Expert Systems With Applications 121 (2019) 418–437

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Efficient on-line anomaly detection for ship systems in operation

Andreas Brandsæter[a,b,*], Erik Vanem[a,b], Ingrid K. Glad[b]

[a] DNV GL, Veritasveien 1, Høvik N-1363, Norway
[b] Department of Mathematics, University of Oslo, P.B.1053, Blindern, Oslo N-0316, Norway

ARTICLE INFO

Article history:
Received 6 December 2017
Revised 20 December 2018
Accepted 21 December 2018
Available online 22 December 2018

Keywords:
Anomaly detection
Condition monitoring
Maritime industry
Auto Associative Kernel Regression (AAKR)
Cluster based AAKR
Sequential Probability Ratio Test (SPRT)

ABSTRACT

We propose novel modifications to an anomaly detection methodology based on multivariate signal reconstruction followed by residuals analysis. The reconstructions are made using Auto Associative Kernel Regression (AAKR), where the query observations are compared to historical observations called memory vectors, representing normal operation. When the data set with historical observations grows large, the naive approach where all observations are used as memory vectors will lead to unacceptable large computational loads, hence a reduced set of memory vectors should be intelligently selected. The residuals between the observed and the reconstructed signals are analysed using standard Sequential Probability Ratio Tests (SPRT), where appropriate alarms are raised based on the sequential behaviour of the residuals.

The modifications we introduce include: a novel cluster based method to select memory vectors to be considered by the AAKR, which gives an extensive reduction in computation time; a generalization of the distance measure, which makes it possible to distinguish between explanatory and response variables; and a regional credibility estimation used in the residuals analysis, to let the time used to identify if a sequence of query vectors represents an anomalous state or not, depend on the amount of data situated close to or surrounding the query vector.

We demonstrate how the anomaly detection method and the proposed modifications can be successfully applied for anomaly detection on a set of imbalanced benchmark data sets, as well as on recent data from a marine diesel engine in operation.

1. Introduction

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behaviour (Chandola, Banerjee, & Kumar, 2009). In other words, anomalies can be defined as observations, or subsets of observations, which are inconsistent with the remainder of the data set (Hodge & Austin, 2004). Depending on the field of research and application, anomalies are also often referred to as outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants (Chandola et al., 2009; Hodge & Austin, 2004). Anomaly detection is related to, but distinct from noise removal (Chandola et al., 2009).

Traditionally, sensor based component control is typically rule-based. A temperature threshold might for example be predefined, forcing the system to automatically shut-down if the temperature surpasses a predefined threshold. The problem with the rule-based approach emerges when we want to analyse multiple signals, and base our decisions on the combined behaviour. To illustrate this, we can consider two signals, x_1 and x_2, where normal behaviour is located on a circle, with an anomaly in the centre of the circle (see Fig. 1). While the anomalous point can be easily identified when we analyse both signals together, it will not be detected as anomalous if we analyse the signals separately. When we want to monitor and analyse a system with many signals, the problem space grows rapidly, making it almost impossible to describe rules that cover every permutation (Flaherty, 2017). Hence, more sophisticated anomaly detection methods are needed.

An extensive number of anomaly detection methods are described in the literature and used extensively in a wide variety of applications in various industries. The available techniques comprise (Chandola et al., 2009; Kanarachos, Christopoulos, Chro-

[*] Corresponding author at: Strategic Research and Innovation, Veritasveien 1, Høvik, Norway.

E-mail addresses: andreas.brandsaeter@dnvgl.com (A. Brandsæter), erik.vanem@dnvgl.com (E. Vanem), glad@math.uio.no (I.K. Glad).

https://doi.org/10.1016/j.eswa.2018.12.040
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

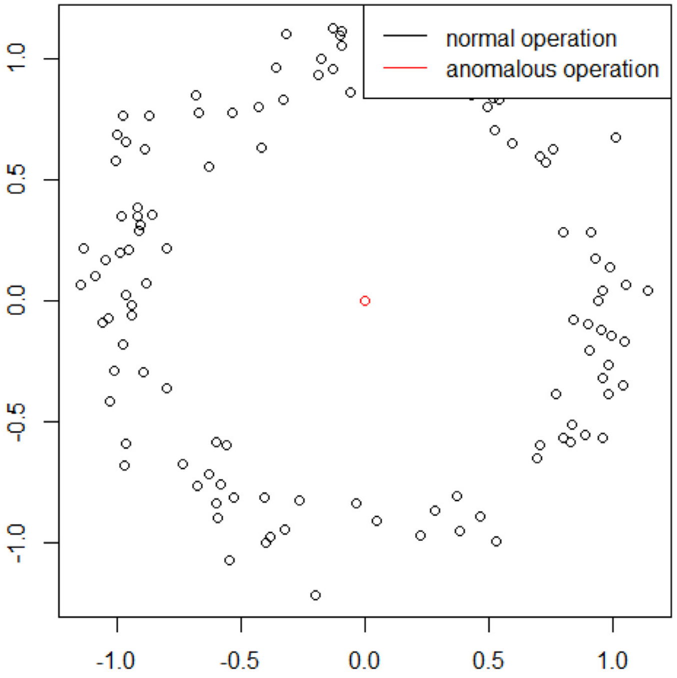


Fig. 1. Points representing normal behaviour is located on a circle. An anomaly is located in the middle.

neos, & Fitzpatrick, 2017; Olson, Judd, & Nichols, 2018; Zheng, Li, & Zhao, 2016): classification methods that are rule-based, or based on Neural Networks, Bayesian Networks or Support Vector Machines; nearest neighbour based methods, including *k* nearest neighbour and relative density; clustering based methods; and statistical and fuzzy set-based techniques, including parametric and non-parametric methods based on histograms or kernel functions.

The fundamental approaches to the problem of anomaly detection can be divided into three categories (Chandola et al., 2009; Hodge & Austin, 2004):

- *Supervised anomaly detection:* Availability of a training data set with labelled instances for normal and anomalous behaviour is assumed. Typically, predictive models are built for normal and anomalous behaviour, and unseen data are assigned to one of the classes.
- *Unsupervised anomaly detection:* Here, the training data set is not labelled, and an implicit assumption is that the normal instances are far more frequent than anomalies in the test data. If this assumption is not true then such techniques suffer from high false alarm rate.
- *Semi-supervised anomaly detection:* In semi-supervised anomaly detection, the training data only includes normal data. A typical anomaly detection approach is to build a model for the class corresponding to normal behaviour, and use the model to identify anomalies in the test data. Since the semi-supervised methods do not require labels for the anomaly class, they are more widely applicable than supervised techniques.

Our main motivation in this study is related to anomaly detection in the maritime industry. Modern ships are a highly complex systems, often equipped with thousands of sensors to monitor various features of the system. Our aim is eventually to identify anomalies and unexpected system behaviour that can represent faults in the system, but in principle, any behaviour that deviates from the behaviour represented in the training data can be discovered, not only faults.

We repeatedly refer to the maritime case study in many of the examples and demonstrations. However, the methods we envisage and the modifications we propose are widely applicable to anomaly detection problems concerning time series data.

In most industries, including the maritime industry, data from normal operating conditions are continuously collected on a large and increasing number of assets. However, comprehensive fault data are more rare, hence we pursue a semi-supervised approach, and present a kernel function based non-parametric statistical anomaly detection technique.

We use an on-line anomaly detection technique, consisting of two steps. In the first step, the observed signal is reconstructed under normal conditions. Secondly, the residuals, i.e. the difference between the observed signal and the reconstructed signal, are analysed. In this study, the signal reconstruction is performed using Auto Associative Kernel Regression (AAKR), (see Section 2.1), and the residual analysis is performed sequentially, with Sequential Probability Ratio Test (SPRT), (see Section 2.2).

One of the main drawbacks with the AAKR signal reconstruction method becomes evident when the set of historical observations grows large. Then the crude approach where all observations are used as memory vectors will lead to unacceptable large computational loads. Therefore, a reduced set of memory vectors should be intelligently selected (Hines, Garvey, & Seibert, 2008; Hines, Garvey, Seibert, & Usynin, 2008), and in this paper we suggest a novel approach to memory vector selection, where the original dataset is represented by sets surrounding a selection of clusters.

In Baraldi, Di Maio, Genini, and Zio (2015), the AAKR signal reconstruction method is compared with other popular signal reconstruction techniques, including Fuzzy Similarity (FS), and Elman Recurrent Neural Network (RNN), and capabilities and drawbacks are discussed. Hence, in this paper we will restrain to comparing the results of the modifications we propose to the crude AAKR method.

The remaining of the paper is structured as follows: The anomaly detection framework mentioned above will be briefly presented in Section 2. In Section 3, we propose three modifications of the standard framework:

A. *Cluster based memory vector selection method:* Perform a cluster analysis on the training data set, which represent normal conditions. Replace the original training data set with rectangular boxes - one for each cluster, centred at the cluster means - and define everything inside the boxes as normal condition.
B. *Modified distance measure between the query vector and the memory vectors:* Modifying the distance measure to enable the possibility of treating the variables differently based on the credibility of the signals, and distinguish between explanatory and response signals.
C. *Credibility estimation:* Regard some regions in the sample space more credible or trustworthy than others. Assume that the reconstruction of a response signal is more credible if the corresponding explanatory signals are similar to previously observed signals.

In Section 4, the performance of the proposed cluster based method is demonstrated on 14 different data sets - 13 benchmark data sets from the KEEL database (Alcalá-Fdez et al., 2011), and one data set from a marine engine in operation, and the results of the proposed cluster based method are compared to the results of the original (crude) method without memory vector selection. To further demonstrate the methodology and the proposed modifications, a more comprehensive study of the data set with the marine engine is presented in Section 5. A short discussion of the assumptions and results is presented in Section 5.8. Finally, in Section 6 some concluding remarks are offered, together with a discussion on further work.

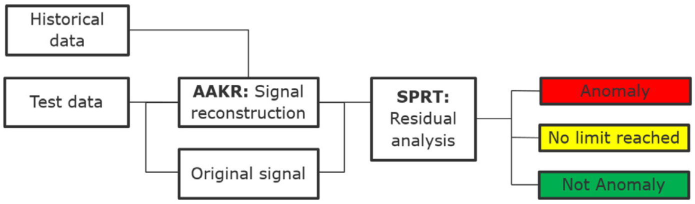


Fig. 2. The methodology can be divided into two main steps: signal reconstruction (via AAKR) and analysis of residuals (via SPRT).

The analysis is conducted in R version 3.3.3 (2017-03-06), using RStudio Version 1.0.136, on a single computer running Windows 10 Enterprise, version 1607, with Intel Core i5-6600 CPU @ 3.30 GHz processor, and 3.02GB installed RAM.

2. Standard framework for anomaly detection with AAKR and SPRT

The classical framework can be divided into two main steps: signal reconstruction and residual analysis (see Fig. 2). In particular, Auto Associative Kernel Regression (AAKR) is used for the reconstruction, and Sequential Probability Ratio Test (SPRT) is used to analyse the residuals between the reconstructed and the observed signal.

At each new time t of the on-line anomaly detection monitoring, both the reconstruction and the residuals analysis are performed in a sequential manner. In the signal reconstruction step, the values of the monitored signals are reconstructed as an estimate of the signals under normal conditions. AAKR is a data driven method where the reconstructed signal is estimated as a weighted linear combination of historical observations. The information from the current observation is used to calculate the weights. In the second step, the residuals, i.e. the difference between the observed test points (queries) and the reconstructed signals, are analysed sequentially, building evidence that the sensors report possibly anomalous behaviour.

2.1. Signal reconstruction using Auto Associative Kernel Regression (AAKR)

Many excellent descriptions of the AAKR method, both comprehensive and more brief, are given in the literature (Baraldi, Canesi, Zio, Seraoui, & Chevalier, 2011; Baraldi, Di Maio, Genini et al., 2015; Baraldi, Di Maio, Pappaglione, Zio, & Seraoui, 2012; Baraldi, Di Maio, Turati, & Zio, 2015; Di Maio, Baraldi, Zio, & Seraoui, 2013; Garvey, Garvey, Seibert, & Hines, 2007; Hines, Garvey, & Seibert, 2008; Hines, Garvey, Seibert, & Usynin, 2008). In the following we will render a basic description, following Brandsæter, Manno, Vanem, and Glad (2016).

The historical observations are collected in an $L \times J$ matrix, where L is the total number of time points of historical observations, and J is the number of sensors. If all historical observations should be taken into account by the AAKR, the reconstruction process will be very computationally expensive when the data set of historical observations grows large. Therefore, more or less intelligent selection methods (Hines, Garvey, & Seibert, 2008; Hines, Garvey, Seibert, & Usynin, 2008) are used to select some $K < L$ historical observations, or memory vectors, and collect them in a new $K \times J$ matrix $\mathbf{X}^{train}$, to be used in the reconstruction procedure.

Note that the reconstruction method does not consider time ordering, not even the sequentiality, of the observations in the training data.

At each test point t, a reconstruction of the test point $\mathbf{x}^{test}(t) = [x(t,1), \dots, x(t,J)]$ is calculated as a weighted linear combination of the observations (the rows) in the training matrix $\mathbf{X}^{train}$. The weight $\mathbf{w}$ of a row k is given by the Gaussian kernel

$$\mathbf{w}_k = \frac{1}{\sqrt{2\pi}h} e^{-\frac{\mathbf{d}_k^2}{2h^2}}, \tag{1}$$

where the parameter h is the bandwidth, and $\mathbf{d}_k$ is the distance between the J signal measurements in the observation $\mathbf{X}^{test}_{(t,)}$ and the kth observation in $\mathbf{X}^{train}$, for $k = 1, \dots, K$. Several distance functions can be used (Garvey et al., 2007), but the most common is the Euclidean norm

$$\mathbf{d}_k = \sqrt{\sum_{j=1}^{J} \left(\mathbf{X}^{test}_{(t,j)} - \mathbf{X}^{train}_{(k,j)}\right)^2}. \tag{2}$$

Finally, the reconstructed value $\hat{\mathbf{X}}^{test}_{(t,j)}$ of the jth observation $\mathbf{X}^{test}_{(t,j)}$, is given as the weighted linear combination of the rows of the training matrix, that is

$$\hat{\mathbf{X}}^{test}_{(t,j)} = \frac{\sum_{k=1}^{K} \mathbf{w}_k \cdot \mathbf{X}^{train}_{(k,j)}}{\sum_{k=1}^{K} \mathbf{w}_k}. \tag{3}$$

The methodology processes the various signals together. To avoid numerical instabilities due to possibly very different range of magnitudes in the different signals, the signal values need to be normalized. Without normalization, the effect of a deviation in one signal cannot be directly compared to the other signals. In the present work we have used the following normalization procedure, sometimes referred to as the z score normalization, encouraged by Di Maio et al. (2013). Having measured a signal $\mathbf{X}_{(t,j)}$, the normalized signal, $\tilde{\mathbf{X}}_{(t,j)}$ is given by

$$\tilde{\mathbf{X}}_{(t,j)} = \frac{\mathbf{X}_{(t,j)} - \hat{\mu}_j}{\hat{\sigma}_j}, \tag{4}$$

where

$$\hat{\mu}_j = \frac{\sum_{k=1}^{K} (\mathbf{X}^{train}_{(k,j)})}{K}, \tag{5}$$

$$\hat{\sigma}_j = \sqrt{\frac{\sum_{k=1}^{K} \left(\mathbf{X}^{train}_{(k,j)} - \hat{\mu}_j\right)^2}{K}}. \tag{6}$$

Alternative normalization procedures should also be investigated, such as the min max-normalization or the decimal scaling, see e.g. Saranya and Manikandan (2013). It is noted that in some situations the choice of normalization technique can influence the results significantly.

2.2. Residuals analysis using Sequential Probability Ratio Test (SPRT)

The residuals, i.e. the differences between the reconstructed value under normal conditions, and the observed test value, $\mathbf{R}_{(t,)} =$

$\hat{\mathbf{X}}^{test}_{(t,\cdot)} - \mathbf{X}^{test}_{(t,\cdot)}$, are analysed sequentially by the standard SPRT to determine if the system is in normal or abnormal state. The methodology will be briefly described in the following. For a more thorough description we suggest (Brandsæter et al., 2016; Cheng & Pecht, 2012; Gross & Lu, 2002; Saxena et al., 2008).

The normal state is described by a null hypothesis H_0, where each component of the residuals, $\mathbf{R}_{(t,\,j)}$, are assumed to be normally distributed with mean 0 and standard deviation σ. The anomalous state is described by an alternative hypothesis H_a, which assumes that the residuals are normally distributed with specified mean and/or standard deviation different from the null hypothesis. The SPRT is performed for each signal $j = 1, \dots, J$ independently.

Based on the residuals $\mathbf{R}_{(t,\,j)}$, an index is calculated and updated sequentially for each new observation. In order to determine the condition of the system, two threshold values, A and B, are specified and at each observation the index is compared to these lower and upper decision boundaries. There are three possible outcomes at each time step:

1. The lower limit is reached, in which the null hypothesis is accepted (normal state), and the test statistic is reset.
2. The upper limit is reached, in which the null hypothesis is rejected (anomalous state), and the test statistic is reset.
3. No limit is reached, in which case the amount of information is not sufficient to make a conclusion.

For each sensor signal j, the analysis is performed on the sequence of residuals $\mathbf{r}_{(i_1,j)}, \dots, \mathbf{r}_{(i_n,j)}$. When either of the limits are reached (outcome 1 and 2), the sequence is reset to zero. If no limits are reached (outcome 3), the sequence is extended with the new residual.

The SPRT index is given as the natural logarithm of the likelihood ratio L_a, given by

$$L_a = \frac{\text{prob of } \mathbf{r}_{(i_1,j)}, \dots, \mathbf{r}_{(i_n,j)} \text{ given } H_a}{\text{prob of } \mathbf{r}_{(i_1,j)}, \dots, \mathbf{r}_{(i_n,j)} \text{ given } H_0} = \prod_{i=i_1}^{i_n} \frac{f_a(\mathbf{r}_{(i,j)})}{f_0(\mathbf{r}_{(i,j)})},$$

where $f(\cdot)$ is the corresponding normal density. Note that this construction is based on an assumption of independence among the residuals.

We consider two alternative hypotheses, i.e. deviations in either direction of the mean, leading to the following indices, for each sensor j

$$SPRT_1 = \frac{M}{\sigma^2} \sum_{i=1}^{n} \left(\mathbf{r}_i - \frac{M}{2} \right) \tag{7}$$

$$SPRT_2 = \frac{M}{\sigma^2} \sum_{i=1}^{n} \left(-\mathbf{r}_i - \frac{M}{2} \right) \tag{8}$$

The standard deviation, σ, is computed from the training data. M is the mean value of the alternative hypothesis, which is decided by the user. M is usually chosen to be several times larger than σ (Cheng & Pecht, 2012).

2.3. Limitations associated with the standard framework

There are some well-known challenges and limitations related to the anomaly detection framework presented above.

An important challenge relates to the efficiency of the AAKR method. When the data set of historical observations grows large, the signal reconstruction procedure becomes very computationally costly (Michau, Palme, & Fink, 2017). To encounter this, various memory vector selection techniques are used (Hines, Garvey, & Seibert, 2008; Hines, Garvey, Seibert, & Usynin, 2008). In this paper, we present a novel cluster based memory vector selection technique, see Section 3.1.

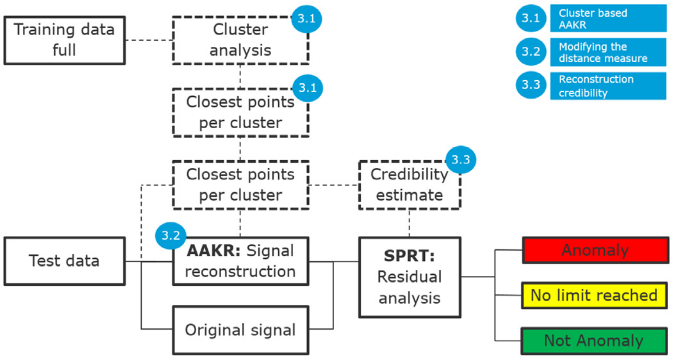


Fig. 3. The modified anomaly detection framework.

When the relative importance of the various signals is known and understood, for example based on physical meaning or by subject matter expert's experience, it should be possible to incorporate this information in the model. We propose to impose the relative importance on the AAKR model by changing the distance measure, see Section 3.2. The proposed generalization of distance measure provides the possibility to distinguish between explanatory and response signals. This also makes it more natural to compare the reconstructions produced with AAKR, with reconstructions based on other regression methods.

With the standard framework, all regions in the sample space are considered equally credible. We suggest to assume that the reconstruction of a response signal is more credible if the corresponding explanatory signals are similar to previously observed signals. In Section 3.3, we describe one possible approach to encounter this.

Other challenges associated with the anomaly detection framework, such as challenges related to time dependency and the need for representative training data, as well as problems associated with evaluating the accuracy when labelled data is lacking, are of general nature and is not addressed here.

3. Proposed modifications

In the following, we propose three novel modifications aiming to improve the anomaly detection framework as presented above, and to address associated challenges. A sketch of the suggested modified anomaly detection framework is shown in Fig. 3, with the new boxes marked with dashed borders.

3.1. Cluster based memory vector selection for AAKR

In the maritime industry, as in many other industries, the amount of available and potentially interesting data is large and growing. In the AAKR method, the distance between the observed query vector and each of the memory vectors have to be calculated, as well as the weights associated with each memory vector and eventually the weighted linear combination of all the memory vectors. Consequently, if we use a naive approach, and let all training data points be represented in the set of memory vectors, the algorithm will be very computationally costly for large training data sets. Hence, intelligent memory vector selection methods are needed.

Several memory vector selection methods exist, including vector ordering, min-max selection, combination of vector ordering and min-max selection (Boechat, Moreno, & Haramura, 2012; Coble, Humberstone, & Hines, 2010; Hines, Garvey, & Seibert, 2008). The methods all strive to adequately represent the operating conditions expected in future fault free operations. If variants of normal operating conditions, such as changes in weather, seasonal variations, are not included in the memory vectors, no confidence can be given to predictions of the model and the memory matrix must either be appended or replaced with new data (Boechat et al., 2012; Hines & Garvey, 2006).

In our experience, a ship's operation pattern can be divided into relatively few sub-operations, such as for example harbour, transit (in a few different speeds) and manoeuvring. This relatively simple operation pattern is typically also reflected in related systems such as the machinery. Hence, we propose to use a memory selection method based on clustering, which exploits this property of the data. Our first experiences with this method was presented in Brandsæter, Vanem, and Glad (2017). Here we elaborate and systematically investigate the methodology.

3.1.1. Clustering for anomaly detection

Several clustering based anomaly detection techniques have been developed (see e.g. Chandola et al., 2009), and various categories of clustering methods for anomaly detection are suggested in the literature. One common approach is to cluster the data first, and then classify the data according to one of the following assumptions:

1. Normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster.
2. Normal data instances lie close to their closest cluster centroid, while anomalies are far away from their closest cluster centroid.
3. Normal data instances belong to large and dense clusters, while anomalies either belong to small or sparse clusters.

The approach we propose in this paper, is somewhat inspired by both 1 and 2 above. In brief, we suggest to first cluster all historical observations. Secondly, the regions surrounding the cluster centroids are identified. The clustering and identification of surrounding sets are performed off-line, prior to operation. Then, during operation, for each new query point, one memory vector from each of the surrounding sets are selected such that the distance between the query point and the representative of the surrounding set is minimized. Finally, the selected memory vectors are used in the AAKR reconstruction procedure. In this way, a new set of memory vectors is selected for each query vector.

3.1.2. Prediction based on representatives from the surrounding sets

After the clustering process is executed on the training data, and the surrounding sets are identified, the reconstruction of the test data can take place. The reconstruction of the query vector, $\hat{\mathbf{X}}^{test}_{(t,)}$, is produced using AAKR as described in Section 2.1, but now the training data $\mathbf{X}^{train}$ which contains selected or all historical observations, is replaced by a matrix $\mathbf{X}^{closest}$ containing the unique closest point per cluster, i.e. the ith row of $\mathbf{X}^{closest}$ is given by

$$\mathbf{p}^* = \arg\min_{\mathbf{p}\in O_i} \sum_{j=1}^{J} \left(\mathbf{p}_j - \mathbf{X}^{test}_{(t,j)}\right)^2, \tag{9}$$

where O_i is the surrounding set of cluster i. Uniqueness follows in the Euclidean space for surrounding sets that are closed and convex (Dattorro, 2010).

Hence, if a test point $\mathbf{X}^{test}$ lies inside a surrounding set O_i, the distance between the test point and the closest point in that surrounding set is 0. If on the other hand, the test point lies outside the surrounding set, the distance between the test point and the closest point in that surrounding set is strictly greater than 0, and the closest point will be on the surrounding set's border. This is illustrated in Fig. 4 a simplistic example in 2 dimensions.

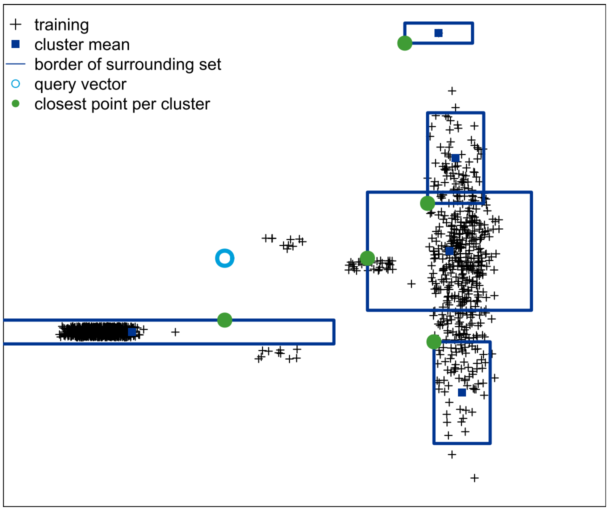


Fig. 4. Illustration of the surrounding hyperrectangles, and their unique closest points to a query vector.

3.1.3. Surrounding sets

One candidate for the surrounding set of a cluster is the convex hull of its members (see left hand plots of Fig. 5). Another suggestion is to use an ellipsoid, centred at the cluster mean with shape parameters based on the standard deviation of the cluster members, for each sensor signal (see the centre plots of Fig. 5). Furthermore, the clustering can be performed using clustering techniques such as Density-based spatial clustering of applications with noise (DBSCAN) (Ester, Kriegel, Sander, Xu et al., 1996), CLARA (Ng & Han, 1994) and CLARANS (Ng & Han, 2002). Such techniques enable identification of clusters with arbitrary shape, that are non-linearly separable, which cannot be adequately clustered with k-means or Gaussian Mixture EM clustering (Ester et al., 1996).

However, for simplicity, and due to the computational cost of calculating the distance between a query vector and the boundary of more complex shapes (Cameron, 1997; Jarvis, 1973), we chose to use axis-aligned hyperrectangles/boxes.

If the data set is in $\mathbb{R}^2$, it is possible to find the set of k axis-aligned rectangles of minimum area that covers the points in the data set using optimization techniques such as for example mixed integer and linear programming (see Ahn et al.; Park & Kim). But to our knowledge, no efficient method exists that applies to large data sets in high dimensions.

Fortunately, we do not need to determine the optimal set of hyperrectangles/boxes and can be satisfied with a good selection. Hence, we will explore the use of well-known clustering techniques to cluster the data. When the data set is divided into clusters the size and position of the hyperrectangles are determined in one of the following ways:

1. *Centred:* The boxes are centred at the mean value of the members of the cluster (in each dimension), where the distance between cluster centroids and boundary are given by the standard deviation.
2. *Enclosed:* The boxes are placed such that they cover all points assigned to each specific cluster.

In addition, a rectangle scaling factor γ is used to increase or decrease the size of the surrounding set.

Four different surrounding sets for a simplistic two dimensional example are illustrated in Fig. 5: convex hulls, ellipses, rectangles centred at the cluster mean and rectangles placed such that they cover all points assigned to each specific cluster. In the upper and lower plots, the number of clusters is set to 7 and 15 respectively.

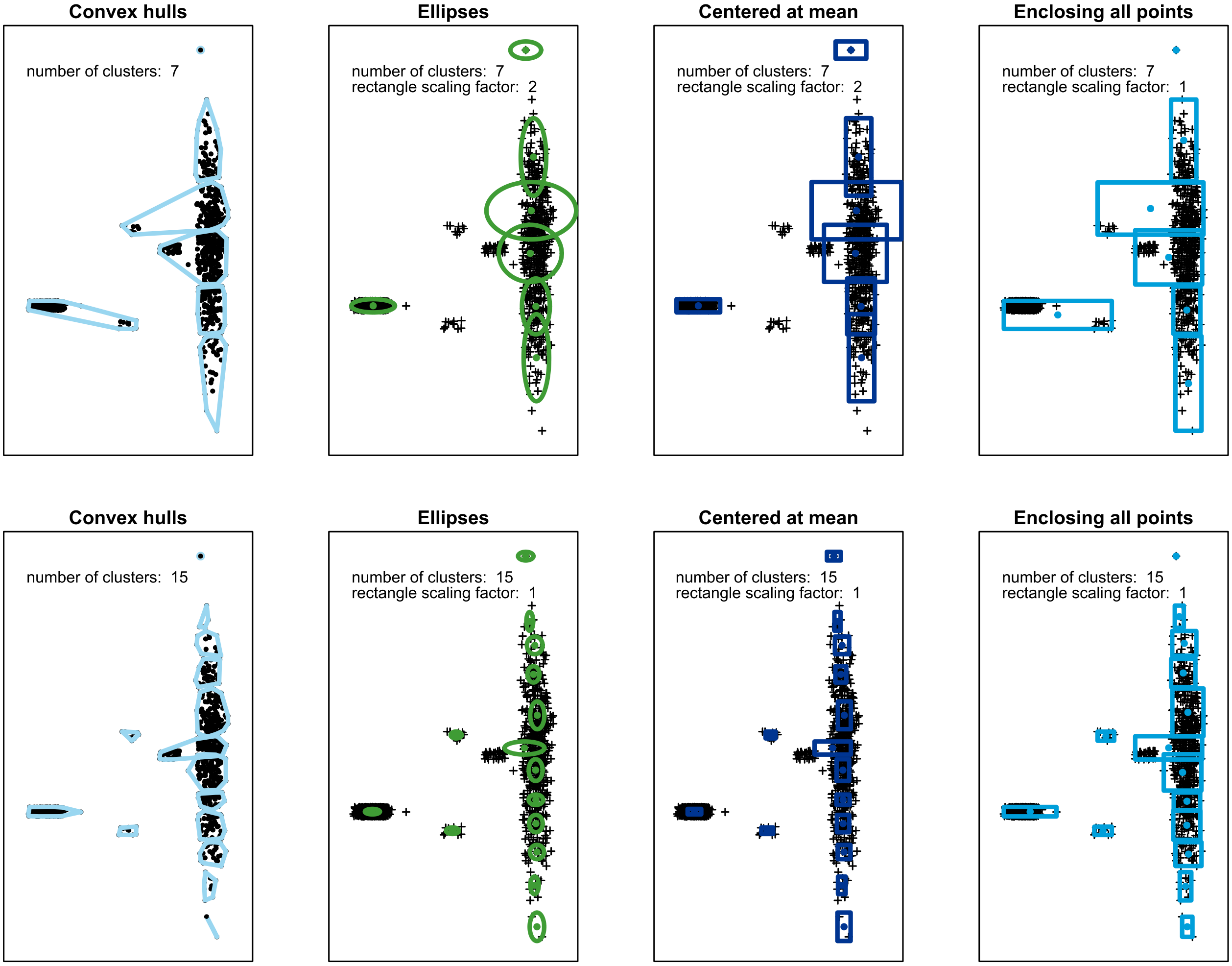


Fig. 5. Illustration of different surrounding sets, with 7 and 15 clusters (upper and lower).

Hierarchical clustering has been used to find the cluster centroids, with the complete linkage criterion, see (Section 3.1.4). The rectangle scaling factor, γ, which adjusts the shape and size of the ellipses and the rectangles is set to 2. In the lower plot, the rectangle scaling factor is set to 1, and the number of clusters is increased to 15.

3.1.4. Clustering techniques

The following clustering techniques are explored (See e.g. Cord & Cunningham, 2008; Friedman, Hastie, & Tibshirani, 2009):

1. *Standard k-means clustering:* In the initialization, k cluster centroids are chosen randomly. Then for each iteration, the observations are reassigned to the closest cluster centroid, before the cluster centroids are updated to reflect the new cluster mean. The iterations continue until the cluster centroids no longer change from one iteration to another.
2. *Agglomerative hierarchical clustering:* Each observation starts in its own cluster, and the pair of clusters with minimum distance, according to a linkage criterion, are merged. To calculate the distance between two points, we use Euclidean distance. We explore two different linkage criteria:
   - *Single:* Where the distance between two clusters A and B, is given as $\mathbf{min}\{d(a,\ b)\colon\ a\in A, b\in B\}$, where a and b are observations assigned to cluster A and B respectively.
   - *Complete:* Where the distance between two clusters A and B, is given as $\mathbf{max}\{d(a,\ b)\colon\ a\in A, b\in B\}$, where a and b are observations assigned to cluster A and B respectively.

3.1.5. Choosing the number of clusters

Unlike in classification tasks, cluster analysis procedures will generally be unable to refer to predefined class labels when employed in real-world applications. Consequently, there is usually no clear definition of what constitutes a correct clustering for a given data set (Cord & Cunningham, 2008). However, since the final goal of our analysis in this study is anomaly detection, which is a classification task, we can claim that the best number of clusters is the one which provides the most accurate anomaly detection. However in practice, this approach can only be utilized through cross validation, on a training set with labelled anomalies.

For standard clustering analysis, not involving classification, a wide variety of validation methods have been proposed (For an overview, see for example Cord & Cunningham, 2008; Friedman et al., 2009; Guha & Mishra, 2016; Wilks, 2011). Cord and Cunningham (2008) organize them into three distinct categories:

1. *Internal validation:* Compare clustering solutions based on the goodness-of-fit between each clustering and the raw data on which the solutions were generated.

2. *External validation:* Assess the agreement between the output of a clustering algorithm and a predefined reference partition that is unavailable during the clustering process.
3. *Stability-based validation:* Evaluate the suitability of a given clustering model by examining the consistency of solutions generated by the model over multiple trials.

In this study, we concentrate on internal validation, which means that we compare the various combinations of clustering methods and number of clusters, based on the goodness-of-fit according to some evaluation function. In addition to well-known methods such as the elbow, silhouette and gap statistic methods, there are more than thirty other indices and methods that have been published for identifying the optimal number of clusters (Charrad, Ghazzali, Boiteau, Niknafs, & Charrad, 2014). We can for example use the NbClust package (Charrad et al., 2014) in R, which provides 30 of the most popular indices for determining the number of clusters for a given data set. The number of clusters is chosen according to the majority rule. However, to allow easy comparison between the various clustering methods, and to illustrate the effect of using different number of clusters, we use a fixed array of number of clusters in the demonstration in Section 4.

As described above, choosing the optimal number of clusters is often ambiguous. Fortunately however, the cluster based AAKR method proposed in this paper, does not require that the optimal number of clusters is found. The motivation behind the clustering is to increase the computational speed. If we increase the number of clusters, we know that we should retain more of the information in the original data. But the number of clusters to use is a trade-off between computational speed and accuracy. With too few clusters, a lot of the information in the data is lost, but with sufficiently many clusters, the assumption is that we can approximate the information in the full training data with sets surrounding the clusters. The aim is to find the right balance between model performance and model run time (Hines, Garvey, & Seibert, 2008). If the model performance turns out to be poor, more clusters should be included to expand the memory matrix coverage of the operational region (Coble et al., 2010).

That being said, we see that in some of the cases presented in Section 4, the results show that the cluster based AAKR outperforms the crude method, where no clustering has been performed. We believe this is due to insufficient training data, and do not regard this performance improvement significant.

3.2. Modified distance measure to distinguish explanatory and response signals

When reconstructions are produced using AAKR, usually all signals are weighted equally when the distance between the query vector and the memory vectors is calculated. In Baraldi, Di Maio, Turati et al. (2015), a new procedure for determining the distance is proposed, where the data are projected into a new signal space, by defining a penalty vector which reduces the contribution of signals affected by malfunctioning. The procedure is motivated by the conjecture that faults or malfunctions causing variations of a small number of signals are more frequent than those causing variations of a large number of signals.

In this paper, we propose to modify the distance calculation, in a fashion inspired by Baraldi, Di Maio, Turati et al. (2015), such that the contribution of the various signals can be weighted differently. Instead of the standard Euclidean norm (see Eq. (2)), we propose to use a weighted version by multiplying the difference in each direction with a penalty vector which we refer to as the distance scaling vector $\mathbf{s} = [s_1, \ldots, s_J]$. This gives the following distance measure

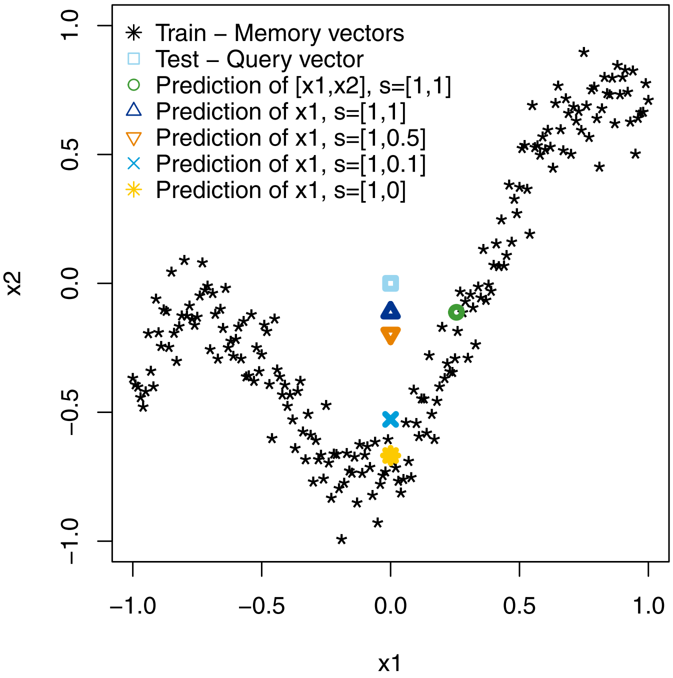


Fig. 6. Illustrating the usage of the modified distance measure, with different distance scaling vectors **s**. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$\mathbf{d}_k^{mod} = \sqrt{\sum_{j=1}^{J} \{ \left[(\mathbf{X}_{(t,j)}^{test} - \mathbf{X}_{(k,j)}^{train} \right] \cdot \mathbf{s}_j \}^2}. \tag{10}$$

If all elements of **s** are equal to 1, the classical distance measure is used. Note that if one of the signals is completely disregarded, i.e. the weight is set to 0, and the weights of the other signals are not changed, then the AAKR reconstruction resembles the traditional Nadaraya–Watson estimator, where the signal with 0 weight is the response variable, and the remaining signals are the explanatory variables. This choice of **s**, also makes comparisons to other regression methods more natural.

This generalization of the AAKR method can be particularly useful when we are not interested in finding anomalies in all the sensor signals, such as sensors measuring environmental conditions. For example, if our aim is to detect anomalies that could be caused by or lead to engine failure, we might find it uninteresting to search for anomalies in the outside air temperature sensor. As long as there is nothing wrong with the sensor, there is obviously nothing wrong with the air temperature, and we are not interested in alarms regarding this. At the same time, this sensor signal could be important in explaining the behaviour in other signals, such as engine temperature or bearing temperature. Hence, we do want to be able to include it in the analysis as an explanatory variable.

In Fig. 6 the usage of the modified distance measure is illustrated with a simplistic example in two dimensions. The black coloured stars are the training data (also referred to as memory vectors), and the light blue coloured square is a query vector (also referred to as test data), located at $[x_1, x_2] = [0, 0]$. The AAKR method with the standard Euclidean distance measure would reconstruct the signal at [0.43,−0.24], as shown by the green circle. If signal x_1 measures an environmental parameter, such as for example outside temperature or wind speed, and we assume that the sensor recordings are without faults, we are not interested in residuals in this dimension. Hence, we would regard signal x_1 as an explanatory variable, and place the reconstruction at the query vector, in this dimension. This is represented by the dark blue triangle. If we reduce the second entry of the distance scaling vector **s**, we reduce the contribution of observations that are near to the query point in the x_2 direction, and far away in the x_1 direction. The orange triangle shows the reconstructions produced with distance scaling vector **s** equal to [1,0.5], while the blue cross, and the yellow star shows the reconstructions produced using distance scaling vector [1,0.1] and [1,0] respectively.

In many real-life applications, the choice of explanatory and response variables is determined by the subject matter experts. Often, it is natural to let **s** take values 0 or 1, but other values are also acceptable. The distance scaling vector can be chosen to achieve acceptable levels of expected detection delay (EDD) and average run length (ARL), as described and demonstrated in Section 5.

3.3. Reconstruction credibility

As the training data is not evenly distributed in the data space, we propose to regard reconstructions from some regions of the sample space more credible or trustworthy than others. The idea is that we should have more confidence in our reconstructions when the query vector is close to, or at least not too far away from, the historical observations for the subset of the signals which we can treat as explanatory variables, such as environmental conditions or similar.

If reconstructions are made using AAKR with the cluster based memory vector selection method presented in Section 3.1, the number of members of a nearby cluster can also be taken into consideration when assessing the credibility of a reconstruction. One can argue that a high number should lead to higher confidence.

To illustrate the idea, we look at the simplistic example in 2 dimensions, shown in the upper plot of Fig. 7. The signal on the horizontal axis, x_1, can for example represent an environmental variable such as wind speed and we decide to treat this as an explanatory variable. Furthermore, the vertical axis, x_2, can for example represent the bearing temperature, and we decide to treat this as a response variable. Now, if we observe a value $[x_1, x_2] = [-0.75, 1.00]$ (see the leftmost red point in Fig. 7), we will be confident that this is an anomaly, since we have many historical observations of x_1 in the area around -0.75, and no corresponding values of x_2 near 1.00. However, for $[x_1, x_2] = [-0.25, 1.00]$ (rightmost red point) we have very few historical observations, hence our confidence in the reconstructions in this area is decreased.

A credibility estimate can be taken into account when the residuals are analysed in the Sequential Probability Ratio Test (SPRT). We suggest to multiply the credibility estimate with the SPRT index (see Eqs. (7) and (8)). This enables the anomaly detection framework to reach a conclusion faster when our confidence in the reconstruction is high, and use more time when our confidence is low. It should be noted, however, that the statistical properties of the SPRT will change.

3.3.1. Suggested formula for credibility estimate calculation

Different estimates can be used to calculate the credibility estimates, and we believe that different estimates should be used in different applications and cases. In the case presented here, we have used the following credibility estimate, ψ, of a query vector $\mathbf{X}^{test}_{(t,)}$,

$$\psi = 1 - \frac{1}{1 + \log(\eta^{\kappa} + 1)} \tag{11}$$

where η denotes the sum of the number of points in the surrounding sets which are close to $\mathbf{X}^{test}_{(t,)}$. A surrounding set is regarded as close if the distance between the point and the cluster centre is less than a predefined parameter λ. We experiment with different values for λ, and in the following section we show results using the following values: inf, 0.5, 0.25, 0.1, 0.05 and 0.01. When λ is infinite, all data points are regarded as close, and the credibility estimate will be constant throughout the data set. A parameter, κ, is set to control the importance of the number of points. Here, for simplicity, we fix κ to 0.1.

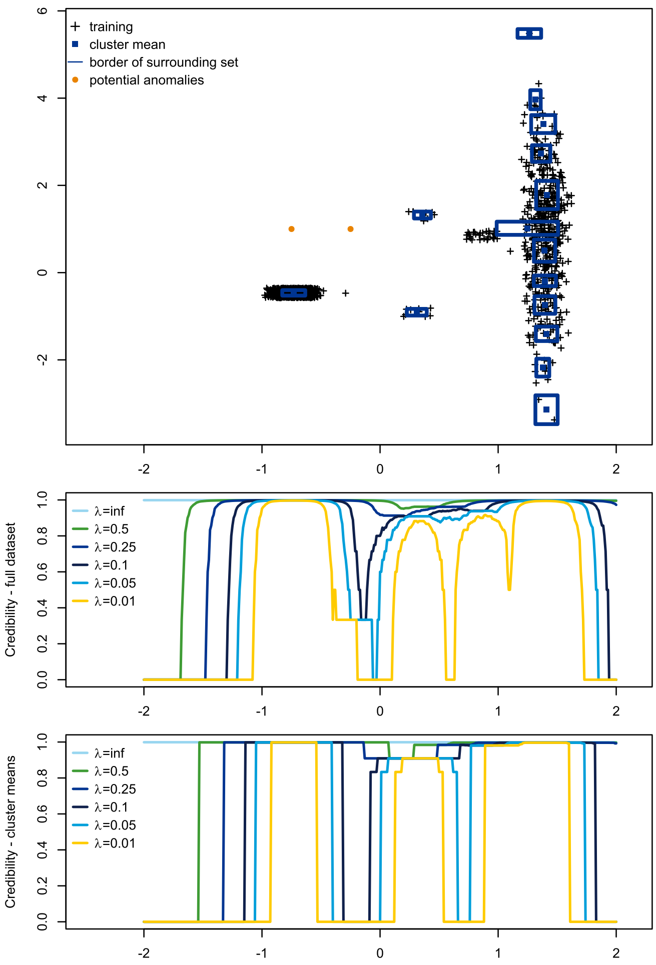


Fig. 7. The upper plot shows a simplistic data set, in two dimensions. In the two lower plots the credibility estimate is calculated for points along the horizontal axis, with different bandwidths. In the middle plot, the distances to all historical observations has been calculated, while the estimates in the lower plot are based on the distance to the unique closest point per cluster and the number of cluster members in that cluster. The number of clusters used in this figure is 15. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We see that the credibility estimate in Eq. (11) requires that the distances between $\mathbf{X}^{test}_{(t,)}$ and all the historical observations are calculated. To avoid this, we replace the full training data set with the clusters as explained in the earlier section. Also the number of points in each cluster is taken into consideration. Hence, the credibility is given by Eq. (11) where η is substituted by $\tilde{\eta}$, the sum of cluster members in clusters with nearby centres, i.e. the distance is less than a specified bandwidth.

The lines in the middle and lower plot of Fig. 7 show the proposed credibility estimates, obtained with different values of λ. The estimates in the middle plot are based on the full training data set, and the estimates in the lower plot are based on the 15 clusters and their surrounding data sets.

Table 1
Data sets used in the analysis.

Data set no.	Data set name	Imbalance ratio	No. of features	No. of training samples
1	vehicle0	3.23	18	428
2	yeast6	53.89	8	963
3	ecoli-0-1-3-7_vs_2-6	14.50	6	186
4	glass5	6.89	9	142
5	shuttle-c0-vs-c4	3.99	9	1218
6	dermatology-6	13.88	32	226
7	shuttle-6_vs_2-3	18.00	9	147
8	winequality-red-4	24.33	11	1034
9	poker-9_vs_7	12.50	10	160
10	yeast1	2.89	8	687
11	segment0	5.99	18	1319
12	vehicle2	3.23	18	409
13	vehicle3	3.04	18	415
14	engine1	1.50	5	10,000[a]

[a] Data set 14 originally includes 175,558 training samples. Due to this high number, computing the results of the crude methods is impractical. Hence, we sample 10,000 training samples without replacement, and use the result of this as an approximation of the crude method.

4. Demonstration on benchmark data sets

In this section we demonstrate the cluster based AAKR method on multiple imbalanced data sets. We present results using different clustering techniques and surrounding sets (see Section 3.1.3), and compare them to the results obtained with the crude AAKR method.

4.1. Data sets

We use 13 imbalanced data sets from the KEEL database (Alcalá-Fdez et al., 2011) (See Table 1). The rows in the data sets are pre-labelled, such that all anomalies are known, and we assume that all datapoints that are not marked as anomalies, represent normal behaviour.

The imbalanced data sets we envisage here, are data sets originated from data sets of multiple classes, where one (or more) of the classes are labelled as anomalous. For example, the imbalanced data set *yeast6* is based on the classification data set *yeast*, which contains information about a set of yeast cells, for predicting the cellular localization sites of proteins. In the classification data set, each instance is classified in 10 different localizations. In the imbalanced version, *yeast6*, the positive examples consist of class EXC and the negative examples consist of the other 9 classes. See Appendix A for a description of the other data sets.

We train on 2/3 of the data, and test on the remaining 1/3. Rows with anomalies occurring in the fraction of the data set used for training are removed.

In addition to the benchmark data sets from the KEEL database, we include another imbalanced data set from a marine engine in operation. The data set originally includes 175,558 rows. Due to the high number of rows, computing the results of the crude methods is impractical. Hence, we sample 10,000 rows without replacement, and use the result of this as an approximation of the crude method. A thorough description of this data set, together with a comprehensive analysis, is provided in Section 5.

The data sets represent various real world applications. In this section, we do not take into account any possible knowledge of the real application, and all columns of the data set are treated as equally important for detecting anomalous behaviour.

4.2. Algorithms

We present results based on the combinations of clustering algorithms and surrounding sets as presented in Table 2. The k-means clustering is performed with the *kmeans* implementation in the *stats* package in R (R Core Team, 2017), with the Lloyd algorithm (Lloyd, 1982). For hierarchical clustering we use the *hclust* implementation, also from the *stats* package, with the following two linkage criteria: single and complete.

Even for the largest data set, the clustering with k-means is performed in less than a second, hence we will not report the time to perform the clustering. For the *engine*1 data set, with 175,558 rows, the hierarchical clustering method cannot be performed due to memory restrictions. It requires that the dissimilarity structure (as produced by the *dist* function in R) is provided, which needs allocation of more than 100GBs memory.

4.3. Simple threshold based residual analysis

Many of the data sets considered in this section are not time dependent, and many of the anomalies occur alone, i.e. the observation imminently before and after are not anomalous. Due to this, we will not use the Sequential Probability Ratio Test (SPRT) when comparing the methods here. A comprehensive demonstration of SPRT will be provided in the maritime case study in Section 5. Here, we will restrain to a simple threshold method when we analyse the residuals. Again to ease the comparison between the methods, we adjust the threshold limit for each feature with a parameter τ, which controls the false alarm rate.

Furthermore, for the data sets we investigate, we have no knowledge about which signals are causing the anomaly, hence we do not distinguish this here. If an alarm is triggered in one of the signals, we consider all signals anomalous at this row/time instance.

The threshold limits obtained using this procedure should be similar to the limits we can obtain with cross validation on a training set, assuming we have known anomalies present in the training set.

4.4. Results

We present results using a range of different number of clusters, k, and a range of 50 different threshold values, τ, between 0.7 and 1. In the following, we highlight a selection of the results. The full table of results can be found in the supplementary material.

4.4.1. Decreased computation time for the cluster based methods

The main goal of the proposed cluster based method is to decrease the computation time of the different methods, and at the same time keeping the performance at an acceptable level. Fig. 8 shows savings in prediction time relative to the crude method

Table 2
Combinations of clustering algorithms and surrounding sets in the presented results.

	Clustering algorithm	Surrounding set
1	**crude**, no clustering	**points**, every point is represented
2	**k-means**, Lloyd's algorithm	**points**, centred at mean with $\gamma = 0$, i.e. every cluster is represented with a single point
3	**k-means**, Lloyd's algorithm	**centred**, centred at mean with $\gamma = 1$, i.e. every cluster is represented with a box centred at the mean of the cluster members, with size based on the standard deviation
4	**k-means**, Lloyd's algorithm	**enclosed**, every cluster is represented with a box which encloses the cluster members
5	**hierarchical**, **complete** linkage criteria	**enclosed**, every cluster is represented with a box which encloses the cluster members
6	**hierarchical**, **single** linkage criteria	**enclosed**, every cluster is represented with a box which encloses the cluster members

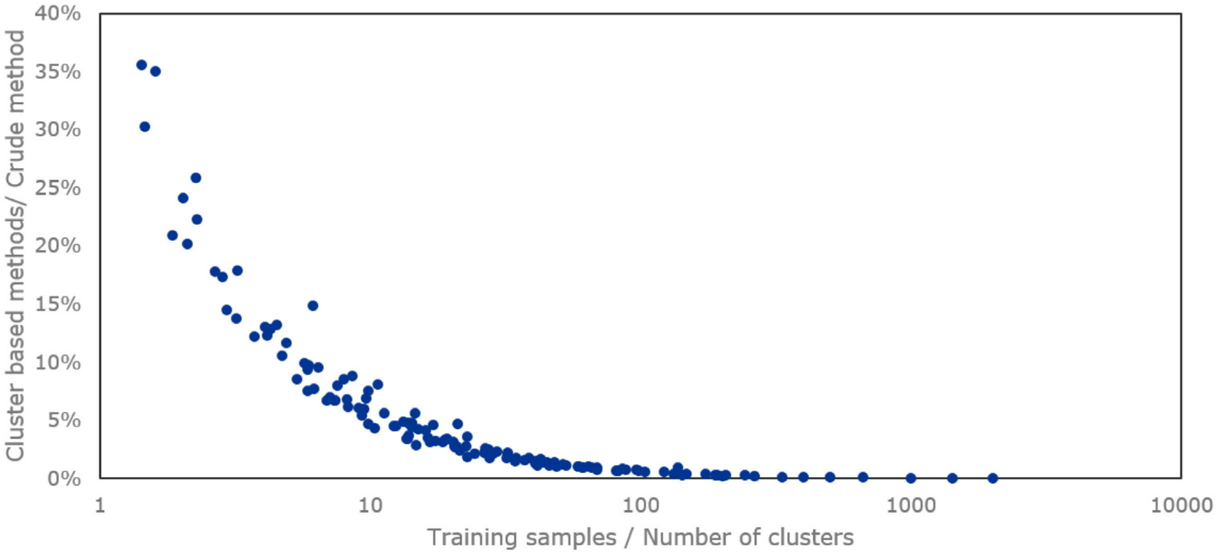


Fig. 8. Decreased computation time per prediction: The vertical axis of the figure shows the maximum computation time, when using the cluster based methods, relative to the computation time when the crude method is used. The horizontal axis represents the number of samples in the training divided by the number of clusters.

achieved with the proposed methods. The horizontal axis in the figure shows the number samples in the original training set divided by the number of clusters. As expected, as this ratio increases, i.e. when we have fewer clusters than training samples, we achieve greater time savings.

4.4.2. Comparing performance

When comparing the different methods ability to classify the anomalies, we have to balance the number of:

- True Positives (TP) - anomalous instance which is correctly identified as anomalous,
- False Positives (FP) - normal instances which are incorrectly identified as anomalous,
- False Negatives (FN) - anomalous instance which is incorrectly identified as normal
- True Negatives (TN) - normal instances which are correctly identified as normal

In this analysis, it is often useful to examine the sensitivity, which is also called the True Positive Rate. It is a measure of the probability of predicting that an instance is anomalous given that the true state is anomalous (Friedman et al., 2009). The True Positive Rate has the following expression

$$TPR = \frac{TP}{TP + FN}. \tag{12}$$

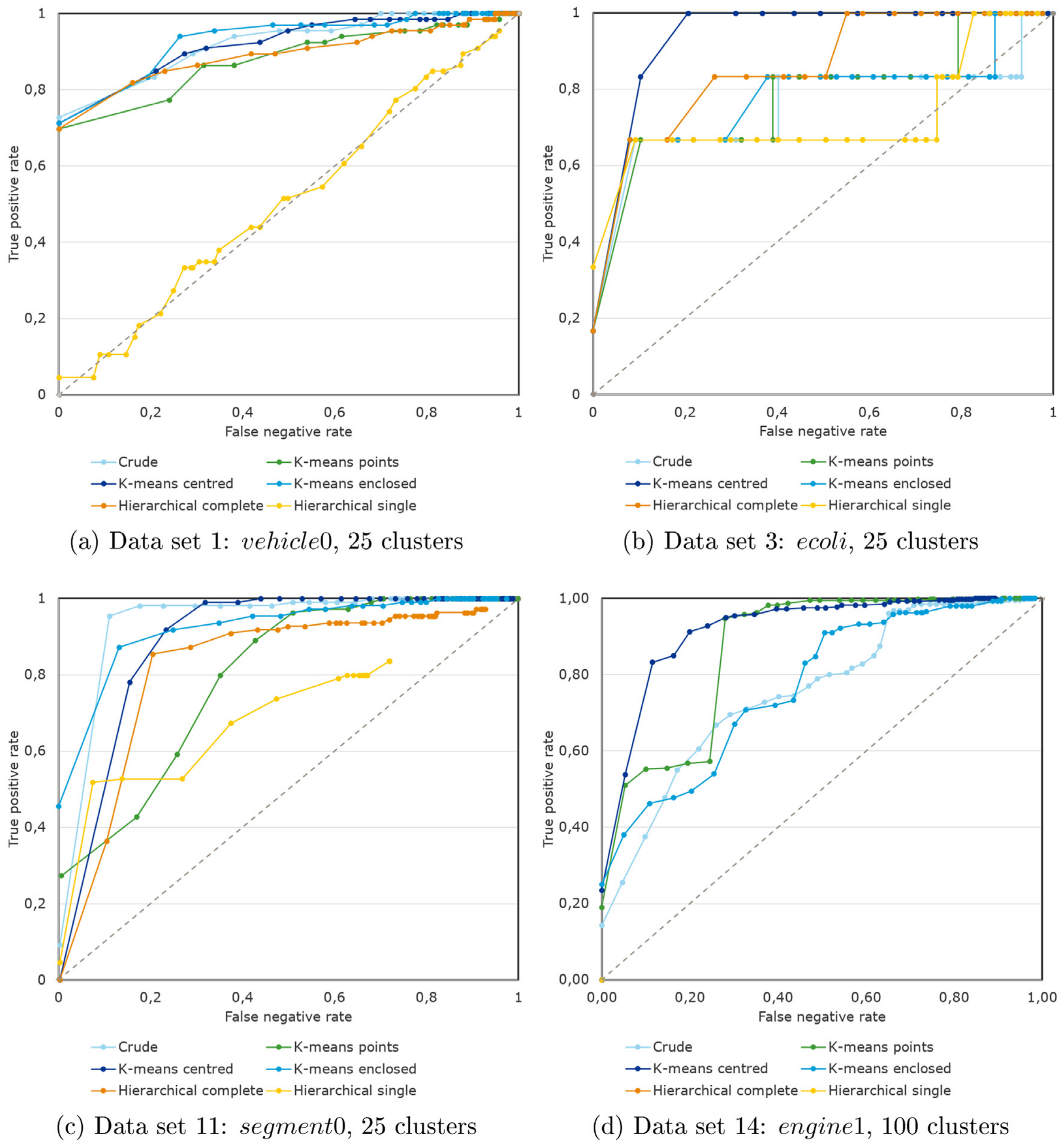
Another useful measure is the specificity, which is the probability of predicting that an instance is normal (non-anomalous) given that the true state is normal (non-anomalous). This information can also be presented as the False Positive Rate, which is given as 1 minus the specificity, that is:

$$FPR = \frac{FP}{FP + TN} = 1 - \text{specificity} \tag{13}$$

The TPR and FPR are often presented in a receiver operating characteristics (ROC) graph, which is a scatterplot with the TPR on the vertical axis, and the FPR on the vertical axis. The ROC graphs have properties that make them especially useful for domains with skewed class distribution and unequal classification error costs, which is important for cost-sensitive learning and learning in the presence of imbalanced classes (Fawcett, 2006).

The ROC graphs of four selected data sets are shown in Fig. 9. We find the most favourable results, of a ROC graph, in the upper left corner, where the FNR is low at the same time as the TPR is high. Similarly, the least favourable results are found in the lower right corner.

From Fig. 9, we observe that the different methods' performance is quite similar, except for the hierarchical clustering method with the single linkage criterion, which is clearly outperformed by the other methods especially on the *vehicle0* and *semgnet0* data sets. On the *engine1* data set, the hierarchical methods are not used due to the computational burden of performing the clustering.



(a) Data set 1: *vehicle*0, 25 clusters

(b) Data set 3: *ecoli*, 25 clusters

(c) Data set 11: *segment*0, 25 clusters

(d) Data set 14: *engine*1, 100 clusters

Fig. 9. The ROC graph for four selected data sets. Results are shown for 50 threshold values τ between 0.7 and 1. Straight lines are drawn between the points for increased readability.

In model selection, the area under the ROC curve is a popular measure, where the model with the highest area under the ROC curve will be selected. The area under the ROC-curve for the 14 data sets is provided in Table 3.

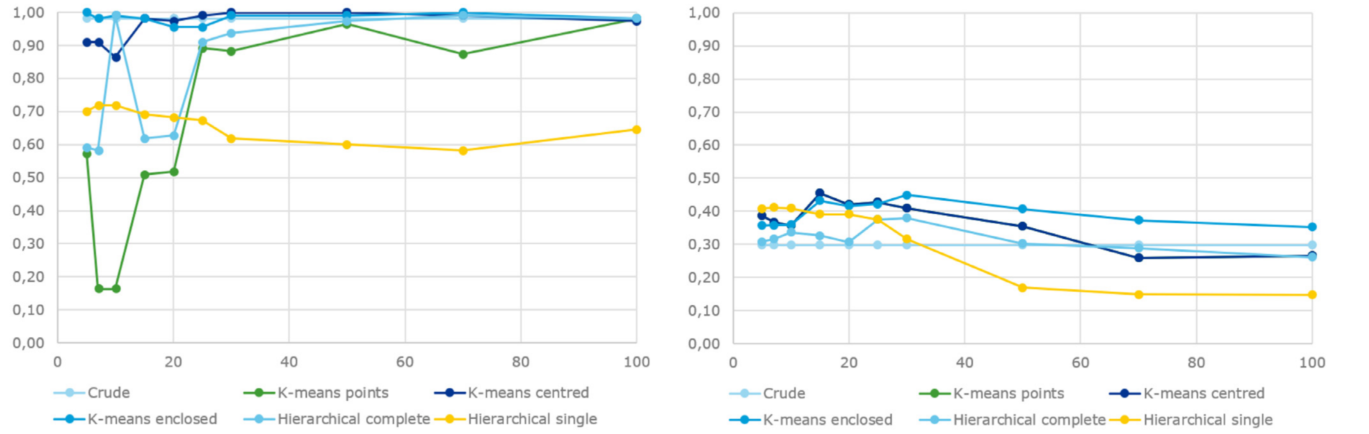
We observe that the area under the curve for the different methods are quite similar, again with a somewhat decreased performance for the hierarchical clustering with the single linkage criterion. The performance differs extensively on the different data sets, with area under the curve as high as 1.00 on some data sets, meaning that all instances are correctly labelled, both the true normal and the true anomalous. On other data sets, however, the performance is quite low, and for some data sets even close to 0.5. That being said, we have not investigated how subtle the anomalies are in the different data sets. In some of the data sets, the anomalies can be very obvious, and in others they can be well-hidden. Hence, the numbers presented here are intended for comparison of performance of the proposed methods with each other, and with the crude method. Our claim is not that the proposed cluster based methods are specifically suitable to solve the particular problems of the specific data sets, but we aim to demonstrate that the best proposed cluster based methods efficiently can achieve performance results comparable to the crude method, while inducing considerable reduction in computation time.

4.4.3. Number of clusters

Fig 10 illustrates how changes in the number of clusters used affects the performance. In figure (a) and (b) respectively, the True Positive Rate and True Negative Rate for the *segment0* data set are shown for various number of clusters. The threshold value τ is kept constant at 0.97. We observe, as expected, that the results converge towards the result of the crude method, as the number of clusters increases. However, we also observe surprisingly good results with very few clusters for all methods, except the hierarchical clustering method which uses the single linkage criterion.

Table 3
Area under the ROC curve. (Hierarchical clustering is not performed for data set 14 due to the large size of the training set). The number of clusters is 25 for data set 1–13. For data set 14, 100 clusters are used.

Dataset	Crude	K-means points	K-means centred	K-means enclosed	Hier. complete	Hier. single	Time crude	Time cluster	Relative time
1	0.92	0.88	0.92	0.91	0.89	0.47	179	5.2	2.9%
2	0.59	0.71	0.52	0.52	0.35	0.16	371	4.2	1.1%
3	0.75	0.77	0.93	0.76	0.83	0.69	11	0.6	5.4%
4	0.41	0.62	0.57	0.55	0.60	0.51	9	0.7	8.3%
5	1.00	1.00	0.98	0.97	1.00	0.71	629	5.8	0.9%
6	1.00	1.00	1.00	1.00	1.00	0.86	64	3.9	6.2%
7	1.00	0.99	0.99	0.97	1.00	1.00	10	0.7	7.5%
8	0.59	0.70	0.59	0.60	0.59	0.33	538	6.0	1.1%
9	0.96	0.86	0.95	0.97	0.93	0.14	12	0.9	7.5%
10	0.61	0.51	0.54	0.48	0.53	0.24	265	4.3	1.6%
11	0.91	0.79	0.88	0.90	0.75	0.45	1511	15.3	1.0%
12	0.94	0.81	0.88	0.88	0.79	0.46	159	5.5	3.5%
13	0.73	0.74	0.77	0.74	0.68	0.55	158	5.0	3.2%
14	0.73	0.82	0.81	0.75			5834	19.6	0.3%



(a) True Positive Rate: TP/(TP+FN)

(b) False Positive Rate: FP/(FP+TN)

Fig. 10. The True Positive Rate and False Positive Rate of data set *segment0* is shown, where the number of clusters used by the cluster based methods vary on the horizontal axis. The threshold value τ is kept constant at 0.97.

5. Marine engine case study with comparisons

In this section, the anomaly detection framework using AAKR in combination with SPRT, both with and without the modifications proposed in Section 3, are applied on the data set consisting of sensor measurements from a large marine diesel engine. The data is collected from a large ocean going ship in operation.

We limit the further analysis to only consider the surrounding sets that are centred at the cluster mean. The size of the surrounding sets are determined by the standard deviation of the cluster members, multiplied with the rectangle scaling factor γ. We present results using three different sizes of γ, and refer to them as points ($\gamma = 0$), small rectangles ($\gamma = 0.5$) and large rectangles ($\gamma = 1$).

5.1. Data description

The data is collected over a period of 10 months, starting in December 2014. A total of 333,144 observations are recorded, which includes idling. In this study, we concentrate on normal operation and use a simple filter based on engine speed [rpm] to remove the idling states, leaving us with a data set consisting of 175,558 rows.

We consider the following sensors:

- engine speed [rpm],
- lubricant oil inlet pressure [bar],
- lubricant oil inlet temperature [C],
- engine power [kW]
- engine bearing temperature [C]

The bearing temperature is considered the response signal, and the others are used as explanatory variables, when this is distinguished. The time series are shown in Fig. 11.

5.2. Operational mode

The ship investigated in this study, is operated in different operational modes, such as transit (in different speeds), port and stand by (with or without anchor), in addition to transient modes. A ship is in a transient mode when its operation changes from one defined mode to another. According to our experience, these modes are the most challenging ones, in respect to anomaly detection.

5.3. Cross validation

When predictions from a statistical model is evaluated on the data set used to train the model, the accuracy estimates tend to be overoptimistic (Arlot & Celisse, 2010). Hence, the data set $\mathcal{D}$ should be divided into exclusive parts where one part, $\mathcal{D}_{train}$, is used to train the model, and the other, $\mathcal{D}_{test}$, is reserved for testing. To build robust and accurate models we ideally want to include all data available in the training data set. The same applies to testing; we want to test our models in many situations. Cross validation introduces various methods of repetitively splitting the

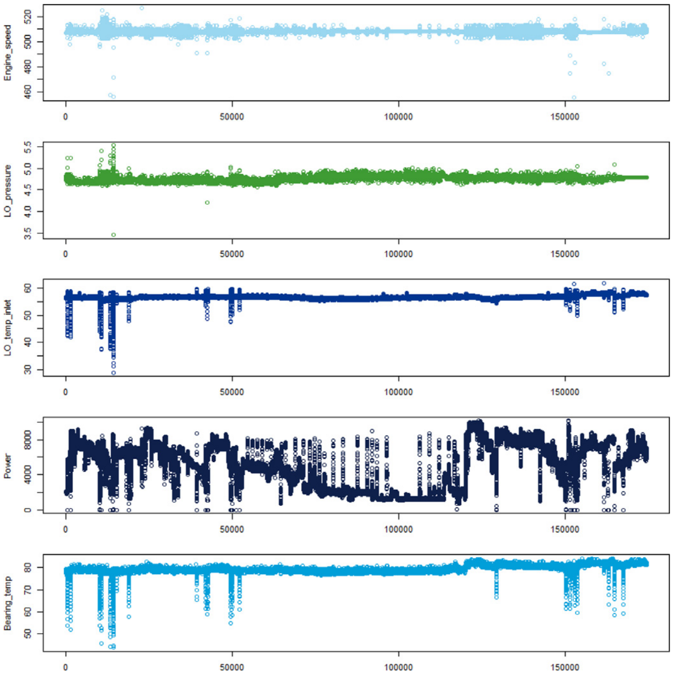


Fig. 11. Time series with training data for the evaluated signals.

data into training and test data sets. A range of different splitting techniques can be applied. See for example (Arlot & Celisse, 2010; Kohavi, 1995) for a brief overview of the most common splitting techniques. We also note that repeated k-fold cross validation can be used to stabilize the error estimation and reduce the variance (Jiang & Wang, 2017; Kohavi, 1995; Rodriguez, Perez, & Lozano, 2010).

In this study, we repeatedly select folds or time intervals containing 1000 query vectors, which constitute the test data set, $\mathcal{D}_{test}$. The remaining 174,000 points constitute the training data set $\mathcal{D}_{train}$. We repeat this procedure 15 times, leaving us with a total of 15,000 tested points.

5.4. Fault simulation

To our knowledge, no faults or anomalies are registered and reported by the crew, shipowner, etc. for the data set we envisage. Hence, we assume that the data set represent normal behaviour and we define normal states based on this data.

To be able to test the anomaly detection framework, we alter some of the signals to simulate faulty states. The anomaly we induce in the test data, is a temperature change in one of the main bearings of the engine. The other signals remain unchanged. For each test set $\mathcal{D}_{test}$, we increase the temperature with A^+ degrees Celsius in the area 200:400, and decrease the temperature with A^- degrees Celsius in the area 600:800. The set up is illustrated in Fig. 12.

The signals are only altered slightly. Fig. 13 shows a scatter plot comparing the training and the test data set, with both A^+ and A^- set to 1.0. The training data are shown in purple, and the test data are shown in blue, green and red, to mark the normal state and the two states with increased and decreased temperatures respectively. On the diagonal, a density plot of each individual signal are shown. The correlations are shown in the upper triangle. We observe that the test values, both in the regions with normal condition, and in the regions were we have altered the signals, lie within the normal operating mode of that specific signal. Hence, a rule based anomaly detection method based on a single threshold would not be able to detect the anomaly.

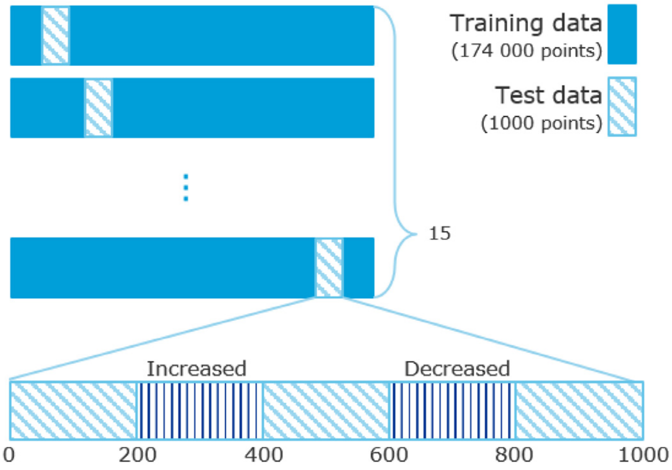


Fig. 12. Illustration of the test set up.

5.5. Evaluating the signal reconstruction

First, we evaluate the signal reconstructions, by comparing the root mean squared error (RMSE) under various conditions. When no anomalies or faults are present in the data, we want the difference between the observed signals and their reconstructions to be as small as possible. The RMSE of the reconstructed temperature signal using the proposed cluster based AAKR is shown in Fig. 14. Due to high computational cost, for very large number of clusters, we select a subset of the available data consisting of 20,000 points, and produce predictions combining different number of clusters and rectangle scaling factors. Here, no anomalies are simulated (A^+ and A^- are set to 0), and the data are assumed to be collected from normal operation.

Note that a rectangle scaling factor of 0 corresponds infinitely small rectangles, i.e. points. Hence, if the rectangle scaling factor is 0, and the number of clusters is equal to the number of historical observations, the reconstruction method resembles the standard AAKR method with the crude memory vector selection where all historical observations are included. The RMSE, using this method, is shown in the lower right hand corner in Fig. 14.

The choice of number of clusters depends on the requirements in calculation time. More clusters will increase accuracy, but computation time will also increase. In this study, we chose to use 100 clusters, and experiment with three rectangle scaling factors 0, 0.5, and 1. We refer to these three options as points, rectangles and large rectangles respectively.

5.5.1. Difference in RMSE with and without anomalies

For the Sequential Probability Ratio Test (SPRT) to be able to successfully detect anomalies, the residuals, i.e. the difference between the observed and the reconstruction signals, should be more pronounced for observations from the anomalous states, compared to observations from normal state. To indicate how the residuals change when we induce anomalies, we reconstruct the signals on the 15 different folds, and calculate the RMSE before and after the anomalies are induced.

The results are shown in the box plots in Fig. 15, for the 15 different folds. Results based on the crude AAKR, where all historical observations are included as memory vectors, and the cluster based version with points (infinitely small rectangles), rectangles and large rectangles are shown. We observe that the calcu-

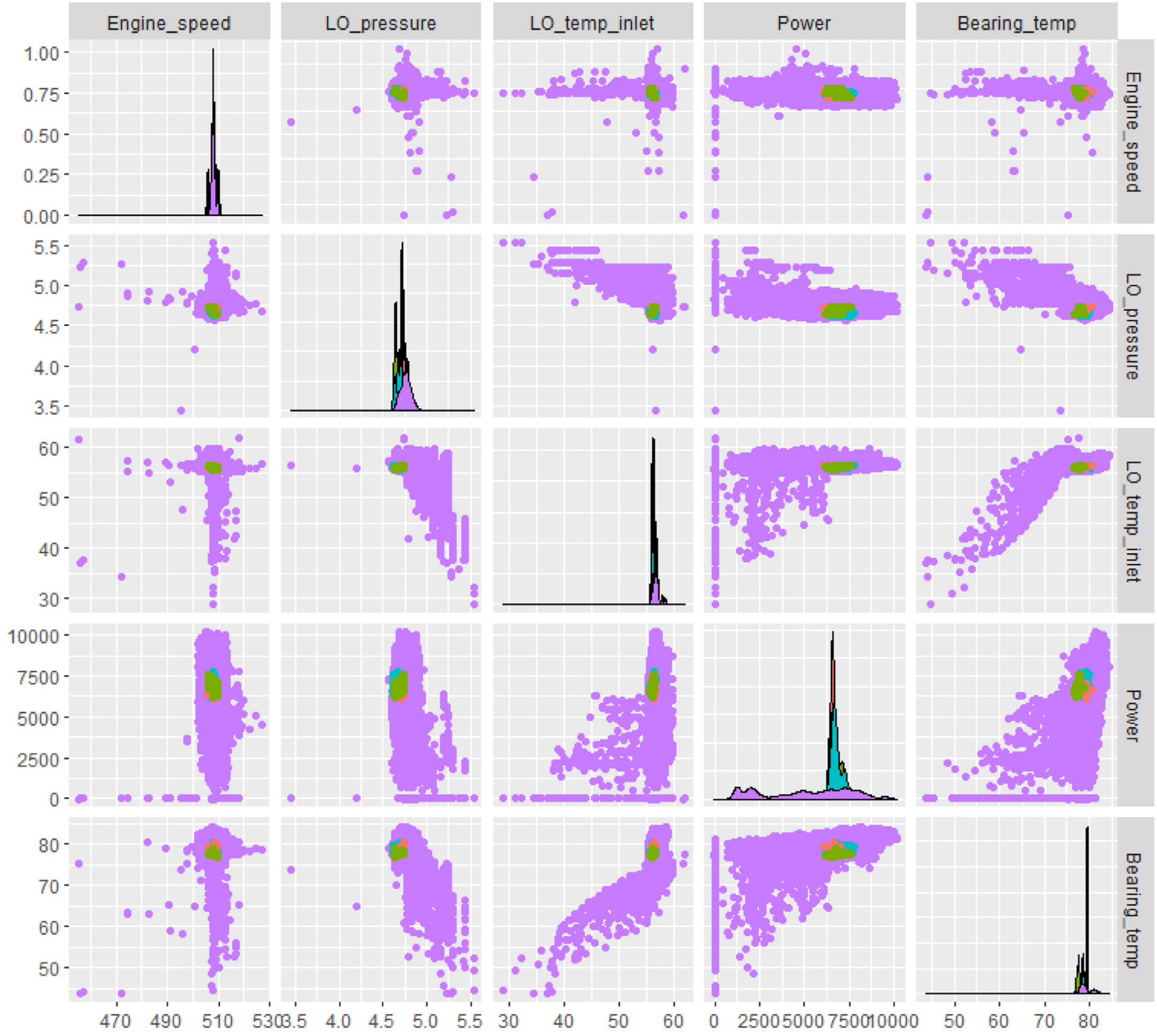


Fig. 13. A scatter plot comparing the training (purple) and the test data set from one of the tested folds, which contains two regions with anomalies (red and green), and the remaining points are considered normal (blue). In this illustration, the training and test data consists of 174,000 and 1000 points respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

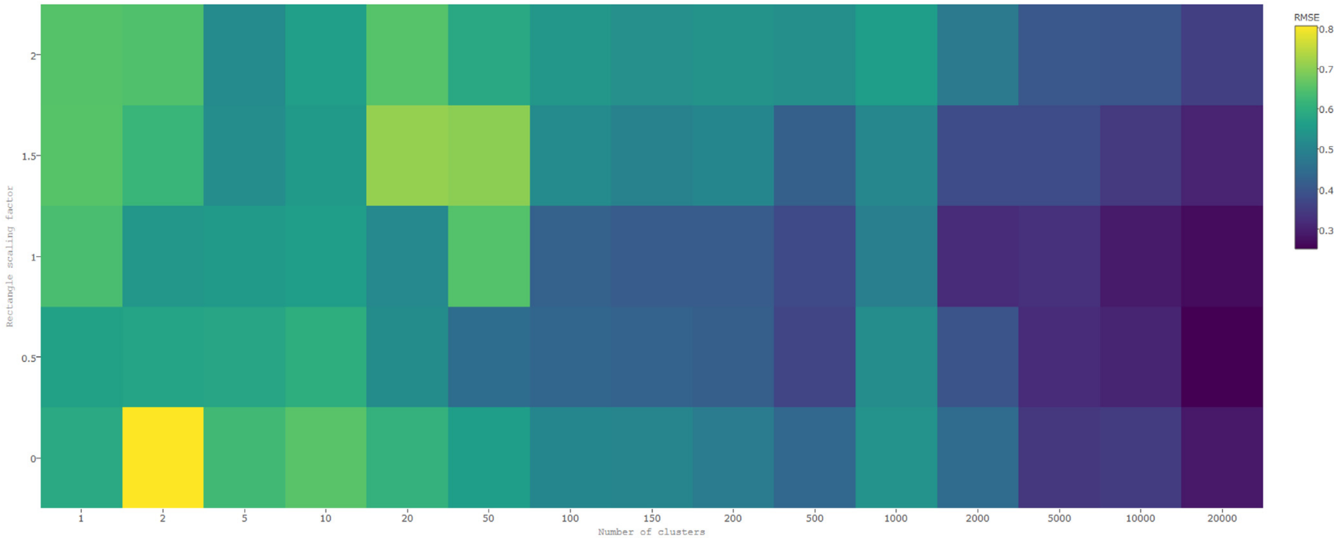


Fig. 14. The root mean squared error (RMSE) of the cluster based AAKR, with different number of clusters and different rectangle scaling factors. Note that when the number of clusters is equal to the number of points, in this example 20,000, and the rectangle scaling factor is set to 0, it resembles the crude AAKR. The kernel bandwidth h is set to 0.2.

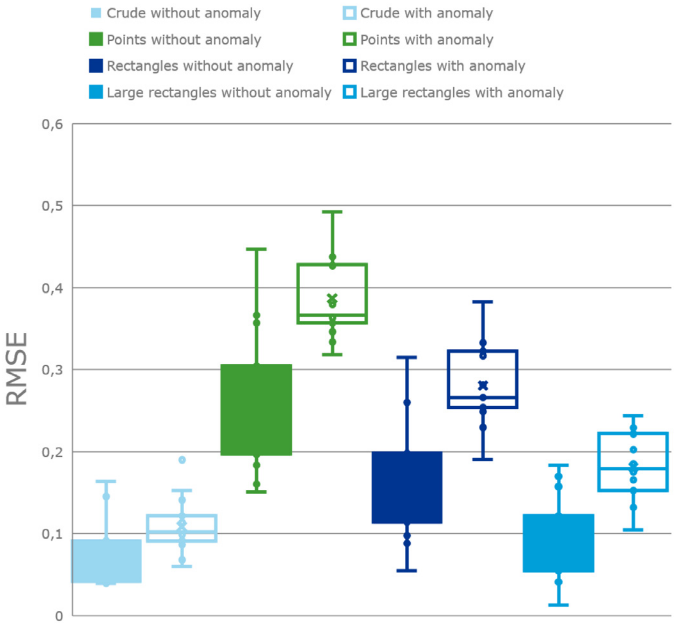


Fig. 15. Box plot of RMSE values calculated with the different memory vector selection methods with and without induced anomalies, on 15 different folds or folds.

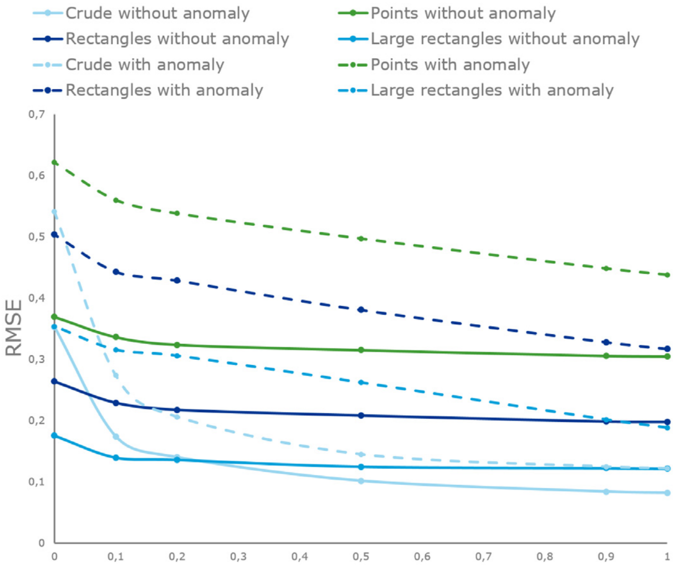


Fig. 16. RMSE values calculated based on reconstructions using the different memory vector selection methods. Values based on calculations with and without anomalies induced are showed with in filled and dotted lines respectively. Here, we vary the *J*th component of the distance scaling vector **s**, and keep the other distance scalings factors constant at 1. The *J*th signal is the bearing temperature.

lated RMSE is greater after anomalies are introduced, which indicates that it should be possible to detect the anomalies. The lowest RMSE is achieved with the crude method, closely followed by the method which use large rectangles. We observe that the differences between RMSE before and after anomalies are induced are more pronounced for reconstructions based on the cluster based methods.

5.5.2. Distance scaling vector

Now we analyse how the distance scaling vector **s**, as introduced in Section 3.2, effects the RMSE before and after anomalies are induced. Fig. 16 shows the average of the RMSE calculated from the different 15 folds. The filled and dotted lines are based on calculations before and after anomalies are induced respectively. Here, we only vary the *J*th component of the distance scaling vector **s**, and keep the other distance scaling vectors constant at 1. The *J*th signal is the bearing temperature.

When the *J*th component of the distance scaling vector is 0, the results of both the crude method and the cluster based methods are small and similar, with values in the range [0.12,0.15]. For larger values of the *J*th component of the distance scaling vector, we observe a significant difference in favour of the cluster based version. Remember, when anomalies are induced we want the AAKR method to produce reconstructions resulting in large residuals, and large RMSE values, while for fault-free signals, without anomalies, we want the RMSE values to be as low as possible.

5.5.3. Analysing the empirical distributions of the residuals

The empirical distribution of the residuals based on reconstructions made with the crude AAKR and the cluster based AAKR, with large rectangles, rectangles and points as surrounding sets, are shown in Fig. 17. As described in Section 5.4, a positive and negative change in mean has been induced in the time intervals 200:400 and 600:800 respectively. Outside of these two time intervals, no anomalies are induced.

The vertical dotted lines in the figure show the means of the three hypotheses; H_0 in the middle, where no anomalies are induced, and the two chosen alternative hypotheses, H_1 on the right hand side and H_2 on the left hand side, for respectively positive and negative changes in mean.

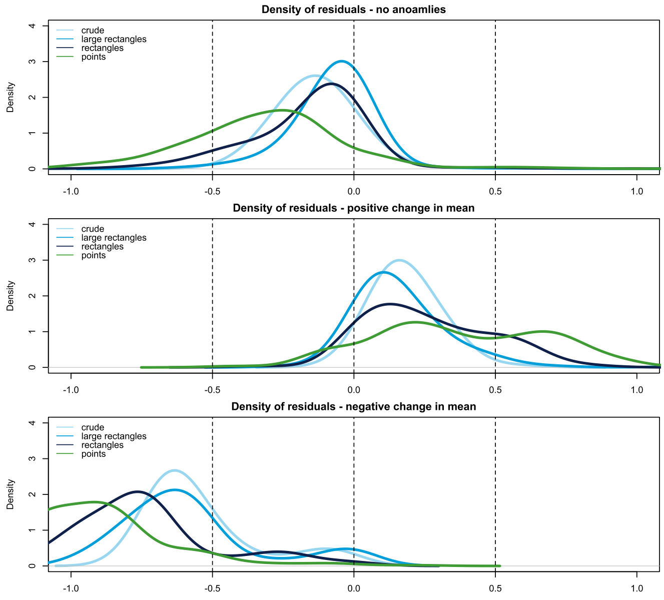


Fig. 17. Estimated densities of the residuals based on the reconstructions from the crude AAKR and the cluster based AAKR, with large rectangles, rectangles and points as surrounding sets. In the upper plot, the densities are based on signals that are not changed. In the middle and lower plot, the densities are based on values from signals that are altered in the positive and negative direction respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

When no anomalies are introduced, we expect the residuals to be small, and centred around zero. The estimated densities of the residuals, when no anomalies are induced, are shown in the upper plot of Fig. 17. We observe that the residuals are mainly situated around zero, but especially the density of the residuals based on

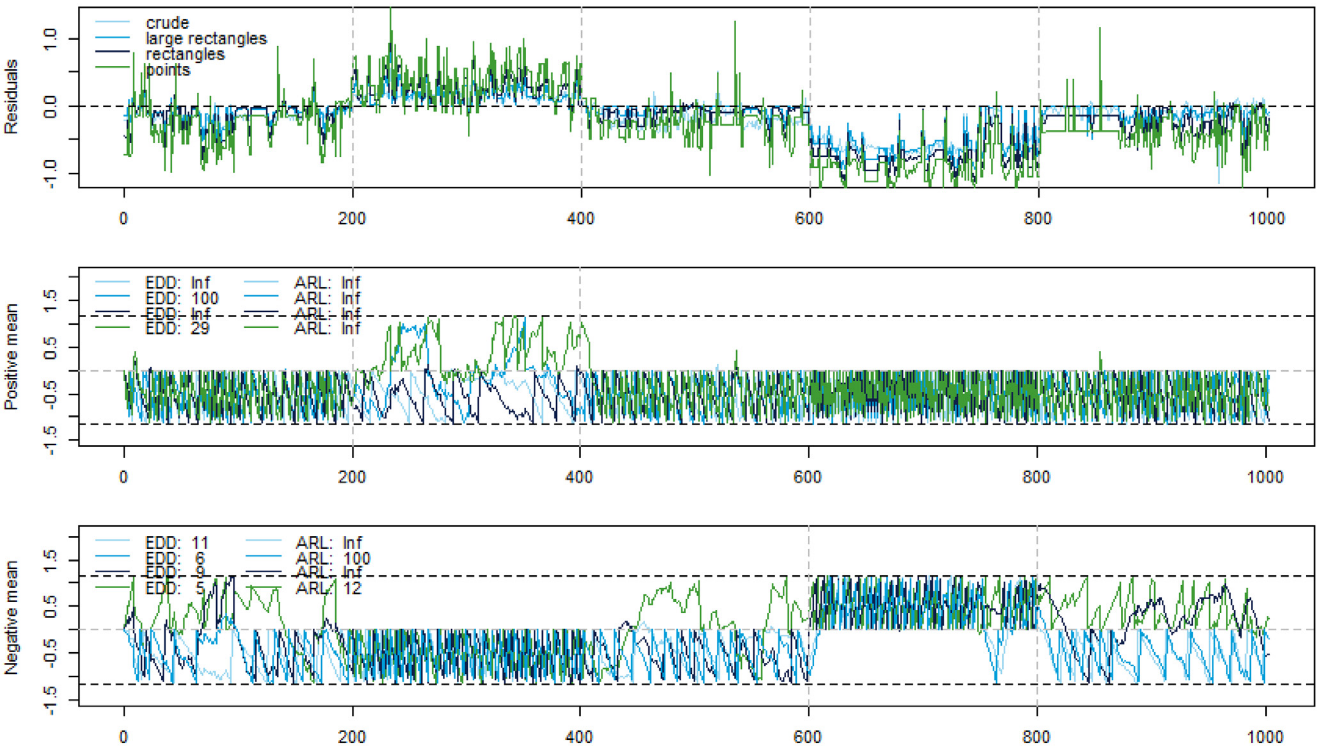


Fig. 18. The residuals are shown in the upper plot. The middle and lower plot show the SPRT indices for positive and negative changes in the mean.

reconstructions using points as surrounding sets (green line) seems to be shifted in the negative direction.

The middle and lower plots show estimated densities from signals which are altered to mimic anomalies. Residuals based on a positive and a negative change in mean are shown in the middle and lower plots respectively. The middle plot shows a slight shift in the positive direction. The shift is most evident in the residuals from reconstructions using the cluster based AAKR with points as surrounding sets. Also the residuals based on reconstructions using rectangles as surrounding sets are quite noticeable. In the lower plot, a shift in negative direction is indisputable, for all reconstructions.

5.5.4. Computation time

The computation time of producing 1000 reconstructions with 175,000 historical observations is about 22 minutes using the crude memory vector selection method. In comparison, the cluster based version, with 100 clusters, produces the 1000 reconstructions in less than 5 s. The time to perform the clustering, using K-means clustering, with the Lloyd algorithm, is about 95 s. However, the clustering only needs to be performed once, and does not need to be performed on-line, hence we believe the time to perform clustering should not be an issue.

5.6. Illustration of the sequence of residuals and the SPRT indices

An example of the residuals analysis using SPRT is displayed in Fig. 18. The residuals are displayed in the upper plot, while the middle and lower plots show the SPRT indices of the positive and negative change in mean respectively. If a value exceeds the upper horizontal dotted line, an alarm is raised, either for positive or negative change in mean, and the sequential test is reset. Similarly, if the value is below the lower horizontal line, the sequential test is reset. But now, confidence of normal state is reached, and no alarm is raised.

The approximated expected detection delay (EDD) and average run length (ARL) of the various reconstruction methods are reported in the figure. The EDD is the expected number of time points from an anomaly is introduced until it is detected, and ARL is the expected number of time points between false alarms.

The induced fault in the example presented in Fig. 18 is a temperature change of +1 °C in the first anomalous time interval and -1 °C in second anomalous time interval. Furthermore, the kernel bandwidth, h, is 0.1, the mean value of the two alternative hypothesis, for positive and negative change in mean, M, is set to 1, and the standard deviation, σ, is extracted from the training data. The distance scaling factor $\mathbf{s}$ is fixed at [1,1,1,1,0.1]. Note that if the last entry is 1, the original AAKR reconstruction will be performed, while if the last entry is 0, a standard Nadaraya–Watson regression will be used. See Figs. 19 and 21 for results with other choices of $\mathbf{s}$.

For positive change in mean, an EDD of 29 is returned when points are used as surrounding sets, while it is 100 when large rectangles are used. Otherwise no alarms for positive change in mean are raised in this example. Neither, no false alarms are raised. For negative change in mean, more alarms are raised. We observer that the lowest EDD is achieved by the use of points as surrounding sets, but this also provides a low ARL of 12. We note that the results are well aligned with Fig. 17.

5.7. Results using multiple surrounding sets, distance scaling vectors and credibility factors

Results of the proposed anomaly detection framework are presented in Figs. 19–21. Multiple surrounding sets are used for the cluster based AAKR reconstruction, and this is combined with multiple distance scaling vectors and credibility factors. All entries in

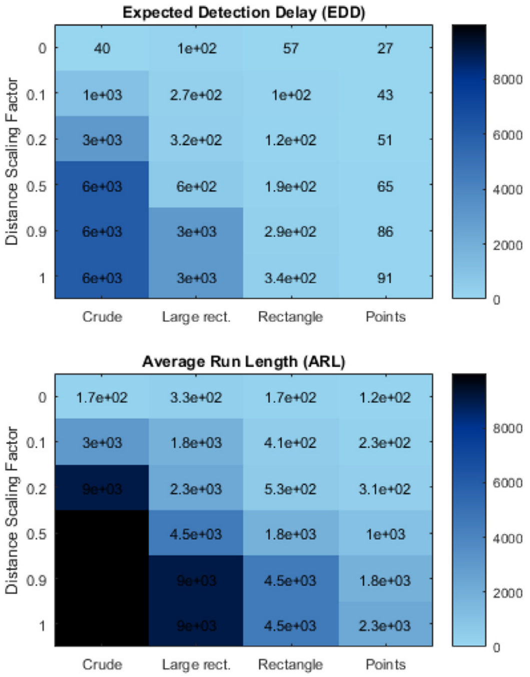


Fig. 19. Surrounding set and distance scaling vector: EDD and ARL at various surrounding sets and distance scaling vectors. If no alarms are raised, the EDD and ARL cannot be calculated. These are represented with black colour. The bandwidth of the credibility estimation is set to infinity, which means that all areas are considered equally credible.

the distance scaling vector can be adjusted, but here we concentrate on the *J*th component. The values in the tables represent approximations of the mean EDD and mean ARL, taken over the whole test period of 15 folds, with 1000 points in each. The presented results are well aligned with our expectations, and show consistent behaviour.

In Fig. 19, the anomaly detection capability of the methodology using the crude and the cluster based AAKR with different surrounding sets for reconstruction, combined with residuals analysis using a range of different distance scaling factors, are presented. We observe that the lowest EDD is achieved by combining points (infinitely small rectangles) as surrounding sets with distance scaling vector 0. Furthermore, the EDD increases when the distance scaling vector is increased. Also, the EDD seems to increase when the size of the surrounding sets is increased. As expected, the ARL follows the same pattern. This illustrate the usual trade-off between EDD and ARL; we want low EDD, but this will of course cause a decrease in the ARL.

Fig. 20 illustrates how changes in credibility factor effects the EDD and ARL. Again, we apply reconstructions produced both with the crude and cluster based AAKR. Here, we fix the distance scaling vector $\mathbf{s}$ at $[1, 1, \dots, 1, 0.1]$, and concentrate on the change in credibility factor. We observe, as expected, that both the EDD and the ARL decreases with when the credibility factor increases.

In Fig. 21, EDD and ARL based on various combinations of distance scaling vectors and credibility factor are presented. We chose to use the reconstruction version with large rectangles as surrounding set.

5.8. Discussion and suggestions for further research

In the following, we discuss some key challenges and suggestions for anomaly detection, with emphasis on the maritime industry.

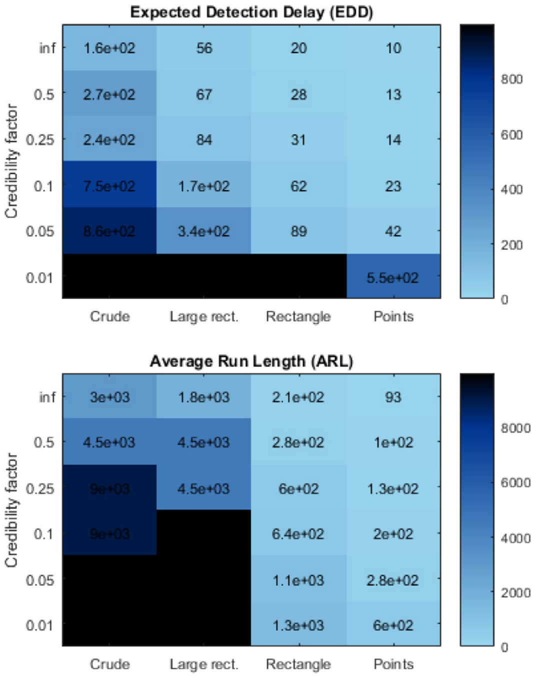


Fig. 20. Surrounding set and credibility factors: EDD and ARL at various surrounding sets and various credibility estimate factors. If no alarms are raised, the EDD and ARL cannot be calculated. These are represented with black colour. The distance scaling vector, $\mathbf{s}_J$, is 0.1.

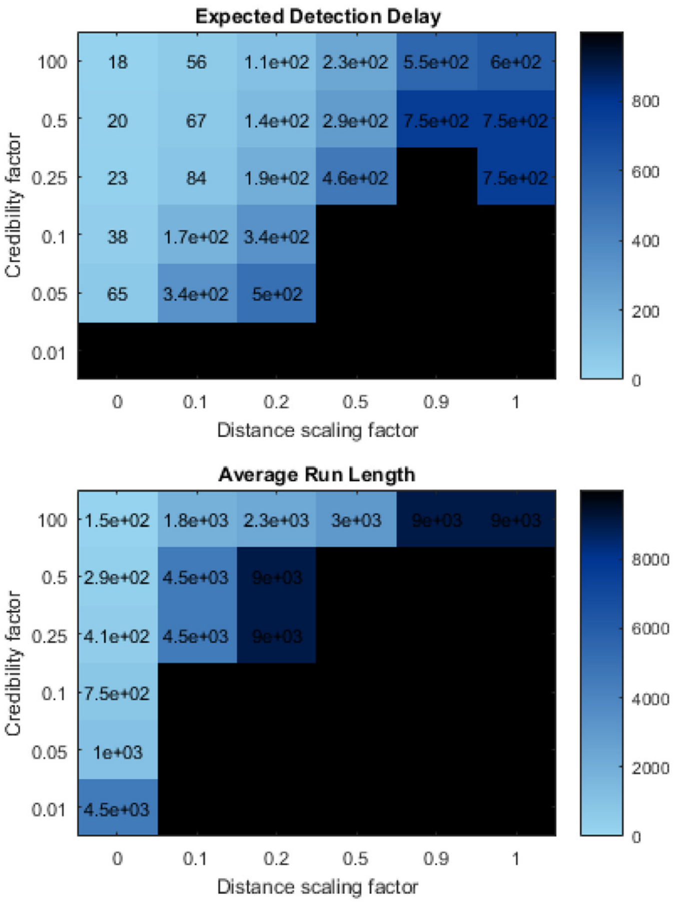


Fig. 21. Distance scaling vectors and credibility factors: EDD and ARL at various distance scaling vectors, and credibility factors. The figure is based on reconstructions produced using cluster based AAKR, with large rectangles.

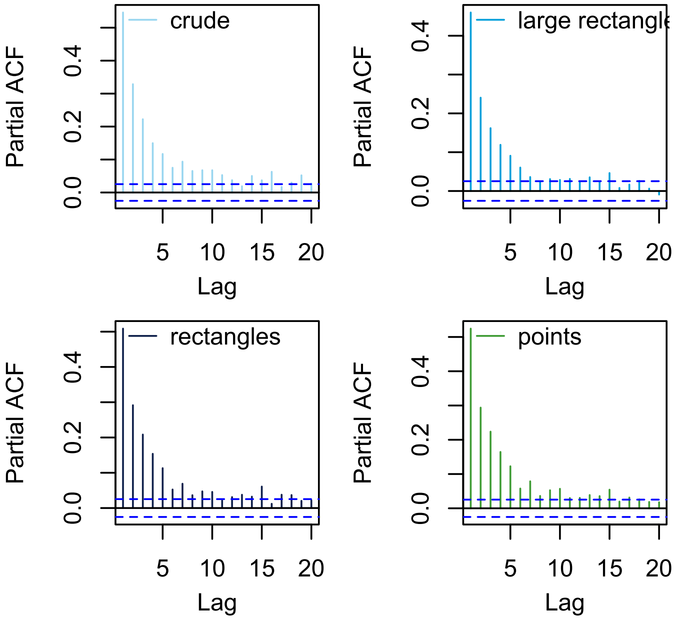


Fig. 22. Partial auto correlation function of the residuals in the bearing temperature sensor.

5.8.1. Extensions to high-dimensional sensor data

In this paper we apply the anomaly detection framework on data sets containing a very limited amount of sensor signals and performed the reconstruction of the measured signals based on distances from the training data in low-dimensional space. However, sensor monitoring of typical ship systems will often consist of hundreds of sensors and it remains to be seen how well the proposed approach scales in higher dimensions. The method will suffer from the curse of dimensionality (Keogh & Mueen, 2011), which will make it more challenging to establish similar models for high-dimensional data. Sensible techniques for dimension reduction will have to be carried out before the signals are analysed with AAKR. Additionally, feature extraction should be investigated further. We believe this is an interesting and important topic for further research.

5.8.2. Operational mode selection

During the different operating modes the behaviour of a ship changes substantially, and it might therefore be advantageous to develop reconstruction models dedicated to the different operational modes. This could also allow the alarm limits to vary in the different modes, depending on the operations criticality. To achieve this, the training data should be divided and used to fit different models. This will result in reduced computational efforts and increased model reconstruction accuracy (Al-Dahidi, Baraldi, Di Maio, & Zio, 2014; Baraldi et al., 2012).

5.8.3. Partial auto correlation in the residuals

The partial auto correlation function of the residuals, made with crude AAKR and cluster based AAKR, with large rectangles, rectangles and points as surrounding sets are shown in Fig. 22. The figure reveals that some time dependence is present in the residuals, for time lags below 5–10 s. We also observe that the dependency structure is similar in the four cases.

5.8.4. Training data extension

Sometimes training data are not available. For instance when a ship is entering a type of operation that has not been tested before, or if a ship is moved to a new geographical area, where it has never operated before, the training data might need to be modified to represent the “new” normal conditions. If the sensors are affected in a deterministic way, new training data can be simulated, based on the other training data. Ships are usually built in sister series. The sensor data collected by the first ship in a series, can possibly be reused by a later ship in the series. Also when the ships are not identical, it is possible that the training data from the first ship can be used on the later one, after necessary calibrations and modifications detailed by simulation software such as for example Dimopoulos, Georgopoulou, Stefanatos, Zymaris, and Kakalis (2014).

6. Conclusion

The paper introduces three generalizations and modifications of an on-line anomaly detection framework consisting of signal reconstruction with Auto Associative Kernel Regression (AAKR) and residuals analysis using Sequential Probability Ratio Test (SPRT).

We demonstrate the ability of the cluster based memory vector selection method for AAKR, which is successfully used for faster signal reconstruction. The methodology is applied to multiple imbalanced benchmarking data sets, in addition to the data set with sensor signals from a marine diesel engine in operation. Many of the anomalies are quite subtle, restrained enough not to easily be revealed by for example analysing scatter plots of the data. Results of the crude and the cluster based methods are presented and compared, and the analysis show that comparable results are achieved, even when very few (<25) clusters are used. The advantage of the cluster based methods is the increased speed. The computation time of the AAKR grows rapidly when the size of the training data increases, and we demonstrate how the presented cluster based memory vector selection technique can be used to dramatically decrease the computation time, at the same time as the performance is kept at an acceptable level.

We also show how the cluster based AAKR can be used in combination with the SPRT, which is used for residuals analysis, to construct a robust and fast anomaly detection framework. The results are well aligned with our expectations, and show consistent behaviour. A generalization of the distance measure used in the signal reconstruction process is proposed, which enables the users system-knowledge to be imposed on the anomaly detection framework to distinguish response and explanatory variables and optimize the weighting of the different features. The distance scaling vector can be chosen to achieve acceptable levels of expected detection delay (EDD) and average run length (ARL).

We also introduce a credibility estimate which enables the SPRT method to reach a conclusion faster when it operates in regions close to instances which are well represented in the training data set, and allows it to use more time to reach a conclusion when it operates in less explored regions.

CRediT authorship contribution statement

Andreas Brandsæter: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Project Administration. **Erik Vanem:** Validation, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition. **Ingrid K. Glad:** Validation, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition.

Acknowledgements

The work is carried out in collaboration with the Big Insight project, and is partly funded by the Research Council of Norway, project number 237718 and 251396.

Appendix A

Abstracts of the original classification data sets is provided below, together with a description of how anomalies are defined for each of the data sets. The descriptions are collected here: Alcalá-Fdez et al., 2011 and Dua and Efi (2017).

Data set	Abstract	Description of anomaly
vehicle0	3D objects within a 2D image by application of an ensemble of shape feature extractors to the 2D silhouettes of the objects	Positive examples belong to class 0 (Van) and the negative examples belong to the rest.
yeast6	Predicting the Cellular Localization Sites of Proteins	Positive examples belong to class EXC and the negative examples belong to the rest.
ecoli-0-1-3-7_vs_2-6	This data contains protein localization sites	Positive examples belong to classes pp and imL and the negative examples belong to classes cp, im, imU and imS.
glass5	From USA Forensic Science Service; 6 types of glass; defined in terms of their oxide content (i.e. Na, Fe, K, etc.)	Positive examples belong to class 5 and the negative examples belong to the rest.
shuttle-c0-vs-c4	The shuttle data set contains 9 attributes all of which are numerical. Approximately 80% of the data belongs to class 1	Positive examples belong to class 0 and the negative examples belong to class 4.
dermatology-6	Aim for this data set is to determine the type of Eryhemato-Squamous Disease.	Positive examples belong to the class 6 and the negative examples to the rest of the classes.
shuttle-6_vs_2-3	The shuttle data set contains 9 attributes all of which are numerical. Approximately 80% of the data belongs to class 1. The task is to decide what type of control of the vessel should be employed.	Positive examples belong to the class 6 and the negative examples belong to the classes 2–3.
winequality-red-4	The data set is related to red variant of the Portuguese Vinho Verde wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).	Positive examples belong to the class 4 and the negative examples belong to the rest of classes.
poker-9_vs_7	Each record of this data set is an example of a hand consisting of five playing cards drawn from a standard deck of 52. Each card is described using two attributes (suit and rank), for a total of 10 nominal attributes. The class attribute describes the Poker Hand obtained	Positive examples belong to the class 9 and the negative examples belong to the class 7.
yeast1	Predicting the Cellular Localization Sites of Proteins	Positive examples belong to class NUC and the negative examples belong to the rest.
segment0	This data set is an image segmentation database similar to a database already present in the repository (Image segmentation database) but in a slightly different form.	Positive examples belong to class 1 and the negative examples belong to the rest.
vehicle2	3D objects within a 2D image by application of an ensemble of shape feature extractors to the 2D silhouettes of the objects	Positive examples belong to class 2 (Bus) and the negative examples belong to the rest.
vehicle3	3D objects within a 2D image by application of an ensemble of shape feature extractors to the 2D silhouettes of the objects	Positive examples belong to class 3 (Opel) and the negative examples belong to the rest.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2018.12.040.

References

Ahn, H.-K., Bae, S. W., Demaine, E. D., Demaine, M. L., Kim, S.-S., Korman, M., et al. (2011). Covering points by disjoint boxes with outliers. *Computational Geometry, 44*(3), 178–190.

Al-Dahidi, S., Baraldi, P., Di Maio, F., & Zio, E. (2014). Quantification of signal reconstruction uncertainty in fault detection systems. *The second European conference of the prognostics and health management society*.

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., et al. (2011). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing, 17*, 255–287.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40–79. doi:10.1214/09-SS054.

Baraldi, P., Canesi, R., Zio, E., Seraoui, R., & Chevalier, R. (2011). Genetic algorithm-based wrapper approach for grouping condition monitoring signals of nuclear power plant components. *Integrated Computer-Aided Engineering, 18*(3), 221–234.

Baraldi, P., Di Maio, F., Genini, D., & Zio, E. (2015). Comparison of data-driven reconstruction methods for fault detection. *IEEE Transactions on Reliability, 64*(3), 852–860. doi:10.1109/TR.2015.2436384.

Baraldi, P., Di Maio, F., Pappaglione, L., Zio, E., & Seraoui, R. (2012). Condition monitoring of electrical power plant components during operational transients.. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, SAGE, *226*, 568–583.

Baraldi, P., Di Maio, F., Turati, P., & Zio, E. (2015). Robust signal reconstruction for condition monitoring of industrial components via a modified auto associative kernel regression method. *Mechanical Systems and Signal Processing, 60–61*, 29–44. doi:10.1016/j.ymssp.2014.09.013.

Boechat, A. A., Moreno, U. F., & Haramura, D. (2012). On-line calibration monitoring system based on data-driven model for oil well sensors. *IFAC Proceedings Volumes, 45*(8), 269–274.

Brandsæter, A., Manno, G., Vanem, E., & Glad, I. K. (2016). An application of sensor-based anomaly detection in the maritime industry. In *2016 IEEE international conference on prognostics and health management (ICPHM)* (pp. 1–8). doi:10.1109/ICPHM.2016.7811910.

Brandsæter, A., Vanem, E., & Glad, I. K. (2017). Cluster based anomaly detection with applications in the maritime industry. *2017 international conference on sensing, diagnostics, prognostics, and control, Shanghai, China*.

Cameron, S. (1997). EnhancinG GJK: Computing minimum and penetration distances between convex polyhedra. In *Robotics and automation, 1997. proceedings., 1997 IEEE international conference on: 4* (pp. 3112–3117). IEEE.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR), 41*(3), 15.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, M. M. (2014). Package ‘nbclust’. *Journal of Statistical Software, 61*, 1–36.

Cheng, S., & Pecht, M. (2012). Using cross-validation for model parameter selection of sequential probability ratio test. *Expert Systems with Applications, 39*(9), 8467–8473. doi:10.1016/j.eswa.2012.01.172.

Coble, J., Humberstone, M., & Hines, J. W. (2010). Adaptive monitoring, fault detection and diagnostics, and prognostics system for the iris nuclear plant. In *Annual Conference of the Prognostics and Health Management Society*.

Cord, M., & Cunningham, P. (2008). *Machine learning techniques for multimedia: Case studies on organization and retrieval*. Springer Science & Business Media.

Dattorro, J. (2010). *Convex optimization & Euclidean distance geometry*. USA: Meboo Publishing.

Di Maio, F., Baraldi, P., Zio, E., & Seraoui, R. (2013). Fault detection in nuclear power plants components by a combination of statistical methods. *IEEE Transactions on Reliability, 62*(4), 833–845. doi:10.1109/TR.2013.2285033.

Dimopoulos, G. G., Georgopoulou, C. A., Stefanatos, I. C., Zymaris, A. S., & Kakalis, N. M. (2014). A general-purpose process modelling framework for marine energy systems. *Energy Conversion and Management, 86*, 325–339.

Dua, D., & Efi, K.T., (2017). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise.. In *KDD: 96* (pp. 226–231).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Flaherty, N. (2017). Frames of mind. *Unmanned Systems Technology, 3*(3).

Friedman, J., Hastie, T., & Tibshirani, R. (2009). The elements of statistical learning. *Springer series in statistics*: 1 (2). New York, NY, USA: Springer-Verlag.

Garvey, J., Garvey, D., Seibert, R., & Hines, J. W. (2007). Validation of on-line monitoring techniques to nuclear plant data. *Nuclear Engineering and Technology, 39*, 133–142.

Gross, K. C., & Lu, W. (2002). Early detection of signal and process anomalies in enterprise computing systems. . In M. A. Wani, H. R. Arabnia, K. J. Cios, K. Hafeez, & G. Kendall (Eds.), *ICMLA* (pp. 204–210). CSREA Press.

Guha, S., & Mishra, N. (2016). Clustering data streams. In M. Garofalakis, J. Gehrke, & R. Rastogi (Eds.), *Data stream management: Processing high-speed data streams* (pp. 169–187)). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-28608-0_8.

Hines, J. W., & Garvey, D. R. (2006). Development and application of fault detectability performance metrics for instrument calibration verification and anomaly detection. *Journal of Pattern Recognition Research*.

Hines, J. W., Garvey, D. R., & Seibert, R. (2008). Technical review of on-line monitoring techniques for performance assessment (NUREG/CR-6895). Volume 3: Limiting case studies. *Technical Report*. United States Nuclear Regulatory Commission, Office of Nuclear regulatory Research.

Hines, J. W., Garvey, D. R., Seibert, R., & Usynin, A. (2008). Technical review of on-line monitoring techniques for performance assessment (NUREG/CR-6895). Volume 2: Theoretical issues. *Technical Report*. United States Nuclear Regulatory Commission, Office of Nuclear regulatory Research.

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

Jarvis, R. A. (1973). On the identification of the convex hull of a finite set of points in the plane. *Information Processing Letters, 2*(1), 18–21.

Jiang, G., & Wang, W. (2017). Error estimation based on variance analysis of k-fold cross-validation. *Pattern Recognition, 69*, 94–106.

Kanarachos, S., Christopoulos, S.-R. G., Chroneos, A., & Fitzpatrick, M. E. (2017). Detecting anomalies in time series data via a deep learning algorithm combining wavelets, neural networks and hilbert transform. *Expert Systems with Applications, 85*(Supplement C), 292–304. doi:10.1016/j.eswa.2017.04.028.

Keogh, E., & Mueen, A. (2011). Curse of dimensionality. In *Encyclopedia of machine learning* (pp. 257–258). Springer.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on artificial intelligence*. In *IJCAI'95: 2* (pp. 1137–1143). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137.

Michau, G., Palme, T., & Fink, O. (2017). Deep feature learning network for fault detection and isolation. In *Proceedings of the annual conference of the prognostics and health management society* (pp. 108–118). Citeseer.

Ng, R. T., & Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In *Proceedings of VLDB* (pp. 144–155). Citeseer.

Ng, R. T., & Han, J. (2002). Clarans: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering, 14*(5), 1003–1016.

Olson, C., Judd, K., & Nichols, J. (2018). Manifold learning techniques for unsupervised anomaly detection. *Expert Systems with Applications, 91*(Supplement C), 374–385. doi:10.1016/j.eswa.2017.08.005.

Park, S. H., & Kim, J.-Y.. Unsupervised clustering with axis-aligned rectangular regions. http://cs229.stanford.edu/proj2009/ParkKim.pdf.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical ComputingVienna, Austria.

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(3), 569–575.

Saranya, C., & Manikandan, G. (2013). A study on normalization techniques for privacy preserving data mining. *International Journal of Engineering and Technology, 5*, 2701–2704.

Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., et al. (2008). Metrics for evaluating performance of prognostic techniques. In *2008 international conference on prognostics and health management* (pp. 1–17). doi:10.1109/PHM.2008.4711436.

Wilks, D. (2011). Chapter 15 – Cluster analysis. In D. S. Wilks (Ed.), *Statistical methods in the atmospheric sciences*. In *International Geophysics: 100* (pp. 603–616). Academic Press. http://www.sciencedirect.com/science/article/pii/B9780123850225000154. doi:10.1016/B978-0-12-385022-5.00015-4.

Zheng, D., Li, F., & Zhao, T. (2016). Self-adaptive statistical process control for anomaly detection in time series. *Expert Systems with Applications, 57*(Supplement C), 324–336. doi:10.1016/j.eswa.2016.03.029.