AIP Conference Proceedings

RESEARCH ARTICLE | OCTOBER 06 2023

Anomaly detection for parabolic trough power plants with density-based outlierness FREE

Josua Braun ✉; Alex Brenner; Gabriele Gühring

Check for updates

AIP Conf. Proc. 2815, 090001 (2023)
https://doi.org/10.1063/5.0149735

View Online Export Citation CrossMark

20 November 2023 12:23:24

AIP Advances
Why Publish With Us?
25 DAYS average time to 1st decision
740+ DOWNLOADS average per article
INCLUSIVE scope
Learn More
AIP Publishing

AIP Publishing

Anomaly Detection for Parabolic Trough Power Plants With Density-based Outlierness

Josua Braun[1, a)], Alex Brenner[1, b)] and Gabriele Gühring[2, c)]

[1]*German Aerospace Center (DLR), Institute of Solar Research, Wankelstraße 5, 70563 Stuttgart, Germany*
[2]*Hochschule Esslingen, Kanalstr. 33, 73728 Esslingen am Neckar, Germany*

[a)] Corresponding author: josua.braun@dlr.de
[b)]alex.brenner@dlr.de
[c)]gabriele.guehring@hs-esslingen.de

Abstract. Defects and faults in parabolic trough power plants often lead to lower energy production. The automated detection of such anomalies could reduce downtimes and increase efficiency. A machine learning anomaly detection approach which is based on processing data recorded during regular plant operation is examined. The measured data are high-dimensional data in a spatio-temporal context. Existing data is preprocessed and useful features are selected and extracted. These features build the input for the anomaly detection. The basic idea of the method is the consideration of multivariant time series on loop level, which are segmented and further quantified in its outlierness. For the segmentation of the time series two different segmentation methods, a window-sliding and a periodic method, are used and compared. For the quantification of the outlierness the local outlier factor is used in both cases. The results of both variants of anomaly detection show differences in the time points of detection, but globally they show a similar behavior and detection.

20 November 2023 12:23:24

INTRODUCTION

In a huge parabolic trough power plant (PTPP) solar field, a variety of failures and defects can occur. Monitoring the enormous amount of measured sensor data manually is nearly impossible and would probably lead to long delays in the detection of failures. An automated anomaly detection and therefore an immediate detection and correction of failures, provides several advantages. First of all, it leads to shorter downtimes of the system and a proper functionality of all components results in a general higher efficiency, availability and reliability of the power plant. Secondly, it prevents from further damage because a failure which is not repaired for a long time can cause other failures. Fewer consequential damages can lead to reduced maintenance work and cost, as well as a longer operational lifetime. In summary, one can assume that detecting failures decreases the levelized cost of electricity (LCOE) for PTPPs.

Anomaly detection with methods of machine learning (ML) is used in a lot of other applications for detecting failures of systems. For example, in [1] methods of ML are used for an anomaly detection of wind turbines. The ML methods are used on sensor data which is measured for the operation of power plants anyway. The approach primarily needs software to be implemented. Besides that, it only requires hardware with suitable computational power. The calculation on a central processing unit (CPU) is sufficient. In comparison to other approaches, which use new sensors or devices, acquisition costs are possibly lower.

The aim of this paper is to present and evaluate a ML-based approach for the anomaly detection on the spatio-temporal raster data of PTPPs. The approach includes a segmentation of multivariate time series followed by a transformation into the feature space for further quantification of the outlierness with the help of the local outlier factor (LOF) from M. M. Breunig et al. [2]. In comparison to other data-based approaches the approach has several advantages. The model does not need knowledge about the physical processes or detailed parameters of the observed

SolarPACES: Solar Power & Chemical Energy Systems
AIP Conf. Proc. 2815, 090001-1–090001-8; https://doi.org/10.1063/5.0149735
Published by AIP Publishing. 978-0-7354-4623-6/$30.00

PTPP as it is required in physical model approaches. Furthermore, it uses well known and understandable machine learning methods that often need less computational power than methods of deep learning.

RELATED WORK

In [3] a research about the failure detection of a PTPP in Chile is presented. The research is based on simulated data, whereas we use real measured data. They train four different artificial neural networks for the detection in the solar field, the heat exchanger, the pump and the complete system. We focus on the anomaly detection of the solar field which is the crucial difference to other power plants. Furthermore, we use lightweight ML methods for the anomaly detection with low run times.

METHODS

The anomaly detection is done in three steps: segmentation, feature extraction and quantification of outlierness (see **FIGURE 1**). The segmentation divides the multivariate time series into elementary segments. The feature extraction transfers the time series segments into the feature space where a quantification of the outlierness takes place.

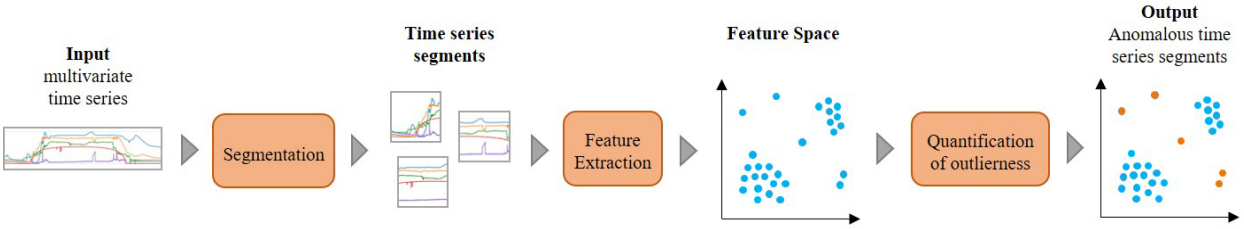


FIGURE 1. Schematic chain of methods for anomaly detection.

Dataset and Data Preparation

The analyzed dataset is operational data of one reference PTPP over one entire year. At first, a subspace of relevant signals for further preprocessing and feature extraction is selected. The selected signals are linearly interpolated, so that the timestamps of all signals are synchronized. Furthermore, the noise and the number of points of the signals are reduced with a piecewise aggregate approximation (PAA) as proposed by Lin et al. [4]. Thereby, the periodic duration is increased to 30 s.

The input data correspond respectively to one loop and contain the following five features: inlet temperature, outlet temperature, volume flow, Solar Collector Assembly (SCA) angle deviation and Direct Normal Irradiance (DNI). Inlet temperature and volume flow are estimated from the subfield measurements. The SCA angle deviation is the mean value of the absolute difference between the measured SCA angle and the calculated track angle of the respective SCAs of the loop. Furthermore, the SCA angle deviation is limited to the value range [0, 3.5] to minimize the distribution of the data. Values greater than 3.5° are set to 3.5°. These values would either way lead to non-focused collectors and a focusing factor of zero. The DNI is spatially interpolated from the three nearest weather stations depending on the position of the loop.

The input time series are trimmed so that only daytime values are considered. This reduces the amount of data and the general value range because the values of the night hours are eliminated. The track angle for the trimming condition is used which allows us to get a time period from shortly after sunrise to shortly before sunset. Hence, the ramp-up and ramp-down of the power plant at sunny days is trimmed.

Segmentation of Input Data

The segmentation splits the time series of long duration in smaller homogeneous pieces, called segments. The data in each segment can be described accurately by a simple model. It reduces the complexity of the representation of the original data. [5]

20 November 2023 12:23:24

It is important that on the one hand the duration of segments is not too short. A small segment might not include the necessary context, leading to many close points in the feature space and to a longer runtime of the subsequent methods. One the other hand the duration of segments should not be too long. In too long segments a relatively short anomalous event might be overseen, because the rest of the data is dominant. Even if an anomalous event is not overseen, it is harder to make an estimation of the exact time point of the anomalous event.

The segmentation has a significant impact on the number and position of points in the feature space. Hence, the anomaly detection is done with two different methods of segmentation: window-sliding segmentation (variant 1) and periodic segmentation (variant 2).

Window-Sliding Segmentation

The window-sliding algorithm as described in [6] is a fast and relatively simple method for change point detection. It is an approximate search method, that yields towards an approximate solution of the change points.

The algorithm computes the discrepancy d between two adjacent windows. The left window $y_{a,t}$ is defined from time point a to t, the right window $y_{t,b}$ from time point t to b (see **FIGURE 2**). The window pair slides along a signal y of length T. For the two adjacent windows and the joined window $y_{a,b}$ are calculated costs for a given cost function $c(y)$. The discrepancy between two windows as described in [6] is given by:

$$d(y_{a,t}, y_{t,b}) = c(y_{a,b}) - c(y_{a,t}) - c(y_{t,b}) \quad (1 \leq a < t < b \leq T) \tag{1}$$

The L2 cost function is used, which is defined by:

$$c(y_{a,b}) = \sum_{d=1}^{D} \sum_{i=a}^{b} (y_i^d - \bar{y}_{a,b}^d)^2 \tag{2}$$

whereby, D is the number of dimensions of the time series and $\bar{y}_{a,b}^d$ denotes the mean value of the segment ranging from a to b in dimension d.

The discrepancy in equation (1) reaches large values, when the segments of the two windows are dissimilar. It is calculated for each time point of the time series and results in a discrepancy curve (see **FIGURE 2**). Peaks in the discrepancy curve are identified as change points. [6]

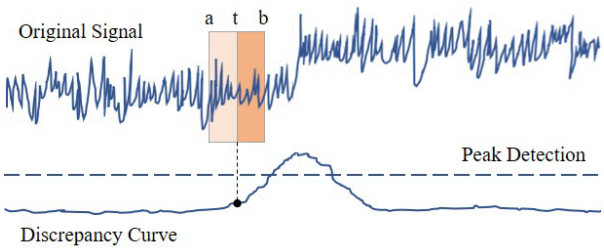


FIGURE 2. Schematic view of window-sliding algorithm.

The window-sliding algorithm helps finding the breakpoints for the segmentation. Therefore, the time points of the discrepancy peaks are sorted by its discrepancy. The time points, considered as breakpoints, are successively added to a list of breakpoints $b = b_1, \ldots, b_N$ beginning with the breakpoint b_1 with the highest discrepancy. For each iteration a reconstruction error for the current segmentation is calculated by:

20 November 2023 12:23:24

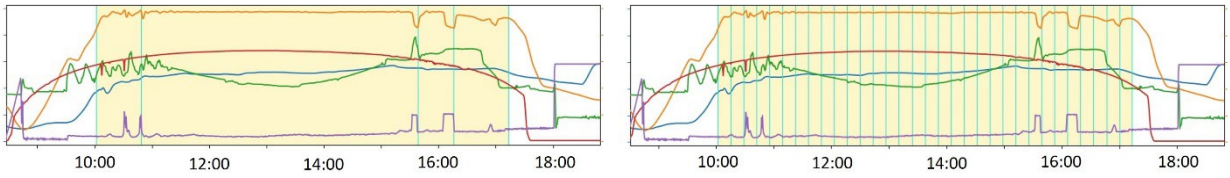
$$error(b) = N \cdot \sum_{i=b_1}^{b_N} Var(x_{i,i+1}) \tag{3}$$

where N is the current number of breakpoints and $Var(x_{i,i+1})$ the variance of the time series segment $x_{i,i+1}$ from i to $i+1$. If $error(b)$ becomes greater than a certain threshold ε the list of breakpoints b is returned and they build the borders for the segmentation. Due to the calculation of the reconstruction error in equation (3), the scale of the features has an impact on the error value. Hence, each feature is scaled with a z-score standardization before the segmentation is applied.

The resulting segmentation depends on the window length and the threshold ε. The window length is set to $n = 6$ so that the sliding window considers a duration of $3\,min$. The duration of $3\ min$ is a value where a significant change of the domain-specific signal can be seen in each dimension without considering a too long duration. The error threshold is set to $\varepsilon = 100$. The threshold value is determined by comparing the results of different threshold values. Lower values lead to smaller segments, higher values lead to longer segments. The determined threshold value results in an average segment duration of $15\,\text{min}\,42\,s$. This ensures a good balance between a huge number of segments and a reasonable split into homogenous segments. A segmentation example can be seen in **FIGURE 3 (a).**

Periodic Segmentation

Another method of segmentation, that is applied, is to divide the time series in regular pieces. It is the simplest way of segmentation. The duration of segments is defined by the number of points n_{seg} for each segment. The number of points is set to $n_{seg} = 27$, which corresponds to a segmentation period of $13\,\text{min}\,30\,s$. This segmentation leads to a similar number of segments as the window sliding segmentation. A segmentation example can be seen in **FIGURE 3 (b).**



(a) (b)

FIGURE 3. Example for trimming (yellow background) and segmentation (turquoise breakpoints) of time series sample. (a) With window-sliding segmentation. (b) With periodic segmentation

Feature Extraction

After the segmentation of the time series, the segments are transformed into a feature space. Therefore, a feature matrix is built by deriving features from each segment. In principle, for time series thousands of interpretable features can be derived. The derived features can be chosen in terms of correlation structure, distribution, entropy, stationarity and scaling properties [7]. The following two features are extracted:

- Mean values of each individual dimension
- Standard deviations of each individual dimension

To avoid the problem of high-dimensional data, only these two features are chosen for each dimension. Furthermore, the features mean and standard deviation are well understandable. Five dimensions and two features per dimension result in a total number of m = 10 extracted features. The distribution of the data in the feature space has a strong impact on the quantification of the outlierness. Therefore, each feature vector X_d is scaled with the z-score standardization

20 November 2023 12:23:24

$$Z_d = \frac{X_d - \mu_d}{\sigma_d} \tag{4}$$

whereby, μ_d is the mean value and σ_d the standard deviation of the feature X_d of dimension d with $d = 1, \dots, m$.

The horizontal concatenation of the standardized feature vectors Z_d build the feature matrix F. The feature matrix has the size $n \times m$, where n is the number of segments or points and m the number of extracted features.

Quantification of Outlierness

The outlierness of each point in the feature space is quantified with the LOF. The LOF identifies how isolated a point is from all other points. It measures the local density deviation of a point with respect to its k-nearest neighbors. Points with a significant lower density than their neighbors can be considered as outliers. [8]

At first, for a given point $\bar{X}$ it is calculated the reachability distance $R_k(\bar{X}, \bar{Y})$ to each neighbor by:

$$R_k(\bar{X}, \bar{Y}) = \max\{dist(\bar{X}, \bar{Y}), D^k(\bar{Y})\} \tag{5}$$

whereby, $D^k(\bar{Y})$ is the distance of the k-nearest neighbor of the point $\bar{Y}$. Then, the average reachability distance $AR_k(\bar{X}, \bar{Y})$ is defined as the average of the reachability distances of all points in the k-neighborhood $L_k(\bar{X})$ of point $\bar{X}$. [8]

$$AR_k(\bar{X}) = MEAN_{\bar{Y} \in L_k(\bar{X})} R_k(\bar{X}, \bar{Y}) \tag{6}$$

The LOF is then simply equal to the mean ratio of $AR_k(\bar{X})$ to the corresponding values of all points in the k-neighborhood of $\bar{X}$. [8]

$$LOF_k(\bar{X}) = MEAN_{\bar{Y} \in L_k(\bar{X})} \frac{AR_k(\bar{X})}{AR_k(\bar{Y})} \tag{7}$$

If a point is in a homogeneously distributed area, where the neighbors show a similar density, the average reachability distance of the point $\bar{X}$ and that of its neighbors is approximately equal. The LOF becomes nearly one and can be considered as inlier. If a point is an outlier, the average reachability distance of the point is greater than the average reachability distance of its neighbors. In that case, the LOF value becomes greater than one. [8]

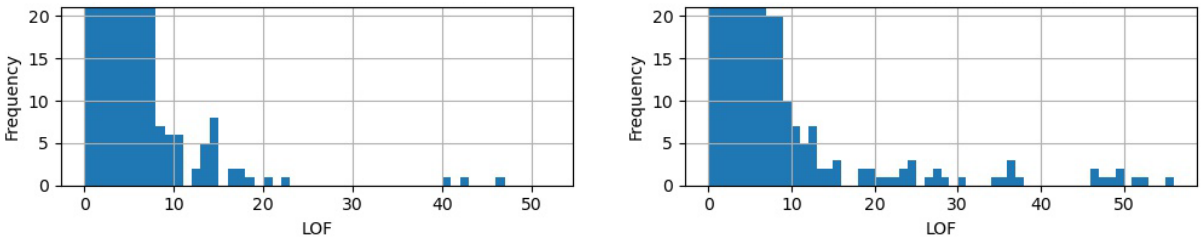
The preceding research [9] used the DBSCAN clustering instead of calculating the LOF. Because of a higher runtime complexity DBSCAN was excluded. Moreover, the LOF gives an outlier score whereas the DBSCAN only gives a binary outlier label.

RESULTS

Both variants of segmentation result in an approximately similar number of segments. The window-sliding segmentation results in 2,022,440 segments. The periodic segmentation results in 2,387,160 segments.

Both variants show a few numbers of segments with a very high LOF as shown in **FIGURE 4**. The variant with the window-sliding segmentation shows 44 segments with a LOF greater than 8. The variant with the periodic segmentation shows 88 segments with a LOF greater than 8.

20 November 2023 12:23:24



(a) (b)

FIGURE 4. Histogram of LOFs. (a) With window-sliding segmentation. (b) With periodic segmentation.

To show the similarity or dissimilarity of the results between both variants the rate of overlapping anomalies depending on the number of assumed anomalies is calculated (see **FIGURE 5**). Therefore, the given number of segments with the highest LOFs are taken as anomalies. The overlapping rate p_{lap} is the number of overlaps divided by the number of assumed anomalies n_a. An overlap is counted, if an anomalous segment of variant 1 shares at least one time point with an anomalous segment of variant 2.

$$p_{lap} = \frac{n_{lap}}{n_a} \tag{8}$$

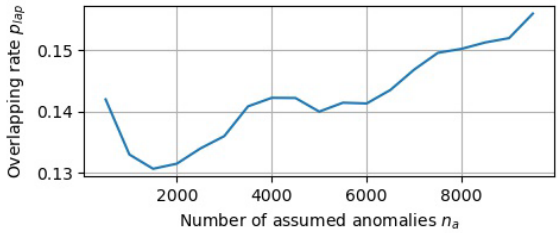


FIGURE 5. Course of overlapping rate depending on number of assumed anomalies.

Generally, the overlapping rate becomes higher for an increasing number of assumed anomalies. For a number of less than 8000 assumed anomalies the overlapping rate is smaller than 15 %. The low overlapping rate indicates that the distribution of the segments in the feature space is highly depending on the segmentation and lead to very different LOFs and detected anomalies. However, with a look at the mean of the features of the top number of outlier points (ranked by the LOF), very similar feature deviations for the anomalies of both variants can be seen. **FIGURE 6** shows the deviation of the mean of the features over the top 500 outlier points from the mean of the features over all points. For example, both variants show a high mean deviation in the standard deviation of the loop inlet temperature (std, inlet_temp). This indicates that many outlier points are caused by a significant change of the loop inlet temperature (see e.g. **FIGURE 8**). The reason for the detection of the anomalies are very similar in both variants. In connection with the low overlapping rate it points that both variants find similar anomalies but at different time points or in different loops.

20 November 2023 12:23:24

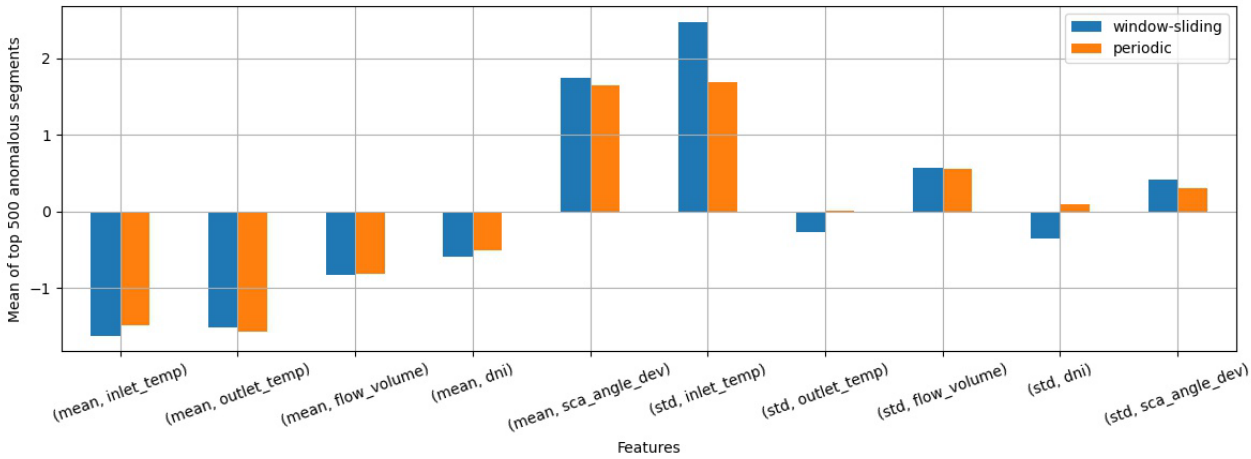
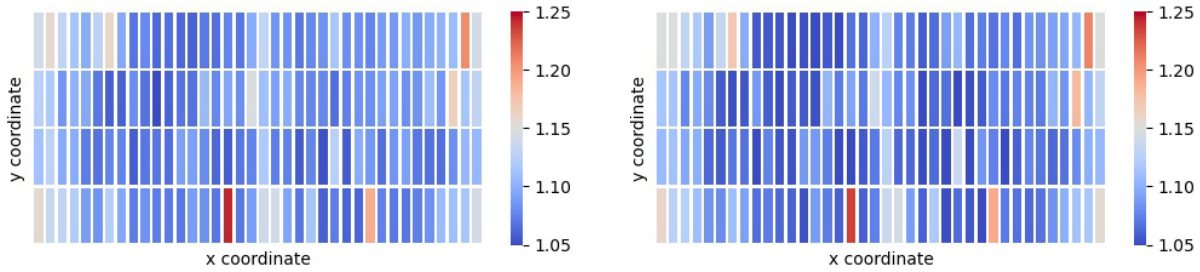


FIGURE 6. Mean deviation of the features of the top 500 anomalies.

Furthermore, the susceptibility to anomalies for specific loops is analyzed. **FIGURE 7** shows a spatial map of the mean LOF values of each loop in the PTPP. The mean is built over all timestamps of the trimmed time series of the entire dataset. The red marked loops have an increased mean LOF. In these loops, anomalies occur more frequently in the given time period. Both variants show very similar mean LOF values. This indicates that the spatial detection of anomalies is less vulnerable to different segmentations.



(a) (b)

FIGURE 7. Spatial map of mean LOFs for each loop. (a) Variant 1 with window-sliding segmentation. (b) Variant 2 with periodic segmentation.

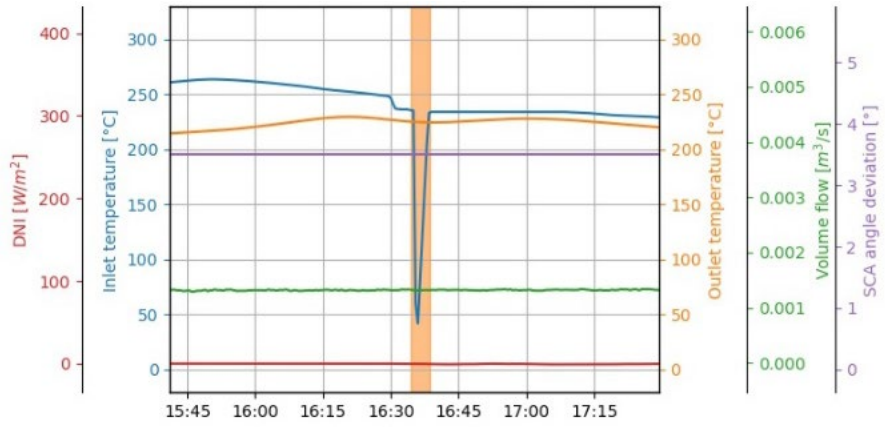


FIGURE 8. Example segment with high LOF in both variants of segmentation. Most likely, the reason for the detection is the significant decrease of the inlet temperature.

20 November 2023 12:23:24

The analysis of the mean deviation of the features of the anomalies, as well as the mean spatial LOFs show that the global detection of anomalies is very similar for both segmentation approaches. The analysis of the overlapping rate shows that there is a low number of shared anomalous segments. Both effects together indicate that both variants find similar anomalies in the same loops but at different time points. Therefore, both variants have their reason for existence: One variant detects anomalies that the other variant does not detect and vice versa. Without having a ground truth, it is hard to decide which variant gives better results. Hence, it is recommended to fuse the results of different variants. This can be done for example by building the mean of the LOF values between the variants. The variation can be implemented via segmentation, feature extraction, feature selection or even other approaches.

DISCUSSION AND CONCLUSION

The examined approach also has several disadvantages. At first, the temporal and spatial context of the segments to each other gets lost in the feature space. Observing one segment in the feature space, only the distances to all other segments are seen. The behavior of previous or following segments or of segments of a nearby loop is not considered for the quantification of outlierness of one segment. The connection between temporally and spatially contiguous sensor data could play a major role in anomaly detection. Especially the large time delay between the inlet and outlet temperature due to the lengthy heating process in the receiver requires an approach that can interpret the correlation between these two features. Furthermore, every loop outlet temperature of one subfield effects the volume flow due to the control of the loop outlet temperature. If the model sees only the data of one loop it cannot know whether a decrease of the volume flow is reasonable or not.

Despite the disadvantages, anomalies which deviate in their features could be detected with a comparative simple approach. With the use of the understandable features mean and standard deviation, irregularities in the highly variable data and therefore possible failures of the PTPP can be detected. Moreover, the approach is capable to identify loops which are prone to errors by building the mean spatial LOFs. Furthermore, results of different variants of the approach can be combined to achieve results which are even closer to the ground truth.

The shown approach is good in detecting failures that occur fast so that the data already changes in a few minutes. Future research could investigate approaches that detect slowly creeping failures which changes the data over hours or days.

REFERENCES

1. Cui, Y., P. Bangalore, and L.B. Tjernberg. *An Anomaly Detection Approach Based on Machine Learning and SCADA Data for Condition Monitoring of Wind Turbines*. in *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. 2018.
2. Breunig, M.M., et al., *LOF: Identifying Density-Based Local Outliers*, in *2000 ACM SIGMOD Int. Conf. On Management of Data*. 2000: Dalles, TX.
3. Muñoz, I., et al., *Development of a Failure Detection Tool Using Machine Learning Techniques for a Large Aperture Concentrating Collector at an Industrial Application in Chile*, in *SolarPACES 2018*. 2018, AIP Conf. Proc, 170008, https://aip.scitation.org/doi/abs/10.1063/1.5117678.
4. Lin, J., et al., *Experiencing SAX: a novel symbolic representation of time series.* Data Mining and Knowledge Discovery, 2007. **15**(2): p. 107-144.
5. Bingham, E., et al., *Segmentation and dimensionality reduction*, in *SIAM International Conference on Data Mining*. 2006, Society for Industrial and Applied Mathematics (SIAM).
6. Truong, C., L. Oudre, and N. Vayatis, *Selective review of offline change point detection methods.* Signal Processing, 2020. **167**: p. 107-299.
7. Fulcher, B. and N. Jones, *Highly Comparative Feature-Based Time-Series Classification.* IEEE Transactions on Knowledge and Data Engineering, 2014. **26**.
8. Aggarwal, C.C., *Outlier Analysis*. 2013, New York, NY, United States: Springer New York.
9. Braun, J., *Anomaly detection for solar thermal parabolic trough power plants with artificial intelligence*, in *Information technology*. 2020, Hochschule Esslingen.

20 November 2023 12:23:24