymbol{S}_\Omega|) - \log(|\boldsymbol{S}_I|)\big) \ , \end{aligned} \tag{8}$$

which we can use for an evaluation of the divergence in a time independent of n, yielding a total computation time of $\mathcal{O}(n' \cdot n)$.

Depending on further assumptions on the distributions, interesting connections to related techniques can be derived. For example, when we assume a global shared covariance matrix ($\boldsymbol{S} = \boldsymbol{S}_I = \boldsymbol{S}_\Omega$), eq. (8) reduces to $\mathrm{KL}_{I,\Omega} = (\boldsymbol{\mu}_I - \boldsymbol{\mu}_\Omega)^T \boldsymbol{S}^{-1} (\boldsymbol{\mu}_I - \boldsymbol{\mu}_\Omega)$ resembling a Mahalanobis distance also used in Hotelling's T^2 test (MacGregor & Kourti, 1995). Furthermore, we can assume identity matrices for the covariances, which reduces eq. (8) to the squared Euclidean distance between the means. The expression in eq. (8) also justifies our choice of $\mathrm{KL}_{I,\Omega}$ instead of $\mathrm{KL}_{\Omega,I}$ or a symmetric version (Liu et al., 2013). For a univariate time series, $\mathrm{KL}_{\Omega,I}$ is given by:

$$\frac{1}{2}\left(\frac{S_\Omega}{S_I} + \frac{(\mu_I - \mu_\Omega)^2}{S_I} - 1 + \log(S_I) - \log(S_\Omega)\right),$$

and we can see that for small values of S_I, we get high values of the KL divergence. Therefore, a maximization of $KL_{\Omega,I}$ would prefer intervals of low variance. In contrast, $KL_{I,\Omega}$ is not affected by this phenomenon since S_Ω is estimated from a large portion of the time series.

Temporal context with time-delay embedding A major drawback of our algorithm so far is the assumption of independent data points in the time series, which is only valid for trivial academic cases. However, modeling the dependency can be done with a simple transformation of the time series. In particular, we use a multivariate time-delay embedding (Smets et al., 2009; Kantz & Schreiber, 2004), where the data points of the new time series are the concatenation of the last k time steps of the original time series, *i.e.* $\boldsymbol{x}'_t = (\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-k+1})$. Whereas this embedding leads to a smoothing of the distance matrix for our MDI KDE approach, it allows our MDI Gaussian approach for calculating and exploiting correlations between subsequent data points. For example, a change of frequency in the time series, can only be detected by our methods with a proper embedding, such as time-delay.

Multiple methods for estimating an "optimal" value for k have been proposed in the literature for univariate time series (Hegger et al., 1999). We developed a method that allows for optimizing k even for multivariate time series. However, the preliminary results we obtained when combining this hyperparameter tuning method with our MDI approach are beyond the scope of this paper. In our current experiments, we therefore fix $k = 3$ for all methods.

3. Experiments

Evaluation criteria, baselines, and implementation details Previous methods in the area of anomaly detection, typically return a novelty score for each of the data points in the time series. Therefore, a common choice for evaluation have been ROC curves and the area under these curves (AUC). However, our method returns scored time intervals directly and therefore differs from previous algorithms. We argue that in this case also a proper detection evaluation criterion needs to be used, which also better reflects the expectations of researchers about an algorithm's performance. Therefore, we count detected intervals as correct if they have an intersection over union ratio of more than $\beta = 0.5$ with a ground-truth anomaly interval. Evaluation is then done using recall-precision curve and the average precision (AP) metric. Figure 1 shows an example, where two algorithms have the same AUC but a significantly different AP performance.

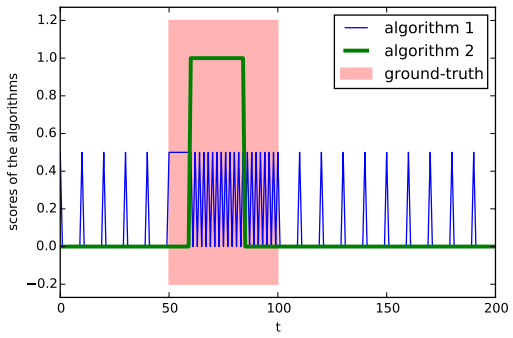

Figure 1. Resulting scores of two algorithms (blue and green) achieving the same AUC performance (0.75) with respect to the red ground-truth interval but a significantly different AP detection performance (0.0 for blue vs. 1.0 for green).

We compare our method to the following baselines: Hotelling's T^2 method (MacGregor & Kourti, 1995) and kernel density estimation operating on single data points in the time series and learned with all of the points. To allow for AP computation and a fair comparison, we use multiple thresholds on the scores to group single point detections into intervals. All of the obtained intervals are then filtered with non-maximum-suppression (NMS) (Neubeck & Van Gool, 2006) and receive as a score the minimum value of the single-point detection score estimated by the baseline.

All of the methods evaluated use a time-delay preprocessing of $k = 3$ and a subsequent NMS to obtain the 5 best scored non-overlapping intervals. Our MDI methods optimize over all possible intervals with sizes from 10 to 50.

Synthetic dataset We first test our algorithms on a synthetic dataset comprised of functions sampled from a Gaussian process prior (Gaussian kernel, $\sigma = 1$). The time series are perturbed at randomly sampled intervals with sizes ranging from 5% to 20% of the length of the whole time

Table 1. Results of our synthetic experiment for several baselines and several methods derived from our maximum divergent interval (MDI) framework. We use average precision (AP) and area under the ROC curve (AUC) as performance measures.

Method/AP	**MS**	**MSH**	**AC**	**FC**	**MS^5**	**FC^5**	**AC^5**
Hotelling's T^2 test (pointwise)	0.88	0.07	0.12	0.18	0.10	0.16	0.06
KDE (pointwise)	0.90	0.10	0.13	0.00	0.18	0.04	0.29
Ours, MDI KDE	0.97	0.12	0.20	0.00	0.82	0.00	0.43
Ours, MDI Gaussian (full cov.)	**1.00**	**0.44**	**0.79**	**1.00**	**1.00**	**0.82**	**0.62**
Ours, MDI Gaussian (no cov.)	0.84	0.14	0.02	0.00	0.45	0.01	0.19
Ours, MDI Gaussian (shared cov.)	0.32	0.06	0.01	0.01	0.10	0.04	0.04
Method/AUC	**MS**	**MSH**	**AC**	**FC**	**MS^5**	**FC^5**	**AC^5**
Hotelling's T^2 test (pointwise)	0.98	0.58	0.92	0.94	0.72	0.82	0.80
KDE (pointwise)	0.98	0.56	0.77	0.72	0.82	0.61	0.82
Ours, MDI KDE	0.95	0.52	0.65	0.39	0.90	0.34	0.71
Ours, MDI Gaussian (full cov.)	**1.00**	**0.76**	**0.94**	**0.97**	**0.99**	**0.90**	**0.87**
Ours, MDI Gaussian (no cov.)	0.90	0.54	0.69	0.46	0.87	0.45	0.75
Ours, MDI Gaussian (shared cov.)	0.90	0.62	0.80	0.58	0.79	0.65	0.71

series, which is set to 250. We simulate the following types of anomalies: (1) mean shift (MS): we set $x'_t = x_t - \mu$ with $\mu \in [3, 4]$ within the anomaly region, (2) mean shift hard (MSH): MS with $\mu \in [0.5, 1]$, (3) amplitude change (AC): multiplying one dimension of the data points with $1 + g(t)$ with g being a Gaussian window centered in the interval and with 2σ matching the interval length, (4) frequency change (FC): one dimension of the time series is sampled from a non-stationary Gaussian process prior (Paciorek & Schervish, 2004) with a change of the kernel hyperparameter within the anomaly interval. All of these types have 20 univariate (MS, MSH, AC, FC) and 20 multivariate ($D = 5$) instances (MS^5, AC^5, FC^5) in the dataset. Note that the anomaly interval can only be detected in one dimension of the multivariate instances and our algorithms do not have information about the dimension. We will release the code for generating the dataset and code for our algorithms to ensure reproducibility.

Results of our synthetic experiments The results of our synthetic experiments are given in Tab. 1 for AP and AUC performance. As can be seen from the AP results, the best method is MDI Gaussian with a full model for the covariance matrices. MDI KDE is not able to deal with frequency changes, since the correlations between the dimensions of subsequent data points are not taken into account. This also holds for the "no cov." and "shared cov" versions of MDI Gaussian. MDI Gaussian "full cov." is also clearly the best method with respect to AUC performance, however, there are a lot of cases where the AUC performance value of another method would not reveal the nearly random detection performance measured by AP.

Application of MDI to real datasets Meteocean data (significant wave height, H_s, wind speed, W and sea level pressure SLP) in a location near the Bahamas in the Atlantic Sea (23.838 N, 68.333 W) were used in these tests. Six months of hourly data, from June 2012 until November 2012 were extracted from the National Data Buoy Center

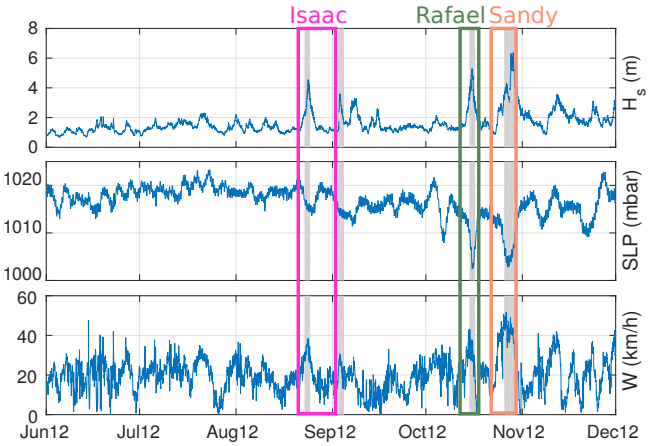

Figure 2. Boxes in colors represent historical hurricanes *Isaac*, *Rafael* and *Sandy* and grey shaded areas MDI Gaussian detections. The false-positive right after Isaac might be related either to a local storm or to the reminiscences from hurricane *Leslie* passing these days by Bermudas.

from the NOAA[1]. This period corresponds to the Atlantic hurricane season, which in that year was specially active with 19 tropical cyclones (winds above 52 km/h) were 10 of them became hurricanes (winds above 64 km/h). In contrast to our synthetic dataset, the anomalies have an effect on multiple variables at once.

We have applied the MDI Gaussian method to these three variables and compared the results with the historical hurricanes at Bahamas (Figure 2). The boxes in color represent the official duration of the three main events of that season that passed near our location, hurricanes *Isaac*, *Rafael* and *Sandy* respectively. Grey shaded areas represent the MDI intervals detected by the model. Note that in general the ground-truth areas are larger than the detections, because they span the entire lifetime of the hurricane and not just its presence at the Bahamas.

4. Discussion and conclusions

We presented methods to detect anomalies in time series. All of our methods maximize a KL divergence criterion that allows for finding intervals in time series that significantly differ from the rest with respect to their data distribution. We propose several variants for modeling the data distribution (kernel density estimation and different Gaussian assumptions) and analyze their particular benefits and drawbacks in experiments. In summary, our methods allow for efficient batch detection of anomalies in multivariate time series and are a useful tool for data discovery in the natural sciences. Future work will be focused on automatically inferring the number of anomalous intervals.

[1]http://www.ndbc.noaa.gov/

Acknowledgements The support of the project EU H2020-EO-2014 project BACI 'Detecting changes in essential ecosystem and biodiversity properties-towards a Biosphere Atmosphere Change Index, contract 640176 is gratefully acknowledged.

References

Ando, Shin. Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection. In *ICDM*, pp. 13–22. IEEE, 2007.

Anscombe, Frank J. Rejection of outliers. *Technometrics*, 2(2):123–146, 1960.

Bodesheim, Paul, Rodner, Erik, Freytag, Alexander, and Denzler, Joachim. Divergence-based one-class classification using gaussian processes. In *BMVC*, pp. 1–11, 2012.

Bodesheim, Paul, Freytag, Alexander, Rodner, Erik, and Denzler, Joachim. Local novelty detection in multi-class recognition problems. In *WACV*, pp. 813–820, 2015.

Byers, Simon and Raftery, Adrian E. Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association*, 93(442):577–584, 1998.

Chandola, Varun, Banerjee, Arindam, and Kumar, Vipin. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

Chiu, Anny Lai-mei and Fu, Ada Wai-chee. Enhancements on local outlier detection. In *Database Engineering and Applications Symposium*, pp. 298–307. IEEE, 2003.

Darkow, Thomas, Dittma, Rainer, and Timm, Helge. Real-time application of multivariate statistical methods for early event detection in an industrial slurry stripper. In *International Federation of Automatic Control*, pp. 8879–8884, 2014.

Dieh, Christopher P, Hampshire, John B, et al. Real-time object classification and novelty detection for collaborative video surveillance. In *IJCNN*, volume 3, pp. 2620–2625. IEEE, 2002.

Duchi, John. Derivations for linear algebra and optimization. Technical report, Berkeley, California, 2007.

Görnitz, N., Braun, M., and Kloft, M. Hidden markov anomaly detection. In *ICML*, pp. 1833–1842, 2015.

Hegger, Rainer, Kantz, Holger, and Schreiber, Thomas. Practical implementation of nonlinear time series methods: The tisean package. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 9(2):413–435, 1999.

Kantz, Holger and Schreiber, Thomas. *Nonlinear time series analysis*, volume 7. Cambridge university press, 2004.

Liu, Song, Yamada, Makoto, Collier, Nigel, and Sugiyama, Masashi. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43: 72–83, 2013.

Ma, Junshui and Perkins, Simon. Time-series novelty detection using one-class support vector machines. In *IJCNN*, volume 3, pp. 1741–1745. IEEE, 2003.

MacGregor, John F and Kourti, Theodora. Statistical process control of multivariate processes. *Control Engineering Practice*, 3(3):403–414, 1995.

Mínguez, R, Reguero, BG, Luceño, A, and Méndez, FJ. Regression models for outlier identification (hurricanes and typhoons) in wave hindcast databases. *Journal of Atmospheric and Oceanic Technology*, 29(2):267–285, 2012.

Neubeck, Alexander and Van Gool, Luc. Efficient non-maximum suppression. In *ICPR*, volume 3, pp. 850–855. IEEE, 2006.

Paciorek, C and Schervish, M. Nonstationary covariance functions for gaussian process regression. *NIPS*, 16: 273–280, 2004.

Shyu, Mei-Ling, Chen, Shu-Ching, Sarinnapakorn, Kanoksri, and Chang, LiWu. A novel anomaly detection scheme based on principal component classifier. *ICDM*, pp. 353–365, 2003.

Smets, Koen, Verdonk, Brigitte, and Jordaan, Elsa M. Discovering novelty in spatio/temporal data using one-class support vector machines. In *IJCNN*, pp. 2956–2963. IEEE, 2009.

Smith, Rasheda, Bivens, Alan, Embrechts, Mark, Palagiri, Chandrika, and Szymanski, Boleslaw. Clustering approaches for anomaly based intrusion detection. *Intelligent engineering systems through artificial neural networks*, pp. 579–584, 2002.

Viola, Paul and Jones, Michael J. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

Zscheischler, Jakob, Reichstein, Markus, Harmeling, S, Rammig, A, Tomelleri, E, and Mahecha, Miguel D. Extreme events in gross primary production: a characterization across continents. *Biogeosciences*, 11(11):2909–2924, 2014.