Nearest neighbors method for detecting transient disturbances in process and electromechanical systems

Inês M. Cecílio[a,*], James R. Ottewill[b], John Pretlove[c], Nina F. Thornhill[a]

[a] *Centre for Process System Engineering, Department of Chemical Engineering, Imperial College London, London SW7 2AZ, UK*
[b] *ABB Corporate Research Center, ul. Starowiślna 13a, 31-038 Kraków, Poland*
[c] *ABB Oil, Gas and Petrochemicals, Ole Deviks Vei 10, 0666 Oslo, Norway*

Abstract

Transient disturbances are increasingly relevant in process industries which rely on electromechanical equipment. Existing data-driven methods for detecting transient disturbances assume a distinct amplitude or time-frequency component. This paper proposes a detection method which is more generic and handles any short-term deviation of a measurement from its overall trend, regardless of whether the trend incorporates features such as oscillations, noise or changes in operation level. The method is based on a nearest neighbors technique and builds a vector of anomaly indices which are high for the period with the transient disturbance. The paper includes analyses of the statistical significance of the threshold proposed and of the sensitivity of the parameters, and it also suggests a color map to visualize the detection results. The method is demonstrated on experimental and industrial case studies.

Keywords: Transients, spikes, plantwide, signal analysis, nearest neighbors, anomaly detection

*Corresponding author
Email address: i.cecilio09@imperial.ac.uk (Inês M. Cecílio)

Published in Journal of Process Control 24(9):1382-1393, 2014

1. Introduction

In the past decade, analysis of disturbances in chemical process systems, for example for detection and diagnosis, has shifted from being focused on single control loops [1] to taking a plant-wide approach [2, 3]. This extension made sense because chemical processes are highly interconnected, thus disturbances often propagate throughout the plant along mass and energy flows, and control signals.

Similarly, it would be meaningful to further extend disturbance analysis to the equipment and utilities which service the process. The reason is that these subsystems also interact with the process through energy and signal paths, and therefore disturbances can also propagate between them [4, 5].

The electrical utility and associated electromechanical equipment are increasingly important sources of disturbances to the process industry. Examples of electromechanical equipment in process plants are the electric motors used to drive pumps and compressors. The use of electrical energy to drive process machinery is increasingly common due to the greater energy-efficiency and easier maintenance of electromechanical equipment compared with traditional gas turbines [6]. Electromechanical equipment has its own fault modes and is susceptible to power quality disturbances, that is, deviations of the voltage or current in the power supply from their ideal behavior [7]. In addition, a recent report [8] indicates that power quality disturbances are becoming more frequent.

The susceptibility of the process to disturbances in the electrical utility and associated electromechanical equipment was evidenced in two recent episodes, in March 2013. A water pump failure in the first case [9] and a power cut in the second [10] disrupted two industrial gas processing plants and temporarily stopped the supply of natural gas to the UK.

The extension of process disturbance analysis to electromechanical measurements makes it essential to detect transient disturbances. The reason is that disturbances related to the electrical utility are mostly of a transient nature, caused by power imbalances in the grid which lead to momentary frequency and voltage instabilities [11]. A transient disturbance is defined in this paper as a sudden and short-lived deviation of a measurement from its previous and subsequent trend. After the transient, the measurement may return to its previous trend or follow a different trend. Examples of the former are voltage spikes and deviations caused by sensor faults [12]. Examples of the latter are the responses of the system to step changes, as in Figure 1.

Transient disturbances are well known to practitioners in process industries but previous research on data-driven plant-wide analysis methods has tended to focus on persistent disturbances, which are repetitive and last for long time horizons [2, 3, 13]. Methods for persistent disturbances, however, are not appropriate for transient disturbances because they rely on the repetition of the abnormal dynamic episode [14, 15, 16, 17].

Typically, the data-driven methods for detecting transient disturbances described in academic literature rely on features such as time-domain amplitude or components in the time-frequency domain to distinguish transients from the normal trend of the measurement. Traditional methods of statistical process monitoring are examples of the former, using thresholds on the amplitude to flag unwanted operation levels. However,

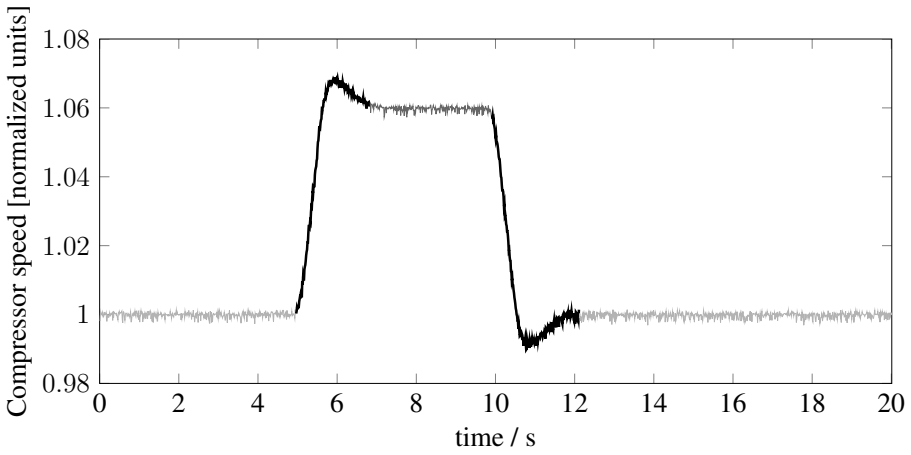


Figure 1: Reference example: time series of the shaft speed of a compressor, with transients around 5 and 11 s. The compressor speed is normalized by its initial value.

this approach is not appropriate if the dynamics of the system generating the measurement is oscillatory or cyclical in nature. Misra et al. [12] use wavelet decomposition to identify the transients in the time-frequency domain. The authors assume that only the parts of the signal affected by transient disturbances map to wavelet coefficients of high amplitude in the lower scales.

The detection method proposed in this paper is, however, more generic than these previous contributions in that it handles any short-term deviation from the overall measurement trend, irrespective of the frequency components or relative amplitude of that deviation. To do that, the basic idea is to consider the measurement as a time series and hence the transient disturbance as the unusual segment. This problem is framed as an anomaly detection problem and solved with nearest neighbors methods. Furthermore, the method proposed does not require the development of models of routine or healthy operation, in contrast to statistical process monitoring techniques.

The contributions of this paper are (i) a method to detect transient disturbances, (ii) statistical analysis of the threshold for detection, (iii) recommendations and analysis of parameter values, and (iv) a color map to visualize the results of the transient detection analysis in a compact way that suggests the propagation of the disturbance through the process.

The paper is organized as follows. Section II presents the background and related work on the nearest neighbors concept. Section III explains the method proposed in detail, analyses the statistics of the threshold for detection, and presents the color map to visualize the results. These contributions are illustrated on a reference example, obtained from experimental work with a gas compression rig. Section IV recommends appropriate values for the parameters of the method and analyses the sensitivity of the detection results to those values. Section V presents an industrial case study and uses it to test the methods. The paper ends with comments and conclusions.

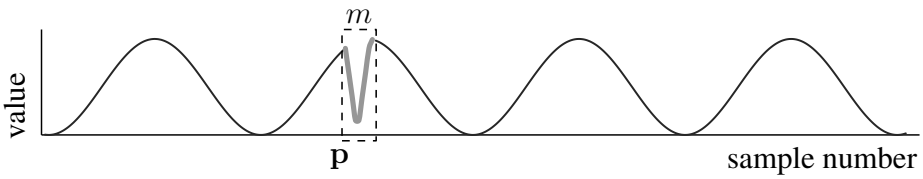


Figure 2: Time series with anomalous segment $\mathbf{p}$ (m-sample long).

2. Background

This section introduces the use of nearest neighbors to detect anomalous segments in a time series. This technique relies on a similarity measure, and on using the similarity assessment to define an anomaly index for each segment of the time series. Hence, this section also discusses similarity measures and anomaly index definitions.

2.1. Detecting anomalous segments with nearest neighbors

If a measurement affected by a transient disturbance is viewed as a time series, then the sequence of samples, or segment, corresponding to the transient will be distinct from the normal trend of the time series. In other words, this segment will be considered anomalous. Figure 2 shows a figurative time series with an anomalous segment, which is emphasized by its thick grey line and the box.

Finding anomalous segments in a time series is a known problem in data mining and is classified as a particular case of anomaly detection, an area which has been comprehensively reviewed in [18]. The techniques proposed for the detection of anomalous segments fall into one of three categories - information theoretic, classification-based or based on nearest neighbors - and all rely on the assumption that the normal behavior of a time series follows a defined pattern hence any segment not conforming to this pattern is an anomaly.

Nearest neighbors techniques are widely applied for anomalous segment detection [18]. The basic idea is to use a similarity measure to evaluate the similarity between each segment in the time series and all other segments. Segments that are similar, *i.e.* with similar sequences of samples, are known as near neighbors. These should be distinguished from near-in-time neighbors, which are adjacent segments with overlapping samples [16]. The similarity of a segment to its nearest neighbors is the basis for its anomaly index. In particular, if a segment is anomalous then its similarity to any other segment will be low.

The effectiveness of the nearest neighbors approach in detecting anomalous segments has been demonstrated on time series with trends which resemble those of process measurements with transient disturbances. Examples include electrocardiogram signals [19, 20] and current measurements from a Space Shuttle control valve [20].

Additional reasons to choose this approach are its autonomy from models of normal and anomalous data, in contrast to classification-based techniques, and the use of a measure that is sensitive even when only one anomaly is present, in contrast to information theoretic techniques [18].

Table 1: Definitions of anomaly index reported in the literature.

Anomaly index	References
Distance to k^{th} nearest neighbor	[19, 20]
Sum of k distances to k nearest neighbors	[24, 25]
Number of nearest neighbors within distance D (inverse)	[26, 27]
Fraction of k nearest neighbors having the segment in their own k nearest neighbors (inverse)	[28]
Density of k nearest neighbors relative to own density	[29, 30]

2.2. Similarity measures for time series

This paper uses the Euclidean distance metric to measure similarity between the segments of the analyzed measurement. The Euclidean distance is defined as the 2-norm of the displacement vector between two points **p** and **q** in an n-dimensional space (1). One such point is the geometrical representation of a time series segment with n samples, where each sample is one spatial coordinate. One reason to use the Euclidean distance is that a distance of zero, in other words, maximum similarity, occurs only when two segments are equal in all n samples, *i.e.* $\mathbf{p} = \mathbf{q}$. Another reason is that the Euclidean distance is affected by the scale of the segments. The scale takes into account the variance of the segment and, in real systems, the units of the measurement it refers to. This is desirable for the detection of transient disturbances since scale reflects variance, and changes in variance may be considered a disturbance.

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{1}$$

Other measures reported in the literature include the cosine similarity [21] and related correlation measure [22], and Dynamic Time Warp (DTW) [23]. However, these measures have properties which are less desirable for the detection of transient disturbances. The cosine similarity and correlation are insensitive to changes in variance between two segments. DTW is a modified version of the Euclidean distance which attempts to maximize the matching of segments **p** and **q** by deforming the time axis. For the detection of transient disturbances, this property leads to less distinction between a short-duration transient and, for example, an oscillation in the background.

2.3. Anomaly index definitions

In nearest neighbor-based anomaly detection, the similarity assessment is used to define an anomaly index for each segment of the time series. This paper uses the distance between that segment and its k^{th} nearest neighbor.

Table 1 gives references for the use of this definition of anomaly index, as well as other definitions also reported in the literature. These other definitions are less adequate for the detection of transient disturbances in industrial process measurements. The reasons are listed below.

- Summing the distances from a segment to its k nearest neighbors is less sensitive to quantized measurements due to the possibility that a large part of k nearest neighbours have the same distance. Quantized measurements are common in industrial data to minimize the storage space.
- Counting the number of neighbors of a segment within a certain distance D requires the parameter D which is less intuitive to set than the integer k. Furthermore, the resulting anomaly index is less discriminating because the number of neighbors is an integer.
- Finding the k nearest neighbors of a segment and counting how many of these neighbors have that segment in their own k nearest neighbours is also less discriminating due to being a ratio of integers.
- The density ρ of the neighborhood of a segment is defined as the inverse of the distance to its k^{th} nearest neighbor, that is, $\rho = {}^{1}/_{d_k}$. This approach compares the density of the neighborhood of a segment to the densities of the neighborhoods of its k nearest neighbors. It is suitable if the anomalous segment has a significantly less dense neighbourhood. However, this is not the case when there are more than one identical anomalous segments, or when the underlying trend of the measurement is neither periodic nor steady.

3. Detection of transient disturbances

This section explains the method proposed to detect and characterize transient distubances. A reference example is first presented and then used to illustrate the explanation.

3.1. Reference example

The time series plotted in Figure 1 represents the shaft speed of a compressor during 20 seconds, measured at 1 kHz. The data was obtained from experimental work using a gas compression rig, located at ABB Corporate Research Centre, Kraków, Poland. Two step changes, around 5 and 11 s, were imposed in the drive of the compressor by changing its speed set-point, resulting in the two transients highlighted in the figure by the black line. The objective of the proposed method is to detect those transients.

3.2. Time series

The method proposed handles the data from a measurement as a time series. A time series X is a finite sequence of n samples, taken at strictly increasing time instants. When the interval between each sample is constant, Δt, the time series can be represented by the sequence of samples in the form $\{x_1, x_2, \cdots, x_n\}$.

The paper additionally deals with two subsets of a time series: segments and embedded vectors. Segment refers to any sequence of samples from the time series that are ordered in time. An embedded vector is a particular type of segment, with m samples τ instants apart, that is $\begin{bmatrix} x_i & x_{i+\tau} & \cdots & x_{i+(m-1)\tau} \end{bmatrix}$. Parameter m is known as embedding dimension and τ as embedding granularity.

3.3. Method

The basis of detecting transient disturbances is the detection of anomalous segments in its time series. Detection of anomalous segments is a data-driven procedure based on computing an anomaly index for each embedded vector of the time series.

3.3.1. Embedding matrix

Equation (2) shows the embedding matrix, $\mathbf{X}$, constructed from the time series $\{x_1, x_2, \cdots, x_n\}$. Each row in the embedding matrix is an embedded vector, defined as above, and each adjacent embedded vector in the matrix lags the previous by δ samples, a parameter known as embedding step. Figure 3 shows the selection, from a symbolical time series represented by dots, of three embedded vectors to be adjacent in the embedding matrix.

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{N_E} \end{bmatrix} = \begin{bmatrix} x_1 & x_{1+\tau} & \cdots & x_{1+(m-1)\tau} \\ x_{1+\delta} & x_{1+\delta+\tau} & \cdots & x_{1+\delta+(m-1)\tau} \\ \vdots & \vdots & & \vdots \\ x_{1+(N_E-1)\delta} & x_{1+(N_E-1)\delta+\tau} & \cdots & x_{1+(N_E-1)\delta+(m-1)\tau} \end{bmatrix} \tag{2}$$

It should be noted that setting parameter δ as a multiple of τ such that $\delta = \alpha\tau$ is equivalent to subsampling the time series by a factor of τ and using $\delta = \alpha$.

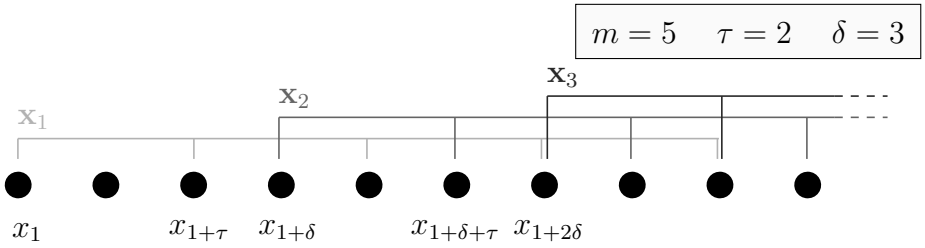


Figure 3: Representation of a time series of samples x_i and the selection of the first three embedded vectors, given parameters m, τ and δ.

The number of embedded vectors, N_E, in an embedding matrix is limited by the total number of samples n of the time series according to Equation (3). The incomplete brackets indicate a floor function, which maps a real number to the largest previous integer.

$$1 + (N_E - 1)\delta + (m-1)\tau \leq n \quad \Leftrightarrow \quad N_E = \left\lfloor \frac{n - (m-1)\tau - 1}{\delta} + 1 \right\rfloor \tag{3}$$

Before the similarity assessment, each embedded vector in the matrix is mean-centered. The effect of this step is that embedded vectors which have identical trends but different ranges of their numerical values will become similar. This is illustrated in Figure 4. Figure 4a plots four embedded vectors from the reference example before mean-centering, and Figure 4b plots the same embedded vectors after mean-centering. The tones of grey relate the embedded vectors to their location in the original time series. The horizontal axis refers to the position of a sample in an embedded vector. The figure illustrates that embedded vectors previously at lower and higher ranges, dark and light grey lines respectively, become close after mean-centering. This step is optional and used when different levels of operation and long-duration ramps should be considered normal.

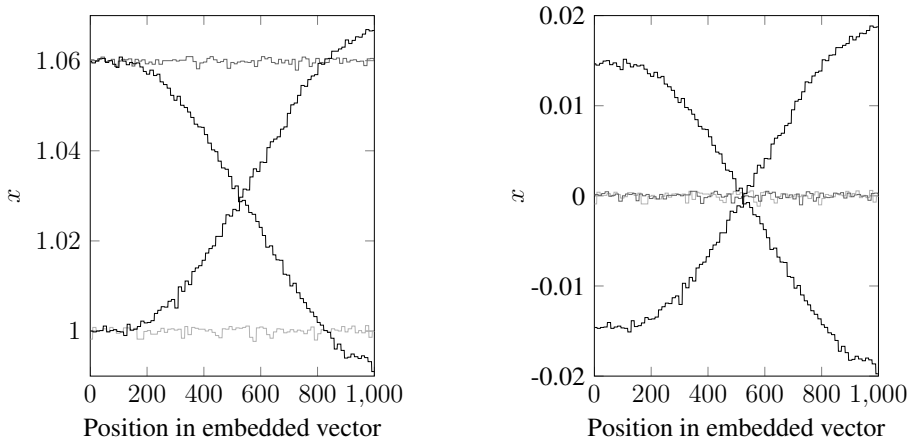
3.3.2. Similarity

Each embedded vector is then compared to every other embedded vector, using the Euclidean distance metric. Section 2.2 discussed the advantages of this measure over other similarity measures for the problem at hand. A simple implementation of this comparison is with two nested loops, running along the rows of the embedding matrix, which will generate a $N_E \times N_E$ matrix of distances.

3.3.3. Anomaly index vector

An anomaly index ai is then attributed to each embedded vector $\mathbf{x}_j$ as the k^{th} smallest distance between $\mathbf{x}_j$ and every other embedded vector. This is denoted by d_k, the distance to its k^{th} nearest neighbor.

It should be noted that distances between $\mathbf{x}_j$ and its near-in-time embedded vectors are excluded from this neighborhood assessment, to avoid their being falsely detected as near neighbors. Near-in-time embedded vectors are defined as embedded vectors



(a) Original embedded vectors.

(b) Mean-centered embedded vectors.

Figure 4: Selected embedded vectors from the reference example, with grey tones relating to their position in the original measurement. Each embedded vector covers 1 s. Parameters are $m = 1001$, $\tau = 1$ and $\delta = 1$.

with at least one sample in common. For example, when one considers $\mathbf{x}_j = \mathbf{x}_1 = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}$, the embedded vector $\mathbf{x}_m = \begin{bmatrix} x_m & x_{m+1} & \cdots & x_{2m-1} \end{bmatrix}$ is the last of its near-in-time embedded vectors, for $\tau = 1$ and $\delta = 1$.

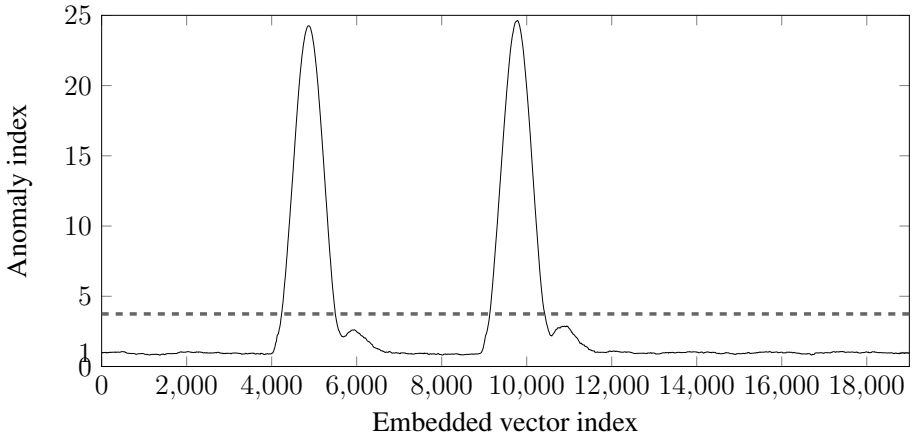
Attributing an anomaly index ai to each embedded vector generates an anomaly index vector $\mathbf{ai}$ of N_E anomaly indices. Figure 5a shows the anomaly index vector for the reference example, with $m = 1001$ and $k = 3$. These parameter values were thought sensible for the initial attempt. The vector was normalized by its median so that $ai = 1$ now approximates the average anomaly index of non anomalous embedded vectors. The next section justifies this statement.

3.4. Significance level

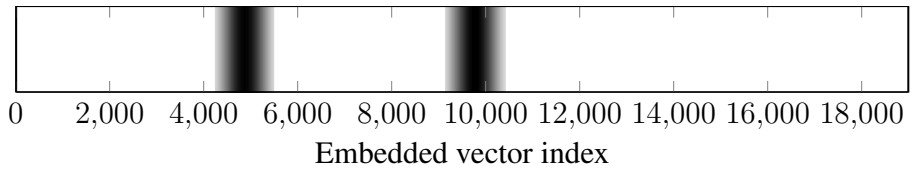
Classifying an embedded vector as anomalous is done based on a threshold on the value of the anomaly indices. This threshold is based on the probability distribution of the anomaly indices, and the reason is that embedded vectors which correspond to transient disturbances have high anomaly indices and are a small fraction of all embedded vectors. This is true because, by definition, transient disturbances are scarce in the time series analyzed. Intuitively, the anomaly indices of anomalous embedded vectors stand out in the anomaly index vector, as in Figure 5a, in which the anomalous embedded vectors have numbers around 5000 and 10000.

In statistical literature, values numerically distant from the rest of the data are treated as outliers [31]. The cutoffs to isolate those values are based on robust statistics, which are less influenced by the outlier values than classical statistics and are, thus, more representative of the rest of the data [32].

The threshold proposed here is based on two robust statistics, the median and the interquartile range, IQR, according to Equation (4). The median is defined as the 50^{th} percentile (Q_2), and IQR as the difference between the 75^{th} and 25^{th} percentiles. When



(a) Normalized anomaly index vector. The dashed line indicates the detection threshold.



(b) Color plot obtained after applying the threshold and mapping the detected indices onto a color scale. Darker tones indicate higher anomaly indices.

Figure 5: Anomaly index vector for the reference example ($k = 3$).

the anomaly index vector is normalized by its median, then the $Q_2(\mathbf{ai})$ term in Equation (4) is one. Figure 5a shows the threshold for the reference example with a dashed line.

$$ai > Q_2(\mathbf{ai}) + 6 \times \text{IQR}(\mathbf{ai}), \tag{4}$$

The performance of the detection threshold can be quantified by analysing its behavior under null hypotheses of time series with no anomalies. The remainder of this section shows that the probability of false positives with the proposed threshold is less than one in a million. This is demonstrated for the cases of (i) steady state operation with only random noise, (ii) operation with non-random variability, and (iii) oscillatory operation.

The underlying assumption is that the anomaly index vector generated from a time series without anomalous embedded vectors approximately fits to a gamma distribution with a skewness of less than 0.77.

Skewness is a measure of asymmetry of a frequency distribution. The anomaly index vector in Figure 5a will have a skewed distribution because there are a few large anomaly indices and many small ones. The statistical gamma distribution commonly describes waiting times between events and, specifically, it models sums of exponentially distributed random variables [33]. For this reason, it is reasonable to assume that this distribution can model the anomaly index vector, given that anomaly indices are distances, hence sums of squared differences. Several authors also refer to the use of the gamma function to model the frequency distribution of nearest neighbor distances [34, 35, 36]. Furthermore, the gamma distribution presents a number of properties that agree with anomaly indices, such as being defined only for positive real numbers, being positively skewed and converging to a Gaussian distribution when its skewness tends to zero.

Appendix A confirms the validity of assuming that the anomaly index vector generated from a time series without anomalous embedded vectors approximately fits to a gamma distribution with a skewness smaller than 0.77. The demonstration includes, as time series without anomalous embedded vectors, (i) steady state operation with only random noise, (ii) operation with non-random variability, and (iii) oscillatory operation.

When this assumption holds, then the probability that the proposed threshold causes false detections is less than one in a million. This logical relation is proved in Appendix B.

3.5. Outputs of the detection

The numerical outputs of the detection method include (i) the number of transient disturbances, (ii) the initial and final time indices of each disturbance, and (iii) a measure of the severity of each disturbance. A fourth output is a color plot.

Each group of consecutive embedded vectors classified as anomalous identifies one transient disturbance. Therefore, the indices j_I and j_F, respectively belonging to the initial and final anomalous embedded vectors in the group, can be used to estimate the initial and final time indices of the corresponding transient. To this end, it is assumed that an embedded vector is anomalous if at least half of it covers the transient. This is

Table 2: Numerical outputs of the detection method for the reference example.

Number of disturbances	initial time index $\hat{i}_I$	final time index $\hat{i}_F$	severity $\mathcal{S}$
2	4733	5996	15.4
	9630	10923	15.1

mathematically described in Equation (5), where $\hat{i}_I$ and $\hat{i}_F$ are the estimated initial and final time indices of the transient disturbance.

$$\begin{aligned} \hat{i}_I &= \left\lfloor 1 + (j_I - 1)\delta + \frac{(m-1)\tau}{2} \right\rfloor \\ \hat{i}_F &= \left\lceil 1 + (j_F - 1)\delta + \frac{(m-1)\tau}{2} \right\rceil \end{aligned} \tag{5}$$

The incomplete brackets in Equation (5) indicate floor and ceiling functions, respectively, which map a real number to the largest previous and the smallest following integer.

The severity $\mathcal{S}$ of a transient disturbance is a single value which captures how distinct the transient disturbance is from the underlying trend of the measurement. It is given by Equation (6), which represents the average of the anomaly indices ai_j from the anomalous embedded vectors with indices between j_I and j_F.

$$\mathcal{S}_t = \frac{\sum_{j=j_I}^{j_F} ai_j}{j_F - j_I + 1} \tag{6}$$

Table 2 exemplifies the numerical outputs of the detection method for the reference example.

An additional output of the method is a color plot showing how the analyzed measurement evolved in time with regards to the presence of transient disturbances. Figure 5b shows the color plot for the reference example. Anomaly indices corresponding to anomalous embedded vectors are mapped onto a two-color scale, thus flagging when a measurement is most affected by a transient disturbance. The maximum anomaly index is mapped onto one color, in this case black, the anomaly index corresponding to the detection threshold onto the other color, in this case white, and the anomaly indices in between are linearly mapped onto a weighed sum of the two colors. Anomaly indices which are below the detection threshold are mapped to the minimum limit of the color scale, for a better contrast. The anomaly index values represented in the color plot can be assigned to time indices in order to visualize the presence of transients over time. This is done by converting from the indices of embedded vectors, j, in to time indices i using the principle in Equation (5).

The value of the color plot is to allow the visualization of the detection results in a way that is compact and that clearly suggests the propagation of the disturbance through the process. This is important when analysing huge data sets [37], particularly in industry.

3.6. Computational effort

The basic nearest neighbors technique is $\mathcal{O}(N_E^2)$, where N_E is the number of embedded vectors. This is because every embedded vector has to be compared to every other embedded vector, giving a total of $^{N_E(N_E-1)}/_2$ operations to compute a distance. It can also be memory intensive if all these distances are stored.

Figure 7 indicates the time taken by the method, using the reference example with different subsamplings. If necessary, the computation time could be shortened based on approximation [20] or pruning [38].

4. Parameter settings and sensitivity

To generate the anomaly index vector, the following parameters have to be selected:

- Embedding granularity τ
- Embedding dimension m
- Embedding step δ
- Number of nearest neighbors k

The objective of this section is to find adequate values for these parameters, given the dynamics of a particular system. The section starts by relating the physical meaning of the parameters with the dynamics of the system, and then recommends the best parameter values and analyses the sensitivity of the detection results to those values.

4.1. Relation between parameters and dynamics of the system

The embedding granularity τ is the number of sampling intervals, Δt, between each sample included in an embedded vector. It should, therefore, be small enough to characterize the main trends of the transient disturbances. For example, if the interesting feature were an oscillation, τ should at least satisfy the Nyquist sampling criterion for that oscillation.

The embedding dimension m is the number of samples in an embedded vector and, together with τ and Δt, defines the duration of the embedded vector according to $(m-1)\tau\Delta t$. This duration should match the time scale of the transient disturbances, otherwise the embedded vectors will capture events at other time scales. An example in the reference case, in Figure 1, would be to detect the whole prominence between 5 s and 12 s as one disturbance if m is large enough. Another example would be to capture noise as a disturbance if m is too small.

The embedding step δ is the number of sampling intervals, Δt, between consecutive embedded vectors. Thus, it can influence the accuracy of the estimated start and end indices of the transient disturbances. This accuracy will be within the range $\pm\delta/2$ samples. The actual accuracy depends on the chosen section of data, in particular, on the elapsed time between the transients and the start of the time series. The reduced time accuracy caused by large values of δ may lead to a failure to detect a transient, if δ is greater than twice the width of that transient. Therefore, one should ideally chose $\delta = 1$, however a larger δ will accelerate the computation of the anomaly index vector since the method is $\mathcal{O}(N_E^2)$ and N_E is proportional to $1/\delta$.

The meaning of k, the number of nearest neighbors, can be described in the following way: if there is a group of at least $k+1$ identical embedded vectors, these will be considered non-anomalous. Therefore, the minimum value for k has to be the number of identical transients, to avoid not detecting the transients. At the same time, k has to be smaller than the number of identical non-anomalous embedded vectors, in order to avoid these being considered anomalous.

Table 3 summarizes the roles of the parameters, as discussed in this section.

4.2. Recommendations for parameters and analysis of sensitivity

This section determines the best values of the parameters and analyses the sensitivity of the detection results in the range of those values. This is done by generating anomaly index vectors from the reference example with different parameter values, and evaluating the performance of the results.

The performance is defined by noting that embedded vectors fall into one of two categories: those which cover a transient, $\mathbf{x}_j$, and those which do not cover a transient, $\mathbf{x}_l$. In the following, the total number of $\mathbf{x}_j$ is denoted as N_E^{trans}, and the total number of $\mathbf{x}_l$ is denoted as N_E^{norm}. Clearly, $N_E^{trans} + N_E^{norm} = N_E$, where N_E is the total number of embedded vectors. For parameter optimization, the indices of $\mathbf{x}_j$ and $\mathbf{x}_l$ are identified visually beforehand. In this section, this is done with the reference example.

If the detection procedure is perfect, then all $\mathbf{x}_j$ are classified as anomalous and all $\mathbf{x}_l$ as normal. However, if the detection is not perfect, some embedded vectors will not be correctly classified. Classification of the embedded vectors as anomalous or normal is based on the threshold proposed in section 3.4.

Metric FN (false negatives) (7) assesses the case of anomalous embedded vectors being incorrectly classified as normal. To that end, it counts the embedded vectors covering a transient, $\mathbf{x}_j$, which were correctly classified as anomalous, N_{TP}, where TP stands for true positives, and compares this number with the total number of $\mathbf{x}_j$, N_E^{trans}. This metric is one if the detection procedure is perfect, and zero if all $\mathbf{x}_j$ were incorrectly classified as normal.

$$FN = \frac{N_{TP}}{N_E^{trans}} \tag{7}$$

Metric FP (false positives) (8) assesses the case of normal embedded vectors being incorrectly classified as anomalous. It counts the embedded vectors not covering a transient, $\mathbf{x}_l$, which were correctly classified as normal, N_{TN}, where TN stands for true negatives, and compares this number with the total number of $\mathbf{x}_l$, N_E^{norm}. This metric is one if the detection procedure is perfect, and zero if all $\mathbf{x}_l$ were incorrectly classified as anomalous.

$$FP = \frac{N_{TN}}{N_E^{norm}} \tag{8}$$

As well as minimizing false negatives and false positives, the default parameter values should also maximize the ratio between the anomaly indices from embedded vectors covering a transient, ai_j, and from embedded vectors from normal operation, ai_l. This metric is denoted as ANR (anomaly-to-normal ratio) and defined by (9).

$$ANR = \frac{\sum_j ai_j}{\sum_l ai_l} \tag{9}$$

The next sections combine the above metrics into a performance index to evaluate the detection results.

4.2.1. Number of samples in a transient, embedding granularity τ and embedding dimension m

Parameters τ and m are directly dependent on the sampling interval, Δt, used to collect the data, hence values for these parameters must be given relative to that sampling interval.

A suitable sampling interval for the detection of transients must lead to enough samples in the transients to guarantee their characterization. The optimal number of samples in a transient was analysed by subsampling the reference time series and considering the number of samples in the first transient, which is the shortest. Figure 6 shows, with the horizontal axis, how the number of samples in the shortest transient influences the performance index. For this optimization, the performance index is the sum of the three objectives, FP, FN and ANR, with ANR normalized by its maximum value.

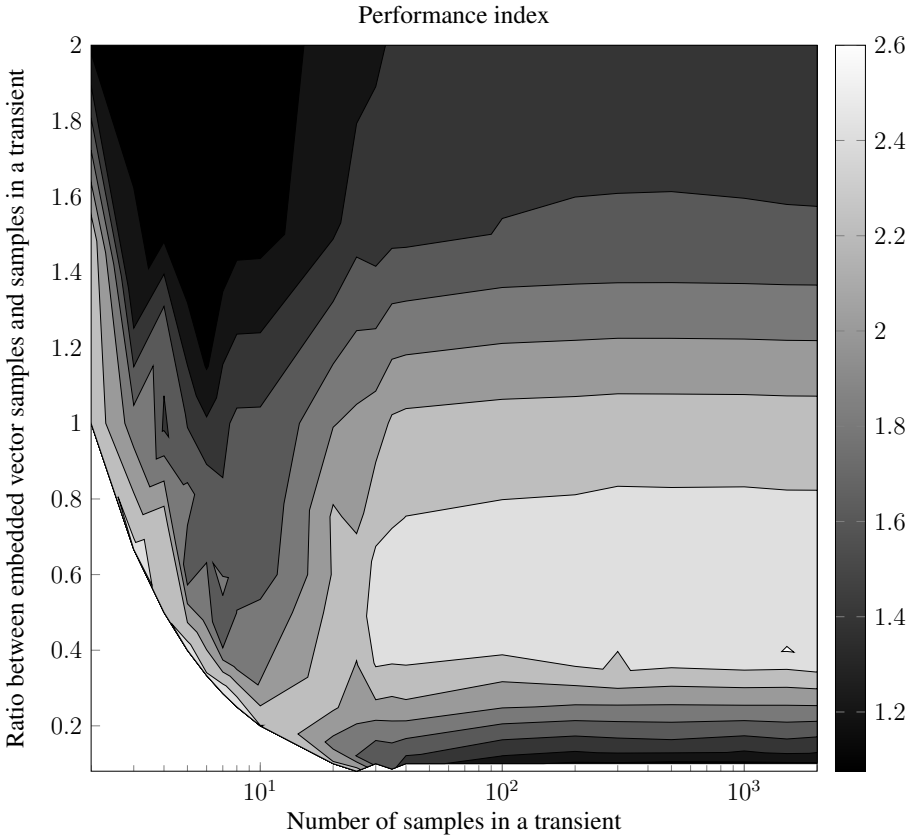


Figure 6: Performance of the detection method as a function of the number of samples in a transient and of the ratio between embedded vector samples and samples in a transient. The performance index can vary between zero and three. Parameters fixed in the analyses were $\tau = 1$, $\delta = 1$ and $k = 3$.

Figure 6 shows higher values of the performance index the more samples charac-

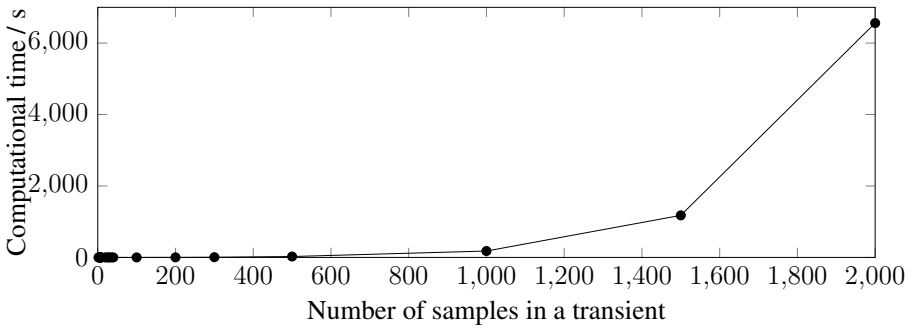


Figure 7: Computational time as a function of the number of samples in a transient. The computational time is an average over the different values of m.

terize the transient, as expected. More importantly, it also shows that above 30 samples the performance is constant. This is important because the fewer samples needed in the time series, the faster is the method, as shown in Figure 7. Since most transient disturbances are abrupt, the use of 30 samples would characterize the transient as well as some of its details.

Users can set such sampling interval based on their knowledge or past experience with the system. For instance, if the experience of the site is that electrical transients have a typical duration of 1 s, then the sampling interval should be 0.033 s. With the suitable sampling interval defined, parameter τ can be set to one.

Figure 6 also analyses, with the vertical axis, the optimal number of samples in the embedded vectors in relation to the number of samples in the transient. It shows that the detection method achieves highest performances when the ratio between the number of samples in the embedding vector and the number of samples in the transient is between 0.4 and 0.75. Choosing the ratio of 0.5 implicitly defines the optimal value for parameter m as 15. Another observation from Figure 6 is that the optimal range 0.4 to 0.75 is consistent above the value of 30 samples in the shortest transients. This is important because it is robust. For instance, if the real dynamics of the system is up to 1.5 times slower than expected, and the transients are characterized by 40 samples instead of 30, $m = 15$ will still lead to similar performance indices.

4.2.2. Embedding step δ

Figure 8 shows the influence of different values of δ on the performance index. For this optimization, the performance index is the sum of the objectives FP and FN because δ influences the accuracy of detection of the start and end of the transient. The values of performance index shown are the average of those from five time series. These are identical to the subsampled time series from the example, but start at different time instants, specifically with a delay of one sample. The reason to calculate performance indices for different starting instants is because the actual accuracy of the results depends on the chosen section of data, specifically the elapsed time between the transients and the start of the time series.

Figure 8 shows that the average performance of the method decreases approxi-

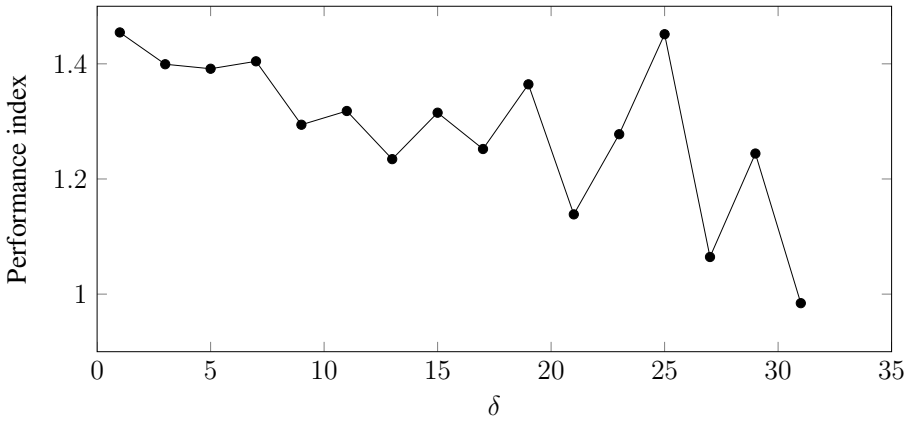


Figure 8: Performance of the detection method as a function of δ. The performance value is an average over five time series with different starting instants, and it can vary between zero and two. Parameters fixed in the analyses were the sampling interval, such that the number of samples in a transient was 30, $\tau = 1$, $m = 15$ and $k = 3$.

mately linearly with δ, as expected from the discussion in section 4.1. Also, the decrease in performance becomes more inconsistent as δ increases. As expected, these results suggest that δ should be set to one. In this example, with 300 samples, the computational time with $\delta = 1$ is below 0.02 s. However, in time series with more samples it may be important to increase δ, given that the method is $\mathcal{O}(1/\delta^2)$. In such cases, it is recommended to keep δ below 5, otherwise Figure 8 shows that the average performance of the method may fall below 90% of the performance achieved with $\delta = 1$. The value of $\delta = 5$ is equal to $m/3$, given that $m = 15$. This means the uncertainty in the time of occurence of the transients is a sixth of the duration of the transient. Ultimately, the uncertainty in the time of occurence of the transients is a user-specific choice.

4.2.3. Number of nearest neighbors k

Parameter k must be at least equal to the number of identical transients. If, at the time of preparation of the data or from historical operation experience, users have an idea of typical numbers of transient disturbances occurring in the site, they may set k accordingly. In principle, however, such numbers are not known a priori, so k should be set high to avoid false negatives.

Figure 9 and Figure 10 show the influence of different values of k on the performance index, which for this optimization is the sum of the three objectives, FP, FN and ANR, with ANR normalized by its maximum value. The seven lines correspond to different values of δ.

Figure 9 focuses on small values of k to show that, for most values of δ, the performance in the reference example increases with decreasing values of k. In particular, $k = 1$ leads to the highest performances in this example, which was expected since the reference measurement has only one transient of each shape. However, the minimum value recommended is $k = 3$, even when only one identical transient is expected. The reason is to allow some margin of error in case the number of identical anomalous em-

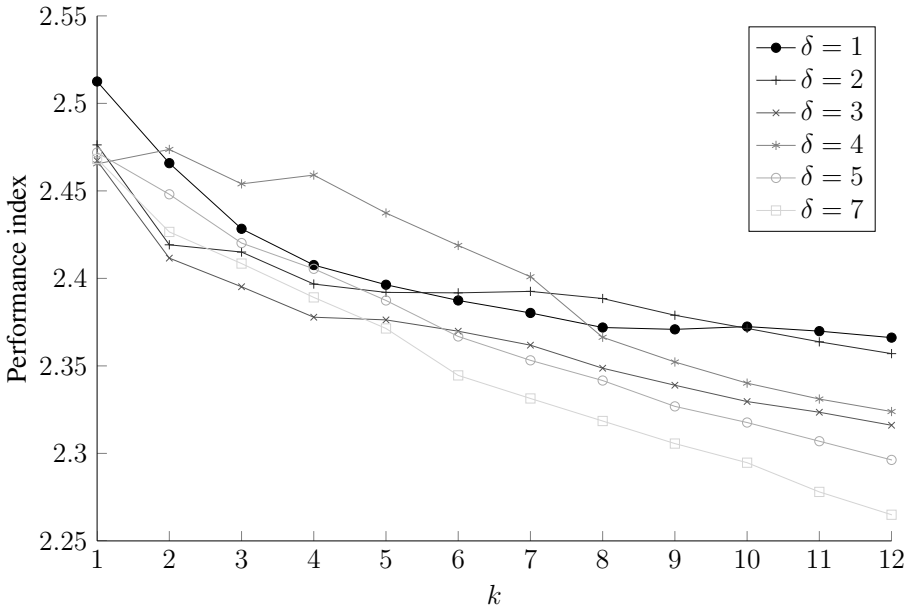


Figure 9: Performance of the detection method as a function of k, for different values of δ. The performance index can vary between zero and three. Parameters fixed in the analyses were the sampling interval, such that the number of samples in a transient was 30, $\tau = 1$ and $m = 15$.

bedded vectors is higher than expected. The reduction in performance, meanwhile, is small, specifically in this example it is less than 3%.

The maximum value of k recommended depends on the dynamics of the system. If time series from routine operation have some periodicity, k should be less than n/N_p, where n is the number of samples in the time series and N_p is the number of samples in one cycle. If, on the other hand, routine time series derive from steady-state operation, then the value of k should not exceed $n/2\delta$. This value is approximately half the number of embedded vectors and derives from the fact that, by definition, the number of non-anomalous embedded vectors needs to be greater than the number of anomalous ones.

Figure 10 supports this last statement. The figure now plots the performance index as a function of $k \cdot \delta/n$, which is approximately the ratio between k and the number of embedded vectors. Ratios above 0.6 show a marked decay in performance, which agrees with the recommendation in the previous paragraph of using ratios below 0.5. For $\delta = 5$, for example, this means that k should not exceed $n/10$.

Table 3 summarizes the values recommended for all the parameters. These values assume the choice of a sensible sampling interval, as discussed in the beginning this section. If the original sampling interval is smaller than necessary, data can be downsampled or the parameters adjusted proportionally.

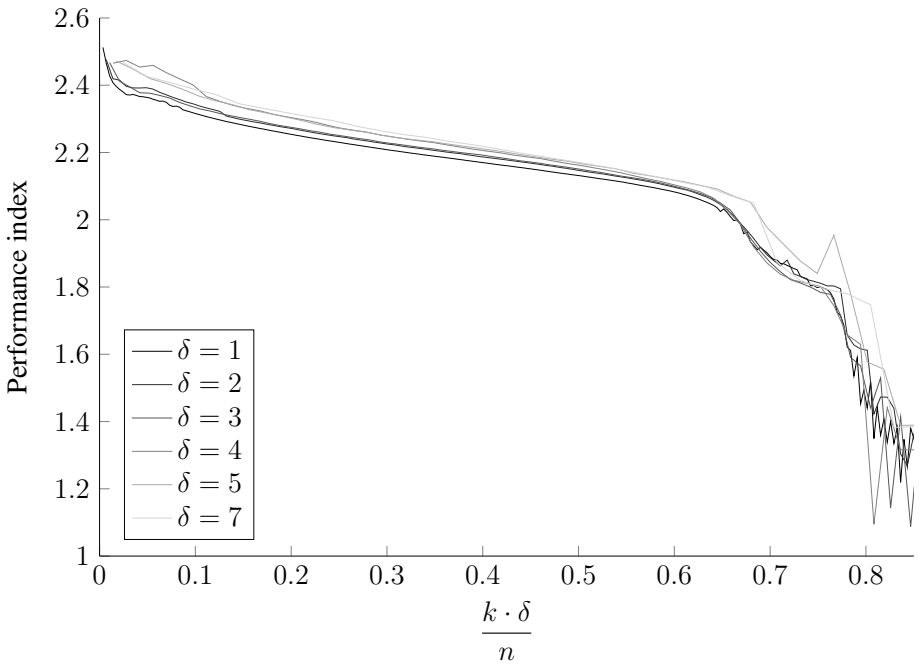


Figure 10: Performance of the detection method as a function of k and δ. The performance index can vary between zero and three. Parameters fixed in the analyses were the sampling interval, such that the number of samples in a transient was 30, $\tau = 1$ and $m = 15$. The number of samples is $n = 300$.

Table 3: Summary of the roles and recommended values for the parameters.

Parameters	Role	Recommended value
number of samples in a transient	characterize the main trends of the transient disturbances	30
τ		1
m	defines the time scale of the transient disturbance	15
δ	defines the time accuracy for the transient disturbance	$\delta \leq 5$
k	defines the maximum recurrence of identical transients	$3 \leq k \leq {}^{n}/_{10}$

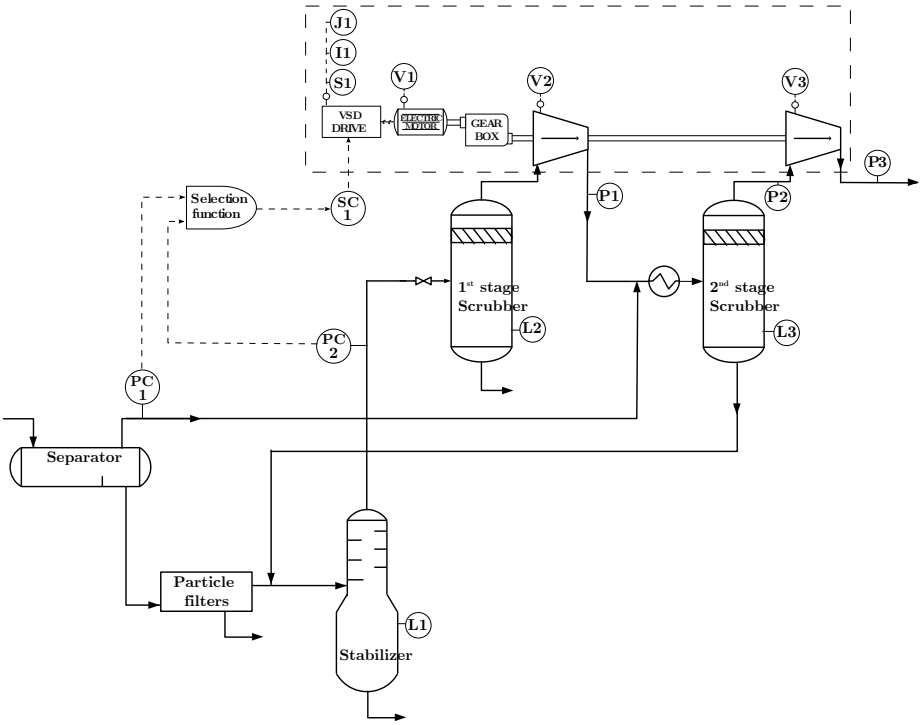


Figure 11: Process schematic for the industrial case study.

5. Application to industrial case study

The test case study is part of an industrial process of gas processing, courtesy of ABB Oil, Gas and Petrochemicals, Oslo, Norway. Figure 11 shows the selected part of the process, which includes a gas-condensate separation section, with a separator, filters and stabilizer, and a gas recompression section, with scrubbers and compressors. The speed of the compressors is used to adjust the pressure in the system, either at the outlet of the separator or at the outlet of the stabilizer.

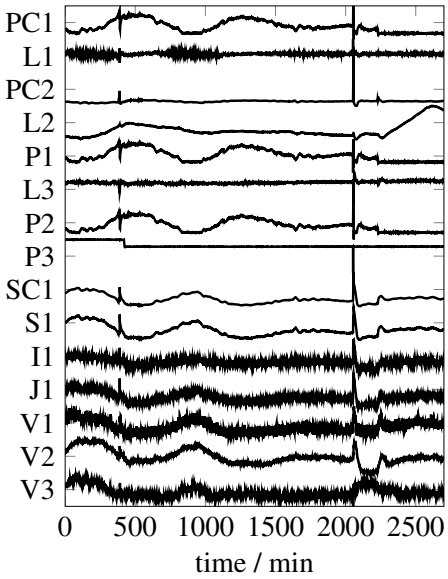
Figure 12a shows the time trends of 15 measurements affected by two sharp spikes happening around minutes 500 and 2000. Each time trend has 5400 samples taken with a sampling interval of 30 s. The measurements include process variables, *i.e.* pressure and level, electrical variables, *i.e.* drive current and power, mechanical variables, *i.e.* speed and vibration, and control signals. It must also be noted that most time trends of the measurements are not constant. Some measurements show oscillations, namely PC1, P1, P2, SC1, S1, J1, V1 and V2, noise, namely I1, V1 and V3, and a ramp, namely L2. This paper excludes ramps from the definition of transients due to being a long-term event.

The expectation for this test case study is that the nearest neighbors method correctly identifies the transient disturbances, while being insensitive to the other behavior, that is, the oscillations, the noise and the ramp. The parameters are defined systemati-

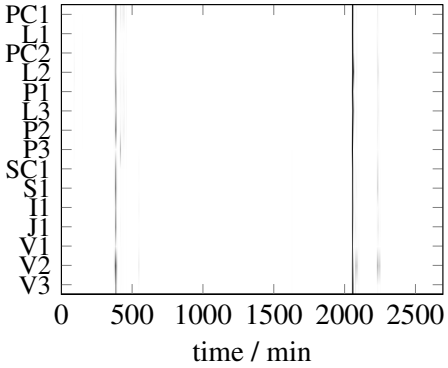
cally, according to section 4. The physical insight used is that the time scale of transient disturbances in process systems is commonly around 10 to 15 min. This is also true in this case study. This insight and the knowledge of the original sampling interval of 30 s indicate that the transients are described by 20 to 30 samples. The recommended number of samples in a transient is 30, which means the data should not be subsampled. The other parameters are set to their recommended values, that is $\tau = 1$, $m = 15$, $\delta = 1$ and $k = 3$. The embedding dimension of $m = 15$ covers 0.75 of the transients with 20 samples, which according to Figure 6 is in the region of highest performances for this number of samples.

The results are shown in Figure 12b using the proposed color plot. The upper limit in the color scale corresponds to the maximum anomaly index in each measurement. These results indicate that the method is able to identify the transient spikes in all measurements, while ignoring noise, oscillations and the ramp. Therefore, these results support the potential of the method for automated uses. It should be noted that the positive detections are true transients and do not result, for example, from noise. The reason is that section 3.4 proved that the method has a probability of false detections of less than one in a million.

The color plot also allows visualization of the order by which the disturbances appeared in each measurement. To that end, Figure 13 reorders the measurements according to the start of the first disturbance, and zooms in to this period. In each measurement, the transient disturbance starts when the color changes from white to light grey. The darker tones reflect the transient becoming more distinct from the rest of the trend. The color plot suggests that the transient first appeared in the gas streams between the gas outlet of the separator (PC1) and the compressors (P1 and P2). The results show that the disturbance is picked up by the control loop which acts on the compressors (SC1). The transient is consequently transmitted to the pressure at the inlet of the first compressor (PC2), and to some electro-mechanical measurements which reflect the operation of the compressors (S1, V2 and J1). Eventually the liquid condensate in the scrubbers is affected (L2 and L3) and the transient recycles back to the stabilizer (L1). This propagation path is suggested by the color plot and it agrees with the physics of the system.



(a) Time trends.



(b) Results of the transient detection shown in the color plot.

Figure 12: Condition of the measurements in the industrial case study.

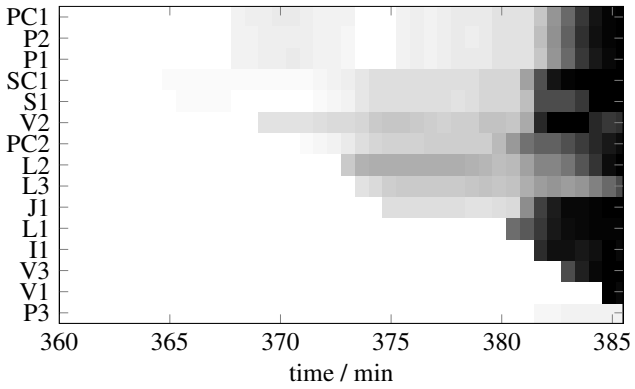


Figure 13: Color plot for the industrial case study, zoomed in to the first transient and reordered to highlight the different times of start of the transient.

6. Conclusion

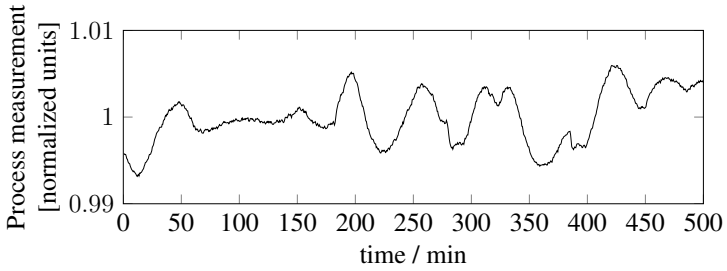
This paper has demonstrated the use of a nearest neighbors technique in the detection of transient disturbances in measurements from chemical process systems and associated electromechanical equipment. The extension of process disturbance analysis to electromechanical measurements is justified by the increasing use of electrically-driven machinery in process industries and the rising number of power quality incidents.

The proposed method is based on the concept of anomaly detection, and was shown to be effective in detecting transients and distinguishing these from other dynamics such as oscillations, ramps, noise and different operation levels. Furthermore, if no transient is present the chosen detection threshold guarantees a probability of less than one in a million of falsely detecting a transient. The method detects step changes as transients, which could be perceived as a limitation of the method. However, in an automated setting, the step changes can be discarded from the list of detected transients because the changes are known interventions.

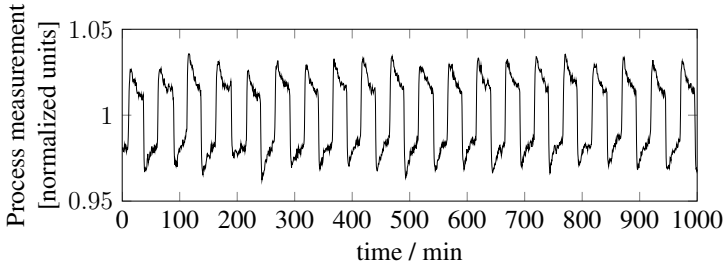
The method requires parameters, thus the paper recommended optimal values and analysed these recommendations. The paper also proposed a color plot to visualize the detection results in a way that is compact and that clearly suggests the propagation of the disturbance through the system.

The detection method is a new contribution to the field of plant-wide disturbance analysis, which has mostly focused on oscillating disturbances. It can also be a preceding step for diagnosis analysis and for the removal of transients in order to enhance the analysis of oscillating disturbances. The case study also showed that the method yields in an automated way what is expected from a visual analysis. As a result, the method is industrially applicable for plant-wide monitoring, as a background tool to explore large amounts of data. Uses in industry for the method and visual output could comprise the understanding of the plant-wide impact of a certain transient, the identification of variables most often affected, and the identification of delays, which suggest a propagation path.

The results show that the method offers informative insights when combined with thorough knowledge of the process.



(a) Case (ii): operation with non-random variability, but no transients.



(b) Case (iii): oscillatory operation.

Figure A.14: Base time series for generation of cases representing normal operation scenarios.

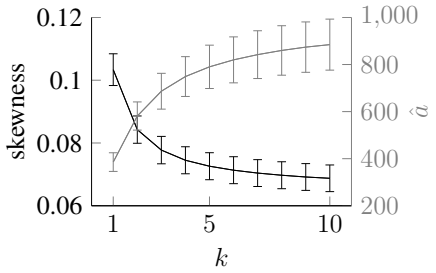
Appendix A. Anomaly index vector fitting to gamma distribution with skewness smaller than 0.77

This appendix confirms the validity of assuming that the anomaly index vector generated from a time series without anomalous embedded vectors approximately fits to a gamma distribution with a skewness smaller than 0.77. This is done using Monte-Carlo simulations, which generate time series without transient disturbances, and observing that a significant fraction of these time series lead to anomaly index vectors with an acceptable fit to the assumption.

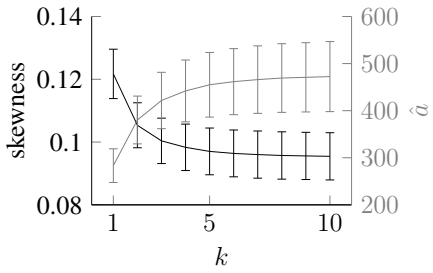
To be representative of different operation scenarios, this demonstration uses three cases: (i) steady state operation, (ii) operation with non-random variability, and (iii) oscillatory operation. The time series used for case (i) is synthetic while the time series for cases (ii) and (iii) belong to physical measurements monitored in a real gas processing plant and are shown in Figure A.14.

To begin with, three sets of N_{MC} random time series, representing Gaussian noise, are generated and added to the three base time series, thus making up $3N_{MC}$ time series. The anomaly index vector is then computed for each of the $3N_{MC}$ time series by the method proposed in section 3, with $m = 50$, $\tau = \delta = 1$ and $k = [1 : 10]$. The choices of these parameters follow the rules proposed in section 4.

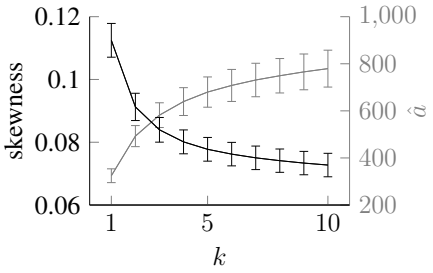
The $3N_{MC}$ anomaly index vectors are assumed to follow gamma distributions. This distribution has two parameters, a and b, respectively describing the shape and



(a) Constant operation.



(b) Operation with non-random variability.



(c) Oscillatory operation.

Figure A.15: Estimated gamma shape parameter (grey line) and derived skewness (black line) for three sets of clean signals.

scale of the distribution. Therefore, this section derives maximum likelihood estimators for the $3N_{MC}$ pairs of parameters a and b. The maximum likelihood method finds the parameters for a given statistical distribution that maximize the likelihood of observing the existing data.

Statistics for the estimated shape parameter $\hat{a}$ and derived skewness (A.1) are then calculated for each set of N_{MC} parameters $\hat{a}$. Figure A.15 shows these statistics for each set. The lines follow the mean of the estimated properties and the error bars represent one standard deviation. The plots indicate skewness values significantly below the assumption limit of 0.77, as we wanted to show.

$$\text{skewness}_{\Gamma} = \frac{2}{\sqrt{a}} \tag{A.1}$$

Appendix B. Probability of false detection in gamma distributions with skewness smaller than 0.77

This appendix proves numerically that, in gamma distributions, a threshold of $Q_{2,\Gamma} + 6 \times \text{IQR}_{\Gamma}$ leads to a detection rate of less than one in a million if that distribution has a skewness of less than 0.77.

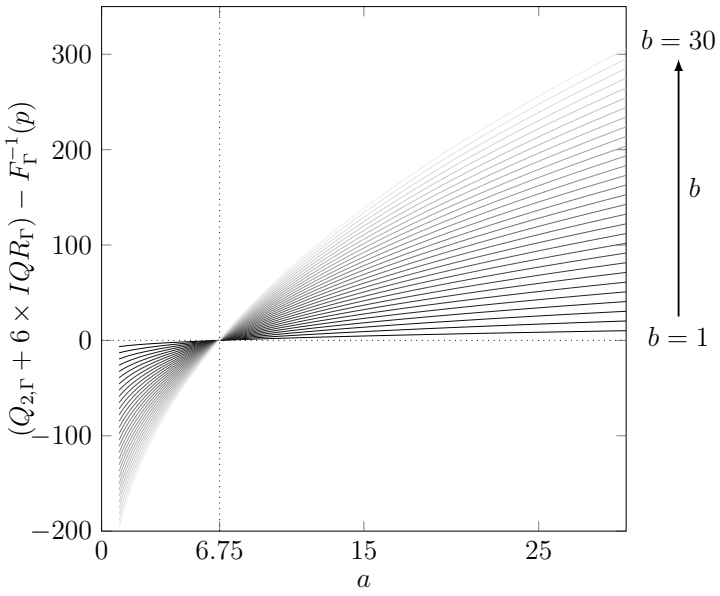


Figure B.16: Influence of gamma parameters a and b on the relation between the proposed threshold and the gamma quantile function, with $p = 0.999999$.

The gamma distribution has two parameters, a and b, respectively describing the shape and scale of the distribution. The gamma quantile function $F_\Gamma^{-1}(p)$ for the probability of detection $p = 0.999999$ and the statistics $Q_{2,\Gamma} + 6 \times \text{IQR}_\Gamma$ can then be calculated for gamma distributions with different pairs of parameters $\{a, b\}$. Because there is no analytical solution to the gamma quantile function, the $F_\Gamma^{-1}(p)$ values were found with the MATLAB function `gaminv`, which uses Newton's method to converge to the solution.

Figure B.16 plots $(Q_{2,\Gamma} + 6 \times \text{IQR}_\Gamma) - F_\Gamma^{-1}(p)$ for each $\{a, b\}$. With the proposed threshold, a detection rate equal or lower than one in a million happens if $F_\Gamma^{-1}(p) \leq Q_{2,\Gamma} + 6 \times \text{IQR}_\Gamma$. Figure B.16 shows that, in gamma distributions, this happens for a shape parameter $a \geq 6.75$ and is independent of the scale parameter b. A gamma distribution with shape parameter $a = 6.75$ has, by definition (A.1), a skewness of 0.77, as had to be proved.

Acknowledgment

The authors gratefully acknowledge the financial support from the Portuguese Foundation for Science and Technology (FCT) under Fellowship SFRH/BD/61384/2009 and the Marie Curie FP7-IAPP project "REAL-SMART - Using real-time measurements for monitoring and management of power transmission dynamics for the Smart Grid", Contract No: PIAP-GA-2009-251304.
Inês M. Cecílio would also like to thank P. Lipnicki, D. Lewandowski, and M. Wojcik of ABB Corporate Research Center, Kraków, Poland, for enabling and helping with the experimental work, and H. Fretheim, and K. Rapp of ABB Oil, Gas and Petrochemicals, Oslo, Norway, for providing data and process insights to support this paper.

References

[1] B. Huang, S. L. Shah, Performance Assessment of Control Loops: Theory and Applications, Springer, 1999.

[2] N. F. Thornhill, A. Horch, Advances and new directions in plant-wide disturbance detection and diagnosis, Control Engineering Practice 15 (10) (2007) 1196–1206.

[3] Y. Shardt, Y. Zhao, F. Qi, K. Lee, X. Yu, B. Huang, S. Shah, Determining the state of a process control system: Current trends and future challenges, The Canadian Journal of Chemical Engineering 90 (2) (2012) 217–245.

[4] I. M. Cecílio, S.-L. Chen, N. F. Thornhill, Importance of auxiliary systems for process fault detection and diagnosis, in: Proceedings of the 19th Mediterranean Conference on Control & Automation (MED), IEEE, Corfu, 2011, pp. 952–957.

[5] A. Lindholm, H. Carlsson, C. Johnsson, A general method for handling disturbances on utilities in the process industry, in: Proceedings of the 18th IFAC World Congress, IFAC, Milano, 2011, pp. 2761–2766.

[6] REAL-SMART, Marie Curie FP7-IAPP project, online (October 2013), http://www3.imperial.ac.uk/realsmart.

[7] M. H. J. Bollen, Understanding Power Quality Problems: Voltage Sags and Interruptions, Wiley-IEEE Press, 2000.

[8] Statnett, Systemdrifts- og markedsutviklingsplan 2012, online (April 2013), http://www.statnett.no/Documents/Kraftsystemet/Systemansvaret/Statnett_SMUP_24.05_lnk_Low.pdf (2012).

[9] S. Pfeifer, Gas price spike underlines UK supply fears, online (October 2013), http://www.ft.com.

[10] S. Pfeifer, Weather and low stocks add to gas worries, online (October 2013), http://www.ft.com.

[11] H. Bevrani, Robust Power System Frequency Control, Springer, 2009.

[12] M. Misra, H. H. Yue, S. J. Qin, C. Ling, Multivariate process monitoring and fault diagnosis by multi-scale PCA, Computers & Chemical Engineering 26 (9) (2002) 1281–1293.

[13] M. Jelali, B. Huang (Eds.), Detection and Diagnosis of Stiction in Control Loops: State of the Art and Advanced Methods, Springer, 2010.

[14] C. Xia, J. Howell, N. F. Thornhill, Detecting and isolating multiple plant-wide oscillations via spectral independent component analysis, Automatica 41 (12) (2005) 2067–2075.

[15] H. Jiang, M. A. A. Shoukat Choudhury, S. L. Shah, Detection and diagnosis of plant-wide oscillations from industrial data using the spectral envelope method, Journal of Process Control 17 (2) (2007) 143–155.

[16] N. Thornhill, Finding the source of nonlinearity in a process with plant-wide oscillation, IEEE Transactions on Control Systems Technology 13 (3) (2005) 434–443.

[17] M. Bauer, J. W. Cox, M. H. Caveness, J. J. Downs, N. F. Thornhill, Finding the direction of disturbance propagation in a chemical process using transfer entropy, IEEE Transactions on Control Systems Technology 15 (1) (2007) 12–21.

[18] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Computing Surveys 41 (3) (2009) 1–58.

[19] M. Chuah, F. Fu, ECG anomaly detection via time series analysis, in: Proceedings of the ISPA International Workshops on Frontiers of High Performance Computing and Networking, Springer Berlin Heidelberg, Niagara Falls, 2007, pp. 123–135.

[20] E. Keogh, J. Lin, A. Fu, HOT SAX: Efficiently finding the most unusual time series subsequence, in: Proceedings of the Fifth IEEE International Conference on Data Mining, IEEE, Houston, 2005, pp. 226–233.

[21] N. F. Thornhill, H. Melbø, J. Wiik, Multidimensional visualization and clustering of historical process data, Industrial & Engineering Chemistry Research 45 (17) (2006) 5971–5985.

[22] S. Lhermitte, J. Verbesselt, W. W. Verstraeten, P. Coppin, A pixel based regeneration index using time series similarity and spatial context, Photogrammetric Engineering and Remote Sensing 76 (6) (2010) 673–682.

[23] D. Fabozzi, T. Van Cutsem, Assessing the proximity of time evolutions through dynamic time warping, IET Generation, Transmission & Distribution 5 (12) (2011) 1268–1276.

[24] F. Angiulli, C. Pizzuti, Fast outlier detection in high dimensional spaces, in: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, Springer Berlin Heidelberg, Helsinki, 2002, pp. 15–27.

[25] J. Zhang, H. Wang, Detecting outlying subspaces for high-dimensional data: the new task, algorithms and performance, Knowledge and Information Systems 10 (3) (2006) 333–355.

[26] E. M. Knorr, R. T. Ng, V. Tucakov, Distance-based outliers: algorithms and applications, The VLDB Journal 8 (3-4) (2000) 237–253.

[27] G. K. Palshikar, Distance-based outliers in sequences, in: Proceedings of the Second International Conference on Distributed Computing and Internet Technology, Springer Berlin Heidelberg, Bhubaneswar, 2005, pp. 547–552.

[28] V. Hautamäki, I. Kärkkäinen, P. Fränti, Outlier detection using k-nearest neighbour graph, in: Proceedings of the 17th International Conference on Pattern Recognition, IEEE, Cambridge, 2004, pp. 430–433.

[29] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, LOF: Identifying density-based local outliers, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, Dallas, 2000, pp. 93–104.

[30] D. Pokrajac, A. Lazarevic, L. J. Latecki, Incremental local outlier detection for data streams, in: Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, IEEE, Honolulu, 2007, pp. 504–515.

[31] P. J. Rousseeuw, A. M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, Inc, 2003.

[32] P. J. Huber, E. M. Ronchetti, Robust Statistics, Second Edition, John Wiley & Sons Inc., 2009.

[33] E. W. Weisstein, CRC Concise Encyclopedia of Mathematics, CRC press, 2010.

[34] S. Chandrasekhar, Stochastic problems in physics and astronomy, Reviews of Modern Physics 15 (1) (1943) 1.

[35] P. Bansal, A. Ardell, Average nearest-neighbor distances between uniformly distributed finite particles, Metallography 5 (2) (1972) 97 – 111.

[36] D. Evans, A. J. Jones, W. M. Schmidt, Asymptotic moments of near–neighbour distance distributions, Proceedings: Mathematical, Physical and Engineering Sciences 458 (2028) (2002) 2839–2849.

[37] D. A. Keim, Information visualization and visual data mining, IEEE Transactions on Visualization and Computer Graphics 8 (1) (2002) 1–8.

[38] S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, Dallas, 2000, pp. 427–438.