Instantaneous failure mode remaining useful life estimation using non-uniformly sampled measurements from a reciprocating compressor valve failure

Panagiotis Loukopoulos [a], George Zolkiewski [b], Ian Bennett [b], Suresh Sampath [a], Pericles Pilidis [a], X. Li[c], and David Mba [c]

[a] School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, United Kingdom
Email addresses: loukopoulospn@gmail.com, s.sampath@cranfield.ac.uk, p.pilidis@cranfield.ac.uk
[b] Shell Global Solutions, Royal Dutch Shell, Rijswijk, Netherlands
Email addresses: george.zolkiewski@hotmail.com, Ian.Bennett@shell.com
[c] Faculty of Technology, De Montfort University, Leicester, United Kingdom
Email address: david.mba@dmu.ac.uk, xiaochuan.li@dmu.ac.uk

Abstract

One of the major targets in industry is minimisation of downtime and cost, and maximisation of availability and safety, with maintenance considered a key aspect in achieving this objective. The concept of Condition Based Maintenance and Prognostics and Health Management (CBM/PHM), which is founded on the principles of diagnostics, and prognostics, is a step towards this direction as it offers a proactive means for scheduling maintenance. Reciprocating compressors are vital components in oil and gas industry, though their maintenance cost is known to be relatively high. Compressor valves are the weakest part, being the most frequent failing component, accounting for almost half maintenance cost. To date, there has been limited information on estimating Remaining Useful Life (RUL) of reciprocating compressor in the open literature. This paper compares the prognostic performance of several methods (multiple linear regression, polynomial regression, Self-Organising Map (SOM), K-Nearest Neighbours Regression (KNNR)), in relation to their accuracy and precision, using actual valve failure data captured from an operating industrial compressor. The SOM technique is employed for the first time as a standalone tool for RUL estimation. Furthermore, two variations on estimating RUL based on SOM and KNNR respectively are proposed. Finally, an ensemble method by combining the output of all aforementioned algorithms is proposed and tested. Principal components analysis and statistical process control were implemented to create T^2 and Q metrics, which were proposed to be used as health indicators reflecting degradation processes and were employed for direct RUL estimation for the first time. It was shown that even when RUL is relatively short due to instantaneous nature of failure mode, it is feasible to perform good RUL estimates using the proposed techniques.

Key words: reciprocating compressor, valve, prognostics, remaining useful life, multiple linear regression, polynomial regression, self-organising map, K-nearest neighbours, instantaneous failure, principal components analysis, statistical process control.

1. Introduction

Reciprocating compressors are of the most essential components in oil and gas industry, being a key element in refining sector as one of most commonly used type of equipment, requiring high reliability and availability [1], [2]. They are widely used, being powerful, flexible, efficient and dependable in many compression applications. Despite their popularity, their maintenance cost can be several times higher than that of other compressor types [3], since the number of moving parts is greater [4], thus they are expected to experience more failures. Bloch and Heinz [1] note that valves are the most common failing part (36%), making them the weakest component, accounting for half the maintenance cost [4].

Valves are an essential part of the reciprocating compressor as they have a significant impact on its performance from both efficiency and reliability perspectives [2]. Their smooth operation is integral since they regulate the gas flow for compression. Valves suffer numerous hardships during their operation as they may come in contact with liquids, foreign particles or debris, corrosive gases or materials depending on application [2]. Furthermore, pulsations, tension, compression and impact created either by the compressor or the valve motion itself can affect proper valve function [2].

In order to decrease downtime and cost, while increasing availability and safety, efficient maintenance is essential [1], [2] since reciprocating failures can cause from production loss to human casualties [3], [5]. Condition Based Maintenance (CBM) [6]–[10] is a method founded on the diagnostics principle and has been increasingly popular over the years, advocating that maintenance should be made only when actually needed depending on unit's health state; it is an effective tool that moves towards this direction [1], with diagnostics being an established area for valve failures [3], [11]–[15]. The equipment of interest is mounted with sensors collecting Condition Monitoring (CM) measurements which are analysed for diagnostics purposes – determine whether healthy or a faulty, and in case of fault identify failure mode – and suggest actions to be taken accordingly.

An extension of CBM is Prognostics and Health Management (PHM) [6]–[9], [16]–[19] which has been gaining traction during recent years and is founded on prognostics principle [6]–[10], [16]–[20]. It predicts the time to failure, known as Remaining Useful Life (RUL), after a fault has occurred, enabling the user to schedule maintenance in advance. PHM's proactive nature can assist optimising maintenance by avoiding any unnecessary action. Since PHM can be employed after a fault has been detected, diagnostics is required and thus its coupling with CBM would be unavoidable, leading to CBM/PHM [6]. To the authors' knowledge, there is limited information about prognostics on reciprocating compressors in open literature. Consequently, the purpose of this project is comparison of several prognostics methods in order to identify most suitable ones based on accuracy and variability.

Prognostics techniques can be divided into two groups [6], [7], [9], [16]–[21]:

  i. *Data-driven. They model the degradation process using historical information, and are suitable when there is limited physical understanding of system under study. They struggle in cases for which they have not been trained like novel events, while their accuracy depends on amount and quality of available data.*
  ii. *Physics based. They create a mathematical representation of system's or failure's physical aspect. They are computationally expensive and tend to be application specific though they can outperform data-driven.*

Similarly, there are two ways for calculating RUL [7], [21], [22]:

  i. *Direct estimation. Relationship between information and RUL is modelled. It requires knowledge of historical and current information, with data being the input and RUL being the output. It is useful in cases lacking failure threshold.*
  ii. *Indirect estimation. Relationship between information and a Health Indicator (HI), reflecting machine's health status, is modelled. In some cases HI can be modelled as function of time. HI is extrapolated until a failure threshold is reached. RUL is estimated as difference between current and failure time. It requires knowledge of historical, current, and future information.*

This project focused on data-driven prognostics and direct RUL estimation due to availability of CM measurements accompanied by historical failures. The techniques employed were:

i.   *Multiple Linear Regression (MLR) and Polynomial Regression (PR) which belong to trend extrapolation, one of the simplest methods and most commonly used one in industry [17], [23]–[25]*

ii.   *Self-Organising Map (SOM) which belongs to Neural Networks (NN) family, one of the most favoured methods in academia [17]. To the authors' knowledge, SOM has yet to be applied for prognostics as a standalone technique. Also, a RUL estimation variation based on SOM was proposed.*

iii. *K-Nearest Neighbours Regression (KNNR) which belongs to similarity-based prognostics, an emerging trend with great potential [26], [27]. Moreover, a RUL estimation variation based on KNNR was proposed.*

iv.   *An ensemble method averaging each of the aforementioned algorithms' output was proposed.*

These methods were applied to non-uniformly sampled historical valve failure data from an industrial reciprocating compressor, retrieved from a server rather than raw sensor measurements commonly used. Use of actual information addressed a major prognostics challenge: limited works utilising real-life data [7], [16]–[21], demonstrating PHM's applicability and benefits in industry, and its implementation to failure modes that are instantaneous in contrast to slowly time varying ones usually examined.

Principal Components Analysis (PCA) and Statistical Process Control (SPC) were employed to create Hotelling T^2 and Q residuals metrics which, to the authors' knowledge, are used for the first time to reflect degradation process of compressor and employed for RUL estimation. PCA/SPC has found limited application in reciprocating compressors as diagnostics tool. Ahmed et al. [28] used experimental raw sensor data, extracted features, fused them with PCA and performed detection of various faults via SPC. They further enhanced their methodology in [29] by extracting more features and utilising contribution plot of Q metric to identify features associated with faults that can assist identification. Prognostics algorithms were benchmarked while utilising these metrics.

The rest of the paper is organised as follows. Section 2 reviews literature of prognostics methods employed. Section 3 analyses HI creation process and overviews prognostics methods. Section 4 describes data acquisition procedure and evaluation metrics used. Section 5 presents results followed by a discussion. Section 6 contains concluding remarks.

2.   *Prognostics methods literature review*

Trend extrapolation is one of the most preferred prognostics method in industry, being the simplest one, though there are limited published works in literature [17]. Zhao et al. [30] used S-transform, Gaussian pyramid, local binary pattern, PCA and linear discriminant analysis for pre-processing along with MLR for RUL estimation for bearings. Li and Nilkitsaranont [31] employed MLR for prognostics of gas turbine engine during early degradation stage while quadratic regression was used when degradation deteriorated. Alamaniotis et al. [32] applied fuzzy sets and MLR for prognostics of power plant turbine blade. Proposed methodology was superior to simple MLR. MLR has also been used extensively as a benchmarking tool, along with PR. In such works, MLR/PR were used either to compare performance of proposed methodology, usually found inferior [33]–[35], or to compare performance of several algorithms [36], [37]. These works used either experimental [30], [33], [34], [36], [37]or simulated [31], [35] or actual [32] raw sensor data.

SOM has yet to be applied for RUL estimation, though it has been used for data fusion creating a HI, namely the Mean Quantisation Error (MQE), for prognostics purposes [38]–[42], where all works used experimental raw sensor measurements.

Inspiration of implementing SOM for direct RUL estimation was taken by its missing data imputation capabilities and similarity based prognostics. Arima et al. [43] trained several SOMs with missing values being imputed as average of their corresponding weights from their best matching units in each map. Fessant and Midenet [44] used SOM to detect outliers as well as to impute missing data in a real transport survey with artificially inserted missing values. Rustum and Adeloye [45] compared imputation performance of SOM, MLR, and backpropagation NN on water treatment time series, with SOM being superior. Folguera et al. [46] applied SOM to impute artificially inserted missing values in water sample dataset.

In similarity based prognostics, a reference data base is created with historical failures which are compared with an ongoing case via distance analysis. Wang et al. [22], used MLR for fusion, curve fitting for smoothing, and segmented failure trajectories. RUL was estimated based on similar reference RULs by measuring distance of ongoing failure trajectory section with historical ones. Zio and Maio [24] segmented and normalised failure signals. During normal operation, RUL was estimated as Mean Time to Failure (MTTF). After fault detection, RUL was calculated as weighted sum of historical RULs based on fuzzy similarity of current segment and reference ones. They further enhanced their methodology in [47] where RUL was calculated continuously and new estimate was compared with previous ones under assumption of stationarity. In case of no significant change healthy state was considered and RUL was replaced by MTTF. Maio and Zio [25] compared Zio and Maio's technique [24] with Monte Carlo based particle filter where it was shown computationally cheaper. Mosallam et al. [23] implemented symmetrical uncertainty method, PCA and EMD for pre-processing and segmented failure signals. RUL was estimated as most similar historical RUL based on K-nearest neighbour analysis of ongoing segment and reference ones, with discrete Bayesian filter used for uncertainty quantification. They also applied the same methodology in [48], and further enhanced it in [49] by adding GPR in RUL estimation process. Zhang et al. [50] used phase space reconstruction trajectory for pre-processing and segmented failure trajectories. RUL was estimated using weighted average of most similar historical RULs, based on distance analysis of ongoing segment and reference ones. Wang et al. [51] applied MLR for fusion, RVM for offline sparse training, estimated RUL as weighted average of historical RULs based on similarity analysis of ongoing trajectory with reference ones, and quantified uncertainty with uncertainty propagation map. Khelif et al. [52] used MLR for fusion and curve fitting for smoothing. RUL was estimated as weighted sum of most similar historical RULs based on distance analysis of current trajectory and reference ones, with most similar cases being favoured and dissimilar ones being penalised. Li et al. [27] used wavelet packet analysis for pre-processing and applied Zio and Maio's methodology [24] where they compared two membership functions which displayed similar performance. You and Meng [26] segmented historical failures. RUL of current segment was estimated based on weighted RUL of similar historical ones. During similarity analysis, more recent measurements within segment had greater importance. Xue et al. [53] estimated RUL by applying local regression on most similar historical RULs based on fuzzy instance modelling of ongoing failure and reference ones, optimised using evolutionary analysis. Lam et al. [54] applied empirical signal to noise ratio method for pre-processing, PCA for fusion, and kernel regression for smoothing. Similarity of ongoing failure with historical ones was computed using various metrics, while RUL was estimated in several ways according to similarity results. Point estimated RUL via Pearson correlation similarity metric outperformed the rest. These works used either experimental [23], [25]–[27], [50], [51], [53] or simulated [22], [24], [47]–[52], [54] raw sensor data. Similarity based prognostics has been implemented on turbofan engines [22], [48], [49], [51]–[54], fission reactor [24], [47], crack propagation [25], lithium-ion batteries [23], [48], bearings [50], contact resistances of electromagnetic relays [27], and ball grid array solder joints of printed circuit boards [26].

Despite its simplicity, KNNR has found limited applications regarding prognostics. Rezgui et al. [55] combined support vector regression with KNNR for diagnostics and prognostics

of reverse polarity fault. Hu et al. [56] extracted features and used KNNR, optimised by particle swarm optimisation and k-fold cross validation, for RUL estimation of lithium-ion battery. Zhao et al. [57] extracted features, and used KNNR with Dempster-Shafer belief theory for RUL estimation local oscillator from an analogue circuit of a high frequency receiver. The method outperformed NN, fuzzy NN, and particle filtering. These works used either experimental [56] or simulated [55], [57] data. On the other hand, KNNR has found popularity in other fields like forestry [58]–[60] or traffic forecasting [61]–[63].

3. *Prognostics methods overview*
3.1 *Health indicator creation*

In data-driven prognostics, data quality is of paramount importance, affecting RUL estimation accuracy [6], [9]. Hence, it is essential data used reflect degradation process adequately. This can be achieved via HIs that can be either features extracted from signals (mean, skewness, kurtosis, etc.), or one-dimensional metrics created by data fusion requiring all useful information be considered [6], [9]. In this work, PCA with SPC were implemented to construct Hotelling T^2 and Q residuals metrics describing compressor's valve degradation, used for the first time as HIs and RUL estimation inputs.

3.1.1 *Principal Components Analysis (PCA)*

PCA is a dimensionality reduction technique that projects a number of correlated variables in a lower space via a linear transformation, while preserving maximum possible variance within original set, creating a new group of uncorrelated, and orthogonal latent variables [64]. Let X be a $n{\times}p$ data matrix (n: number of measurements, p: number of variables), its PCA transformation is [64]:

$$X = P'T + R, \text{ Equation 1}$$

Where T, the $n{\times}k$ score matrix, is the projection of X from p-dimensional space to k-dimensional, with $k \leq p$. P, the $p{\times}k$ component matrix, is the linear mapping of X to T. R is the $n{\times}p$ reconstruction error matrix. Calculation of principal components can be done with use of singular value decomposition [64].

Selection of appropriate k was done employing Cumulative Percentage of Variance (CPV) [64], where k first components leading to a model capturing a predefined variance percentage are kept. A typical value is 90% [64].

3.1.2 *Statistical Process Control (SPC)*

SPC is used to monitor a process for diagnostics purposes. A univariate process is considered to be healthy when its value lies within some statistical limits decided by control chart used [65]. For multivariate process, SPC assumptions of variable independency are inadequate. Hence, Multivariate Statistical Process Control (MSPC) is introduced, where a single control chart is created using information from all variables. A common tool used to facilitate MSPC is PCA by reducing number of monitored variables and decorrelating them. Some good reviews describing application of PCA and MSPC can be found in [66]–[69].

After PCA model has been created, its scores and residuals can be used for SPC. Control charts employed in this work are Hotelling T^2 and Q residuals, most widely used ones regarding PCA/SPC [66]–[69]. Hotelling metric for score matrix T is [66], [69], [70]:

$$T^2 = \sum_{i=1}^{k} \frac{t_i}{s_i^2}, \text{ Equation 2}$$

With t_i ith principal component scores, s_i^2 its variance, and control limit [66]–[69]:

$$T_a^2 = \frac{k(n^2-1)}{n(n-k)} F_a(k, n-k), \text{ Equation 3}$$

With $F_a(k, n-k)$ the $(100-1)a\%$ *upper critical point of F distribution with k and $n-k$ numbers of freedom.*

Q metric for residual matrix R is [66], [67], [69]:

$$Q = \sum_{i=1}^{n}(x_i - \hat{x}_i)^2, \text{ Equation 4}$$

With $\hat{x}_i$ *reconstructed values of x_i, and control limit [66], [67]:*

$$Q_a = g x_{h,a}^2, \text{ Equation 5}$$

Where $g = \frac{var(R)}{2mean(R)}$, $h = \frac{2(var(R))^2}{var(R)}$, and $x_{h,a}^2$ the $(100-1)a\%$ upper critical point of x^2 distribution with h numbers of freedom.

Metrics created by PCA/SPC were used as HIs for prognostics purpose. Procedure of employing PCA/SPC to create HIs is described in a compact form as follows. In phase I healthy data are centred and scaled to unit variance, and PCA model is created, along with control limits for T^2 and Q. In phase II new data are centred and scaled using healthy means and variances, projected on healthy PCA model calculating their scores and residuals, and their metrics are estimated creating HIs.

3.2 Prognostics methods

As already mentioned, there is lack of literature about prognostics on reciprocating compressors. Ergo, several prognostics methods were compared on valve failure data from an operation industrial compressor.

3.2.1 Multiple Linear Regression (MLR)

MLR belongs to trend extrapolation family being its simplest representation. Let Y be a $n \times 1$ response vector and X a $n \times p$ regressor matrix. MLR is used to predict the dependent variable as linear combination of independent ones [71]:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \text{ Equation 6}$$

With $\beta_0, \beta_1, \dots, \beta_p$ *regression coefficients to be estimated, ε the residuals assumed to be uncorrelated and normally distributed, and $i = 1, \dots, n$. Parameters are calculated utilising least squares algorithm [71]:*

$$\hat{\beta} = (X'X)^{-1}X'Y, \text{ Equation 7}$$

Fit of model on data can be assessed using coefficient of determination R^2 that measures amount of variability captured [71]:

$$R^2 = 1 - \frac{SS_E}{SS_T}, \text{ Equation 8}$$

with $SS_E = \sum E^2$ *[71], and $SS_T = \sum(Y - \bar{Y})^2$ [71]. Another criterion is the adjusted coefficient of determination $R_{adjusted}^2$ [71]:*

$$R_{adjusted}^2 = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}, \text{ Equation 9}$$

Both metrics range from zero indicating bad fit to one indicating perfect fit.

MLR was trained using historical failures and applied for direct RUL estimation, with HIs being independent variables and RUL dependent one.

3.2.2 Polynomial Regression (PR)

PR also belongs to trend extrapolation class. It can be seen as an extension to MLR where predictors are also included in power form. Polynomial order depends on desired power. A second order polynomial for two regressors is [71]:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11} x_{i1}^2 + \beta_2 x_{i2} + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon, \text{ Equation 10}$$

Estimation of coefficients is done as in 3.2.1.

Depending on polynomial order, number of parameters can be significantly large leading to overfitting. Stepwise regression is most widely used selection process for including an optimum number of regressors [71]. It is an iterative procedure where terms are included or removed from the model based on a partial F-test. Considering that f_{in} is F-value for including a term and f_{out} for removing one, for a variable to be included it should be $f \geq f_{in}$ and to be excluded $f \leq f_{out}$ [71]. During the initial step, a model is constructed using only most correlated regressor with the dependent variable as it will have highest f value. The process concludes when no variables can be included or excluded [71], leading to polynomial stepwise regression (PSR). Adequacy of model can be examined using R^2 and $R_{adjusted}^2$ metrics. Prognostics application of this method is the same as for MLR.

3.2.3 Self-Organising Map (SOM)

SOM is a form of NN used for unsupervised learning, introduced by Kohonen [72]. It is employed for clustering, and dimensionality reduction, used to project multidimensional data on a two-dimensional structure resembling a map [38], [41], [42], [44]–[46], [72]–[79]. SOM consists of multidimensional input and competitive output. Let X be a $n \times d$ data matrix. The output represents a grid of M neurons, each with a weight vector $W_i = [w_{i1}, \ldots, w_{id}]$, interconnected via a neighbourhood relation. M can be determined as [42], [45], [72], [75]: $M = 5\sqrt{n}$. Dimensions d_1 and d_2 can be found using two largest eigenvalues of covariance matrix [42], [45], [72], [75]: $d_1/d_2 = \sqrt{e_1/e_2}$.

Training is an iterative process. X is centred and scaled to unit variance, and weight vectors are initialised given random values limited within subspace of e_1 and e_2 [42], [72], [75]. A random sample is presented to the map, and its similarity to every neuron is calculated to identify the Best Matching Unit (BMU) [38], [41], [42], [44]–[46], [72]–[79]. A common similarity metric is Euclidean distance [38], [41], [42], [44]–[46], [72]–[79]:

$$D_{ki} = \|X_k - W_i\| \sqrt{\sum_{j=1}^{d} (x_{kj} - w_{ij})^2}, \text{ Equation 11}$$

With $k = 1, \ldots, n$, $i = 1, \ldots, M$, $j = 1, \ldots, d$, X_k the kth sample, W_i weight vector of ith neuron, and D_{ki} their distance. BMU's and its neighbours' weight vectors are adjusted to better resemble input sample [38], [41], [42], [44]–[46], [72]–[79]:

$$W_i(t + 1) = W_i(t) + \alpha(t) h_{BMUi}(t) D_{ki}(t), \text{ Equation 12}$$

With t current time step, $h_{BMUi}(t)$ neighbourhood function centred at BMU, and $\alpha(t)$ learning rate. A typical neighbourhood function is Gaussian [75]: $h_{BMUi}(t) = e^{-d_{BMUi}^2/2\sigma(t)^2}$ with $\sigma(t)$ neighbourhood radius and d_{BMUi} Euclidean distance between BMU and neuron i on the map lay out. Learning rate is [75]: $\alpha(t) = \alpha_0 (0,005/\alpha_0)^{t/T}$, with α_0 initial rate and T training length. Both $h_{BMUi}(t)$ and $\alpha(t)$ are monotonically decreasing functions as iterations increase.

Fit of map on data can be evaluated using Mean Quantisation Error (MQE) [42], [45], [75]:

$$q_e = \frac{1}{n}\sum_{i=1}^{n}\|X_i - W_{BMUi}\|, \text{ Equation 13}$$

Which is average of Euclidean distance of all input data and their BMUs. Its range is [0, ∞) with zero indicating perfect fit. Another metric used is Topographic Error (TE) [42], [45], [75]:

$$t_e = \frac{1}{n}\sum_{i=1}^{n}u(X_i), \text{ Equation 14}$$

With u a binary variable yielding one if first and second BMUs of X_i are not bordering and zero if they are. Its range is [0, n] with zero indicating perfect fit.

After SOM construction, each neuron is able to recognise inputs that are similar to itself, earning the name of self-organising map. Data with similar patterns are associated with same neurons or their neighbours, preserving topology and relations of measurements.

Although SOM is a form of NN, one of the most favoured prognostics methods in academia [17], it has yet to be used for RUL estimation. Inspiration of utilising SOM in such way was drawn from its imputation capability along with similarity based prognostics. When used for imputation, SOM is constructed using observed measurements. The sample containing missing data is presented to the map and its BMU is determined using distances of its observed values and their corresponding neuron weights. Missing values are imputed as their equivalent BMU weight values. Similarity based prognostics [26], [27] is an emerging trend over recent years. For an ongoing fault, RUL is estimated as sum of historical RULs, weighted based on similarity analysis between current information and historical failures [22], [23], [26], [27], [47]–[49], [51], [52], [80]–[82]. It is noted that as time passes, estimated RUL converges with actual one [26], [47]. Its requirements are [22], [23], [26], [27], [48], [52], [81]: sufficient amount of historical failures, continuous monitoring of information, and information reflecting system degradation through time. It is a simple method as there is no complex algorithm used, making it generic, though is highly affected by data quality [26], [27], [49], [80], [82].

Based on above, an offline SOM is constructed using historical T^2 and Q measurements and their corresponding RUL values. In an online step, new T^2 and Q statistics are presented to the map, and their similarity to every neuron is calculated to identify the BMU. Then RUL is calculated as its equivalent BMU weight values. As with similarity based prognostics, the purpose is to identify similar degradation patterns with historical cases utilising SOM's structure. RUL estimation in this study was carried out by performing pointwise similarity analysis on the T^2 and Q statistics rather than similarity analysis between segments as usually done (section 2). In pointwise similarity analysis, only the latest information was compared with historical samples, since more recent samples contain richer information about degradation [26], [27], [52]. Moreover, RUL is calculated solely based on most similar case. This method shall be denoted as SOM 1.

3.2.4 Proposed variation of SOM based RUL estimation
A variation of RUL estimation process based on SOM is also proposed. Instead of creating a single map from all historical failures, an individual map is trained for each case. For an ongoing fault, its information is presented to each SOM and RUL is calculated as average of imputation result of all maps. This variation shall be denoted as SOM 2.

3.2.5 K-Nearest Neighbours Regression (KNNR)
KNNR is a form of similarity based prognostics, belonging in nonparametric regression family. It estimates the regression function without making any assumptions about underlying relationship of dependent and independent variables [59], [62], [83], [84] by utilising similarities of current sample to historical points for prediction [63]. KKNR is a distribution free, multivariate method that preserves variable relations and local structure within data, easy to

use, fast and computationally cheap [85], but highly affected by amount of historical data available [56].

Let X be a $n \times q$ regressor matrix, Y its $n \times 1$ response vector and u a new sample. Resemblance of new sample's predictors and historical ones is calculated via similarity analysis. Euclidean distance [55], [58], [61]–[63], [85]–[88] is most commonly used similarity metric [56], [59], [85]:

$$d(u, x_i) = \|u - x_i\| = \sqrt{\sum_{j=1}^{q} (u_j - x_{ij})^2}, \text{ Equation 15}$$

with $i = 1, \ldots, n$. u's response value is [56], [59], [85]:

$$y_u = \frac{\sum_{l=1}^{K} w_l y_l}{\sum_{l=1}^{K} w_l}, \text{ Equation 16}$$

With K number of most similar historical points to current sample according to $d(u, x_i)$, w_l and y_l weight and response value of lth neighbour. Hence, response value is weighted sum of response values of K closest historical samples based on their predictor similarities. About weighting there is no straightforward formula and can be done in various ways [84]. Formulation used here was:

$$w_l = 1 - d_l / \sum_{l=1}^{K} d_l. \text{ Equation 17}$$

Optimum K can be found via k-fold cross validation [58], [83]–[87]. Historical data are partitioned into k new sets of approximately equal length. For a range of Ks, a model is trained with k-1 sets, leaving one out for validation estimating an error criterion. This is repeated until all subsets are left out once creating k new models. Mean error for each K is calculated and smallest one yields optimum K [56], [83], [89].

As with SOM, pointwise similarity was used instead of segmented. Furthermore, RUL was estimated using K most similar samples from all historical data, meaning that one failure might have more than one common points with current sample while another might have none. This method shall be denoted as KNNR 1.

3.2.6 Proposed variation of KNNR based RUL estimation

A variation of RUL estimation process based on KNNR is also proposed. As with SOM 2, instead of applying KNNR on all historical data, it is implemented on each historical case. RUL is weighted sum of RULs from each case based on similarity results. In this variation, instead of using K most similar points from each case only most similar one was used. This variation shall be denoted as KNNR 2.

3.2.7 Ensemble method

Output of each prognostics algorithm was also combined via averaging leading to an ensemble method:

$$Ensemble = (MSR + PR + SOM1 + SOM2 + KNNR1 + KNNR2)/6, \text{ Equation 18}$$

The purpose is to improve prognostics results by combining strengths of multiple techniques, refining their results.

4. Data acquisition

4.1 Data preparation

Information employed in this work came from an operational industrial two-stage, four-cylinder, double-acting reciprocating compressor that has been used in various applications (compressing different gasses). The machine is instrumented with sensors collecting both process (temperature, pressure, speed, etc.) and mechanically (bearing vibration, bearing temperature, seal pressure etc.) related measurements, that stream continuously, via internet, to a central location. They are stored, pre-processed, and analysed for CBM purposes. Considering each sensor's sampling frequency, a large volume of data is created every second thus a huge amount of storage is required. To mitigate this issue, a rule set was created deciding which values should be stored, creating non-uniformly sampled sets. Linear interpolation was utilized in the data retrieval tool kit to resample non-uniformly sampled data. The fault mode under study was a valve failure. A ring valve was the defective component with cause of failure: broken valve plate leading to leakage. There were 13 defective cases available that all took place in the same cylinder within a period of one and a half years. Depending on case, the failing valve was either Head End (HE) or Crank End (CE) discharge valve. In all failures, valves were of same type, model, and manufacturer. Failure was denoted as the point when it was deemed as incapable of performing its intended function.

Historical information of 16 temperature measurements, one for each valve (two suction (HE/CE) and two discharge (HE/CE) per cylinder, four cylinders), was extracted from a server with sampling period $T_s = 1s$ ($f_s = 1Hz$). Each case contained roughly two and a half days' worth of data, consisting of both healthy and failing states. Table 1 summaries fault duration of each case (moment of detection until moment of failure). The instantaneous nature is evident as failure occurs in a matter of minutes. Prior to proceeding with analysis, data were scanned for missing values, a common phenomenon in industry, utilising SOM for imputation.

Table 1 Data set specifications

Failure Case	1	2	3	4	5	6	7	8	9	10	11	12	13
Fault Duration (s)	333	119	280	245	125	242	114	233	494	131	246	73	254

In order to mitigate the impact of external factors such as air temperature or rotational speed on the temperature measurements, their ratios were employed for the calculation of HIs. Temperature ratios were calculated for suction and discharge of each cylinder, as follows $T_r = T_{HE}/T_{CE}$. Healthy data from each case were centred and scaled to unit variance, and used to create a PCA model of 3 components ($CPV = 95\%$) while calculating T^2 and Q control limits. Failure data, after centring and scaling, were projected on the model calculating their T^2 and Q metrics creating HIs (Figure 1). Both metrics were divided with their respective statistical limits in order to be comparable.
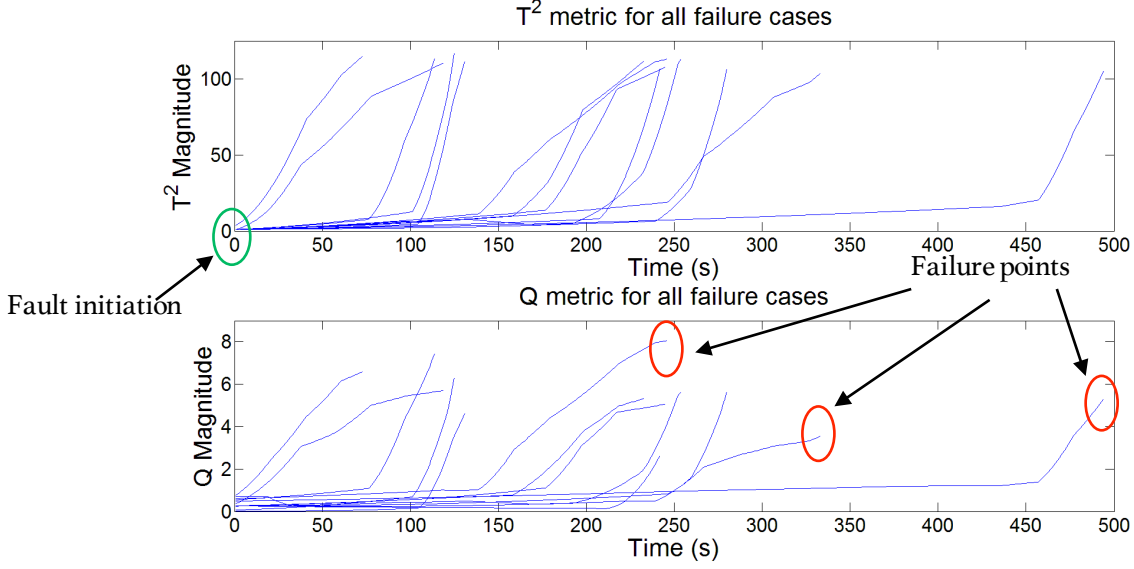
Figure 1 T^2 and Q health indicators

An appropriate HI, needs to be monotonic and encapsulate degradation evolution through time [21], [23], [30], [34], [48], [49]. If this is satisfied, estimated RUL is expected to be accurate [82]. Furthermore, it is desired that HI is of low variability [30], [34], present roughly same value during failure under same failure mode and operating conditions, and have resembling pattern [21], [30]. Figure 1 confirms suitability of both metrics by fulfilling aforementioned perquisites, adequately reflecting fault propagation.

4.2 Prognostics metrics

In order to quantitatively benchmark performance of methods several criteria were used, as there is no universal criterion available yet [36]. The metrics can be separated into two categories: a) accuracy (NMSE, MAPER, CRA) measuring distance between estimated and actual RUL with higher accuracy desired, and b) precision (MAD) measuring error variability with low volatility desired. Let $RUL(t)$ be actual RUL at time t, $t = 1, ..., N$ number of available samples, $\widehat{RUL}(t)$ be estimated RUL, and $\Delta RUL(t) = RUL(t) - \widehat{RUL}(t)$ be difference of actual and estimated RUL. Employed metrics are:

i. Normalised Mean Square Error (NMSE) [90]:

$$NMSE = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\Delta RUL(i)^2}{(RUL(i)-\overline{RUL})^2}, \text{ Equation 19}$$

With $\overline{RUL}$ the mean value of RUL.

ii. Mean Absolute Percentage Error (MAPER) [36]:

$$MAPER = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{100\Delta RUL(i)}{RUL(i)}\right|, \text{ Equation 20}$$

iii. Cumulative Relative Accuracy (CRA) [36]:

$$CRA = \frac{1}{N}\sum_{i=1}^{N}RA(i), \text{ Equation 21}$$

With $RA(i)$ the Relative Accuracy at each time instance [36], [91]:

$$RA(i) = 1 - \frac{|\Delta RUL(i)|}{RUL(i)}, \text{ Equation 22}$$

iv.   Mean absolute deviation [36]:

$$MAD = \frac{1}{N}\sum_{i=1}^{N}|\Delta RUL(i) - median(\Delta RUL(i))|, \text{ Equation 23}$$

NMSE and CRA range in (-∞, 1] with 1 indicating perfect score, while MAPER and MAD range in [0, ∞) with 0 indicating perfect score.

5. Prognostics results
5.1 Application of prognostics methods

RUL was estimated directly with T^2 and Q being independent variables and RUL dependent one. Prognostics methods attempted to model this relationship so that RUL could be calculated accurately. RUL was logarithmically transformed to improve the fit of the models described in section 3. During training, 12 cases were used for model building while the 13th was kept for testing. Results for representative cases 8 and 11 are presented. Training outcome of each method can be found below. All methods were implemented in Matlab [92]–[94].

5.1.1 Multiple linear/ polynomial regression

For PR, third order was maximum order examined. **Table 2** *contains R^2 and $R^2_{adjusted}$ metrics. Both methods have an adequate fit with PR being superior having greater values.*

Table 2 MLR/PR R^2 and $R^2_{adjusted}$ metrics

Algorithm	MLR		PR	
Failure case	8	11	8	11
R^2	0,68	0,67	0,76	0.77
$R^2_{adjusted}$	0,67	0,67	0,76	0,77

5.1.2 Self-organising map

Data were centred and scaled to unity. Maps were constructed using Gaussian neighbourhood function with starting radius $\sigma = \max(d_1, d_2)/4$, an initial learning rate $\alpha_0 = 0.5$, and Euclidean distance. Table 3 and

Table 4 contain MQE and TE metrics for SOM 1 and SOM 2. Both methods yield high accuracy having low metric values.

Table 3 Mean quantisation and topological errors for SOM 1

Failure case	8	11
MQE	0,13	0,12
TE	0,02	0,04

Table 4 Mean quantisation and topological errors for SOM 2

Failure case	1	2	3	4	5	6	7	8	9	10	11	12	13
MQE	0.04	0,09	0,05	0,05	0,06	0,06	0,06	0,04	0,04	0,06	0,05	0,10	0,06
TE	0.31	0,27	0,40	0,29	0,22	0,43	0,16	0,20	0,27	0,21	0,40	0,18	0,19

5.1.3 K-nearest neighbours regression

Data were centred and scaled to unity, and Euclidean distance was used. Table 5 contains optimum K for KNNR 1, selected via 10-fold cross validation ranging from 1 to 200, while for KNNR 2 optimum K was decided a priori as K=1.

Table 5 Optimum K for KNNR 1

Failure case	8	11
k	15	14

5.2 Results

Figure 2 and Figure 3 contain prognostics results for both historical failures, giving a qualitative perspective of each method's performance. X-axis indicates time while y-axis RUL at each time stamp, with $t = 0$ the moment fault was detected (RUL=233 case 8 and 246 case 11, Table 1) and $t = 233$, or $t = 246$, the moment of failure (RUL=0). Graphs consist of a number of lines. Black indicates actual RUL through time, as observed in-situ, and rest correspond to each algorithm's estimations. All methods perform comparably well with best performing being the ensemble technique (magenta line) as it tracks closely RUL evolution in both cases, followed by polynomial regression (continuous blue line), while worst performing seems to be SOM 1 (dashed red line) which demonstrates great variation. KNNR 1 (dashed green line) performs adequately, while KKNR 2 (continuous green line) and SOM 2 (continuous red line) consistently underestimate RUL. It can be noted that all methods converge to actual RUL as time passes.
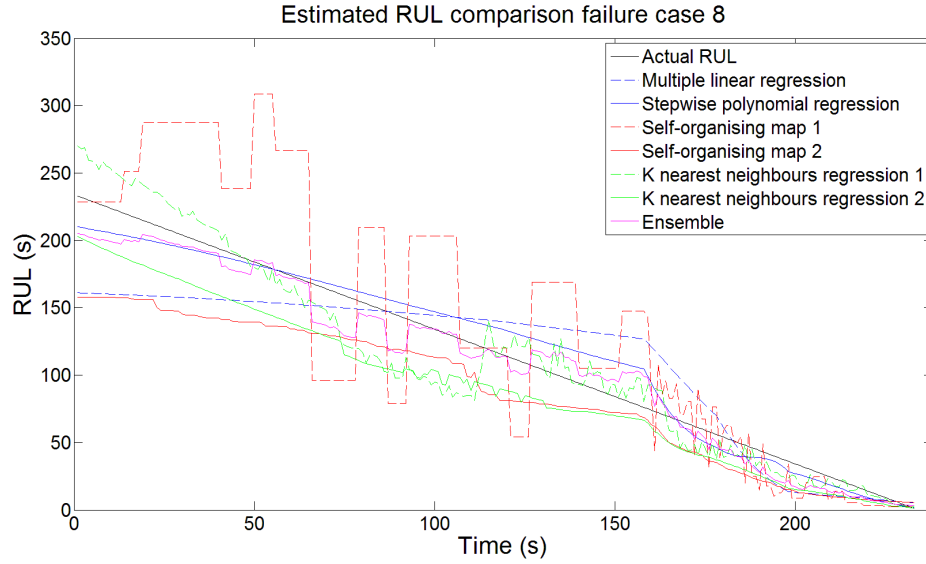

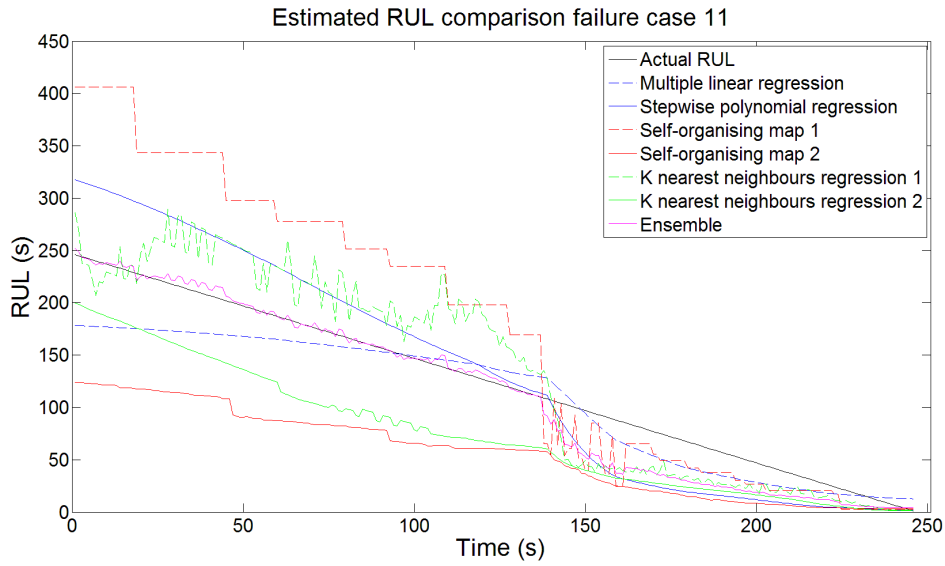*Figure 2 RUL estimation for failure case 8*

13

Figure 3 RUL estimation for failure case 11

Quantitative inspection of methods' performance can be done via metrics found in Table 6. The prognostics horizon for all metrics is from moment of fault detection until failure, meaning all available samples were considered in calculation. From results it is evident that ensemble method consistently outperforms the rest being superior in most metrics for both failures, while in case where another technique prevails, ensemble follows closely. Although PR performs well in case 8, in 11 it is outperformed by others. KNNR 2 displays, lowest variability, an attribute highly desired, followed closely by ensemble one. SOM 1 has lowest accuracy and highest volatility. Overall, quantitative results are in accordance with qualitative ones. Furthermore, results confirmed the claim of lack of universal metrics since the same method might be suitable or not depending on metric used. This calls for more effort to be put towards this direction.

Table 6 Evaluation metrics

Performance metrics	NMSE		MAPER		CRA		MAD	
Failure case	8	11	8	11	8	11	8	11
Method								
MLR	0.77	0.86	30.50	29.39	0.70	0.71	26.10	16.90
PR	**0.96**	0.66	**14.17**	40.14	**0.86**	0.60	11.3	35.61
SOM 1	0.35	-0.39	40.45	51.88	0.60	0.48	41.85	62.05
SOM 2	0.73	0.09	31.43	61.21	0.69	0.39	14.73	26.41
KNNR 1	0.90	0.71	16.94	35.53	0.83	0.65	16.43	33.53
KNNR 2	0.83	0.46	27.96	49.13	0.72	0.51	**9.33**	**14.36**
ENSEMBLE	**0.96**	**0.92**	16.14	**25.17**	0.84	**0.75**	9.35	15.21

Based on prognostics results presented in this section, there are some comments that can be made:

- *PSR and MLR performed similarly well with PSR being superior based on both qualitative and quantitative results, as it could better reflect the complex relationship*

between RUL and HIs by including interaction and higher order terms of HIs, overcoming MLR's rigidness.

- SOM and KNNR displayed similar performance, both belonging to similarity based prognostics family and using the same distance metric (Euclidean) for similarity analysis.

- SOM 1 performed poorly from both accuracy and variability perspective due to considering only most similar case for estimation lacking versatility. SOM 2, KNNR 1, and KNNR 2 performed better since they considered more information during RUL calculation.

- The difference between SOM 2 and both KNNR versions can be attributed to the pooling procedure where SOM 2 averaged RULs while the rest weighted them being more flexible.

- KNNR 1 tended to outperform KNNR 2 indicating that even when considering more than one similar case from the same failure during RUL estimation can increase accuracy. On the other hand, KNNR 2 displayed lower variation hinting that considering each case separately can reduce volatility.

- Ensemble method's performance is highly dependable on individual performance of compromising methods. Its components performed well thus it displayed the best overall performance based on both qualitative and quantitative results. Its output could be seen as refinement of prognostics estimations of its elements.

Importance of HI quality should be noted, as performance of algorithms is also heavily dependent on quality of HIs used since they reflect degradation process. The HIs that were used (T^2 and Q) encapsulated adequately failure evolution confirmed by good results, tracking closely fault propagation through time.

6. Conclusions

In this project, four prognostics techniques (MLR, PR, SOM 1, and KNNR 1), along with two RUL estimation variations (SOM 2 and KNNR 2), and an ensemble method combining aforementioned algorithms' output, were benchmarked using valve failure data from an operational industrial reciprocating compressor. To the authors' knowledge this was the first attempt of RUL estimation on reciprocating compressor valves. Furthermore, use of actual data addressed lack of works regarding implementation of prognostics in industrial applications demonstrating PHM's potency. Moreover, it was the first known implementation of SOM in RUL estimation as standalone prognostics method, and the first time that T^2 and Q metrics were used as HIs and utilised in direct RUL estimation process.

Analysis showed that all methods performed comparably well both in qualitative (graphs) and quantitative (metrics) analysis, with ensemble outperforming the rest by better tracking RUL evolution and having high metric values. SOM 1 performed poorly being less accurate and highly volatile considering only most similar case, while SOM 2, KNNR 1, and KNNR 2 performed closely being all similarity based methods using the same distance metric, with KNNR 1 performing the best. Also, quality of HIs used was deemed satisfactory given good results of techniques, confirming suitability of T^2 and Q metrics to be used as such. Moreover, results demonstrated that all methods were able to cope with instantaneous nature of failure mode under study.

References
[1]   H. P. Bloch, *A Practical Guide to Compressor Technology*, Second. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

[2]   H. P. Bloch and J. J. Hoefner, *Reciprocating Compressors: Operation & Maintenance.* Butterworth-Heinemann, 1996.

[3]   V. T. Tran, F. AlThobiani, and A. Ball, "An approach to fault diagnosis of reciprocating compressor valves using Teager–Kaiser energy operator and deep belief networks," *Expert Syst. Appl.,* vol. 41, no. 9, pp. 4113–4122, Jul. 2014.

[4]   W. A. Griffith and E. B. Flanagan, "Online continuous monitoring of mechanical condition and performance for critical reciprocating compressors," in *30th turbomachinery symposium, Texas: Houston,* 2001.

[5]   Keerqinhu, G. Qi, W.-T. Tsai, Y. Hong, W. Wang, G. Hou, and Z. Zhu, "Fault-Diagnosis for Reciprocating Compressors Using Big Data," *2016 IEEE Second Int. Conf. Big Data Comput. Serv. Appl.,* pp. 72–81, 2016.

[6]   G. Vachtsevanos, F. Lewis, M. Roemer, A. Hess, and B. Wu, *Intelligent Fault Diagnosis and Prognosis for Engineering Systems.* Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.

[7]   A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mech. Syst. Signal Process.,* vol. 20, no. 7, pp. 1483–1510, Oct. 2006.

[8]   A. J. Guillén, A. Crespo, M. Macchi, and J. Gómez, "On the role of Prognostics and Health Management in advanced maintenance systems," *Prod. Plan. Control,* vol. 7287, no. June, pp. 1–14, May 2016.

[9]   J. Yan, *Machinery Prognostics and Prognosis Oriented Maintenance Management.* Singapore: John Wiley & Sons Singapore Pte. Ltd, 2014.

[10]   S. Kadry, *Diagnostics and Prognostics of Engineering Systems.* IGI Global, 2013.

[11]   B.-S. Yang, W.-W. Hwang, D.-J. Kim, and A. Chit Tan, "Condition classification of small reciprocating compressor for refrigerators using artificial neural networks and support vector machines," *Mech. Syst. Signal Process.,* vol. 19, no. 2, pp. 371–390, Mar. 2005.

[12]   C. Annicchiarico, A. Babbini, R. Capitani, and P. Tozzi, "Numerical and Experimental Testing of Composite Rings for Reciprocating Compressor Valves," in *ASME 2013 Pressure Vessels and Piping Conference,* 2013.

[13]   F. Gu, Y. Shao, N. Hu, A. Naid, and A. D. Ball, "Electrical motor current signal analysis using a modified bispectrum for fault diagnosis of downstream mechanical equipment," *Mech. Syst. Signal Process.,* vol. 25, no. 1, pp. 360–372, Jan. 2011.

[14]   H. Cui, L. Zhang, R. Kang, and X. Lan, "Research on fault diagnosis for reciprocating compressor valve using information entropy and SVM method," *J. Loss Prev. Process Ind.,* vol. 22, no. 6, pp. 864–867, Nov. 2009.

[15]   K. Feng, Z. Jiang, W. He, and B. Ma, "A recognition and novelty detection approach based on Curvelet transform, nonlinear PCA and SVM with application to indicator diagram diagnosis," *Expert Syst. Appl.,* vol. 38, no. 10, pp. 12721–12729, Sep. 2011.

[16]   J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications," *Mech. Syst. Signal Process.,* vol. 42, no. 1–2, pp. 314–334, Jan. 2014.

[17]   J. Z. Sikorska, M. Hodkiewicz, and L. Ma, "Prognostic modelling options for remaining useful life estimation by industry," *Mech. Syst. Signal Process.,* vol. 25, no. 5, pp. 1803–1836, Jul. 2011.

[18]   B. Sun, S. Zeng, R. Kang, and M. G. Pecht, "Benefits and challenges of system prognostics," *IEEE Trans. Reliab.,* vol. 61, no. 2, pp. 323–335, 2012.

[19]   A. Heng, S. Zhang, A. C. C. Tan, and J. Mathew, "Rotating machinery prognostics: State of the art, challenges and opportunities," *Mechanical Systems and Signal Processing, vol. 23, no. 3.* pp. 724–739, 2009.

[20]   M. S. Kan, A. C. C. Tan, and J. Mathew, "A review on prognostic techniques for non-stationary and non-linear rotating systems," *Mech. Syst. Signal Process., vol. 62–63,* pp. 1–20, 2015.

[21]   E. Zio, "Prognostics and Health Management of Industrial Equipment," in *Diagnostics and Prognostics of Engineering Systems,* S. Kadry, Ed. IGI Global, 2013, pp. 333–356.

[22]   T. Wang, Jianbo Yu, D. Siegel, and J. Lee, "A similarity-based prognostics approach for Remaining Useful Life estimation of engineered systems," in *2008 International Conference on Prognostics and Health Management,* 2008, pp. 1–6.

[23]   A. Mosallam, K. Medjaher, and N. Zerhouni, "Bayesian approach for remaining useful life prediction," *Chem. Eng. Trans., vol. 33,* pp. 139–144, 2013.

[24]   E. Zio and F. Di Maio, "A data-driven fuzzy approach for predicting the remaining useful life in dynamic failure scenarios of a nuclear system," *Reliab. Eng. Syst. Saf., vol. 95, no. 1,* pp. 49–57, Jan. 2010.

[25]   F. DI MAIO and E. ZIO, "Failure prognostics by a data-driven similarity-based approach," *Int. J. Reliab. Qual. Saf. Eng., vol. 20, no. 1,* Feb. 2013.

[26]   M.-Y. You and G. Meng, "A generalized similarity measure for similarity-based residual life prediction," *Proc. Inst. Mech. Eng. Part E J. Process Mech. Eng., vol. 225, no. 3,* pp. 151–160, Aug. 2011.

[27]   L. L. Li, D. J. Ma, and Z. G. Li, "Residual Useful Life Estimation by a Data-Driven Similarity-Based Approach," *Qual. Reliab. Eng. Int.,* 2016.

[28]   M. Ahmed, F. Gu, and A. D. Ball, "Fault Detection of Reciprocating Compressors using a Model from Principles Component Analysis of Vibrations," *J. Phys. Conf. Ser., vol. 364,* May 2012.

[29]   M. Ahmed, M. Baqqar, F. Gu, and A. D. Ball, "Fault detection and diagnosis using Principal Component Analysis of vibration data from a reciprocating compressor," in *2012 UKACC International Conference on Control,* 2012, pp. 461–466.

[30]   M. Zhao, B. Tang, and Q. Tan, "Bearing remaining useful life estimation based on time–frequency representation and supervised dimensionality reduction," *Measurement, vol. 86,* pp. 41–55, May 2016.

[31]   Y. G. Li and P. Nilkitsaranont, "Gas turbine performance prognostic for condition-based maintenance," *Appl. Energy, vol. 86, no. 10,* pp. 2152–2161, 2009.

[32]   M. Alamaniotis, A. Grelle, and L. H. Tsoukalas, "Regression to fuzziness method for estimation of remaining useful life in power plant components," *Mech. Syst. Signal Process., vol. 48, no. 1–2,* pp. 188–198, 2014.

[33]   Y. Xing, E. W. M. Ma, K. L. Tsui, and M. Pecht, "An ensemble model for predicting the remaining useful performance of lithium-ion batteries," *Microelectron. Reliab., vol. 53, no. 6,* pp. 811–820, 2013.

[34]   T. H. Loutas, D. Roulias, and G. Georgoulas, "Remaining useful life estimation in rolling bearings utilizing data-driven probabilistic E-support vectors regression," *IEEE Trans. Reliab., vol. 62, no. 4,* pp. 821–832, 2013.

[35]   F. Lasheras, P. Nieto, F. de Cos Juez, R. Bayón, and V. Suárez, "A Hybrid PCA-CART-MARS-Based Prognostic Approach of the Remaining Useful Life for Aircraft Engines," *Sensors, vol. 15, no. 3,* pp. 7062–7083, 2015.

[36]   A. Saxena, J. Celaya, B. Saha, S. Saha, and K. Goebel, "Evaluating algorithm performance metrics tailored for prognostics," IEEE Aerosp. Conf. Proc., 2009.

[37]   M. A. A. Wahab, M. M. Hamada, and A. Mohamed, "Artificial neural network and non-linear models for prediction of transformer oil residual operating time," Electr. Power Syst. Res., vol. 81, no. 1, pp. 219–227, 2011.

[38]   S. Hong, Z. Zhou, E. Zio, and W. Wang, "An adaptive method for health trend prediction of rotating bearings," Digit. Signal Process. A Rev. J., vol. 35, pp. 117–123, 2014.

[39]   C. Lu, L. Tao, and H. Fan, "An intelligent approach to machine component health prognostics by utilizing only truncated histories," Mech. Syst. Signal Process., vol. 42, no. 1–2, pp. 300–313, 2014.

[40]   G. Niu and B.-S. Yang, "Intelligent condition monitoring and prognostics system based on data-fusion strategy," Expert Syst. Appl., vol. 37, no. 12, pp. 8831–8840, Dec. 2010.

[41]   J. Yu, "Machine health prognostics using the Bayesian-inference-based probabilistic indication and high-order particle filtering framework," J. Sound Vib., vol. 358, pp. 97–110, Dec. 2015.

[42]   R. Huang, L. Xi, X. Li, C. Richard Liu, H. Qiu, and J. Lee, "Residual life predictions for ball bearings based on self-organizing map and back propagation neural network methods," Mech. Syst. Signal Process., vol. 21, no. 1, pp. 193–207, 2007.

[43]   K. Arima, N. Okada, Y. Tsuji, and K. Kiguchi, "Evaluations of a multiple SOMs method for estimating missing values," in 2014 IEEE/SICE International Symposium on System Integration, 2014, pp. 796–801.

[44]   F. ¸oise Fessant and S. Midenet, "Self-Organising Map for Data Imputation and Correction in Surveys," Neural Comput. Appl., vol. 10, no. 4, pp. 300–310, Apr. 2002.

[45]   R. Rustum and A. J. Adeloye, "Replacing Outliers and Missing Values from Activated Sludge Data Using Kohonen Self-Organizing Map," Journal of Environmental Engineering, vol. 133, no. 9. pp. 909–916, 2007.

[46]   L. Folguera, J. Zupan, D. Cicerone, and J. F. Magallanes, "Self-organizing maps for imputation of missing data in incomplete data matrices," Chemom. Intell. Lab. Syst., vol. 143, pp. 146–151, 2015.

[47]   E. Zio and F. Di Maio, "A fuzzy similarity-based method for failure detection and recovery time estimation," Int. J. Performability Eng., vol. 6, no. 5, pp. 407–424, 2010.

[48]   A. Mosallam, K. Medjaher, and N. Zerhouni, "Data-driven prognostic method based on Bayesian approaches for direct remaining useful life prediction," J. Intell. Manuf., Jun. 2014.

[49]   A. Mosallam, K. Medjaher, and N. Zerhouni, "Component based data-driven prognostics for complex systems: Methodology and applications," in 2015 First International Conference on Reliability Systems Engineering (ICRSE), 2015, vol. 56, pp. 1–7.

[50]   Q. Zhang, P. W.-T. Tse, X. Wan, and G. Xu, "Remaining useful life estimation for mechanical systems based on similarity of phase space trajectory," Expert Syst. Appl., vol. 42, no. 5, pp. 2353–2360, Apr. 2015.

[51]   P. Wang, B. D. Youn, and C. Hu, "A generic probabilistic framework for structural health prognostics and uncertainty management," Mech. Syst. Signal Process., vol. 28, pp. 622–637, 2012.

[52]   R. Khelif, S. Malinowski, B. Chebel-Morello, and N. Zerhouni, "RUL prediction based on a new similarity-instance based approach," in 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE), 2014, pp. 2463–2468.

[53]   F. Xue, P. Bonissone, A. Varma, W. Yan, N. Eklund, and K. Goebel, "An Instance-Based Method for Remaining Useful Life Estimation for Aircraft Engines," *J. Fail. Anal. Prev.*, vol. 8, no. 2, pp. 199–206, Apr. 2008.

[54]   J. Lam, S. Sankararaman, and B. Stewart, "Enhanced Trajectory Based Similarity Prediction with Uncertainty Quantification," in *PHM 2014 - Proceedings of the Annual Conference of the Prognostics and Health Management Society 2014*, 2013, pp. 623–634.

[55]   W. Rezgui, N. K. Mouss, L.-H. Mouss, M. D. Mouss, and M. Benbouzid, "A regression algorithm for the smart prognosis of a reversed polarity fault in a photovoltaic generator," in *2014 First International Conference on Green Energy ICGE 2014*, 2014, pp. 134–138.

[56]   C. Hu, G. Jain, P. Zhang, C. Schmidt, P. Gomadam, and T. Gorka, "Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery," *Appl. Energy*, vol. 129, pp. 49–55, 2014.

[57]   Jianguang Zhao, Hongbo Li, Fanjing Zeng, and Tiefeng Li, "Prognostics of high frequency receiver based on evidential regression," in *Proceedings of the IEEE 2012 Prognostics and System Health Management Conference (PHM-2012 Beijing)*, 2012, no. 0, pp. 1–5.

[58]   G. Chirici, A. Barbati, P. Corona, M. Marchetti, D. Travaglini, F. Maselli, and R. Bertini, "Non-parametric and parametric methods using satellite images for estimating growing stock volume in alpine and Mediterranean forest ecosystems," *Remote Sens. Environ.*, vol. 112, no. 5, pp. 2686–2700, May 2008.

[59]   A. Haara and A. Kangas, "Comparing k nearest neighbours methods and linear regression – is there reason to select one over the other?," *Math. Comput. For. Nat. Sci.*, vol. 4, no. 1, pp. 50–65, 2012.

[60]   R. E. McRoberts, E. O. Tomppo, A. O. Finley, and J. Heikkinen, "Estimating areal means and variances of forest attributes using the k-Nearest Neighbors technique and satellite imagery," *Remote Sens. Environ.*, vol. 111, no. 4, pp. 466–480, Dec. 2007.

[61]   S. Li, Z. Shen, and G. Xiong, "A k-nearest neighbor locally weighted regression method for short-term traffic flow forecasting," in *2012 15th International IEEE Conference on Intelligent Transportation Systems*, 2012, pp. 1596–1601.

[62]   Tao Zhang, Lifang Hu, Zhixin Liu, and Yuejie Zhang, "Nonparametric regression for the short-term traffic flow forecasting," in *2010 International Conference on Mechanic Automation and Control Engineering*, 2010, pp. 2850–2853.

[63]   Z.-W. Yuan and Y.-H. Wang, "Research on K Nearest Neighbor Non-parametric Regression Algorithm Based on KD-Tree and Clustering Analysis," in *2012 Fourth International Conference on Computational and Information Sciences*, 2012, vol. 1, pp. 298–301.

[64]   I. T. Jolliffe, *Principal Component Analysis*, Second. New York: Springer-Verlag, 2002.

[65]   S. Bersimis, S. Psarakis, and J. Panaretos, "Multivariate statistical process control charts: an overview," *Qual. Reliab. Eng. Int.*, vol. 23, no. 5, pp. 517–543, Aug. 2007.

[66]   T. Kourti, "Application of latent variable methods to process control and multivariate statistical process control in industry," *Int. J. Adapt. Control Signal Process.*, vol. 19, no. 4, pp. 213–246, 2005.

[67]   P. Nomikos and J. F. MacGregor, "Multivariate SPC Charts for Monitoring Batch Processes," *Technometrics*, vol. 37, no. 1, pp. 41–59, Feb. 1995.

[68]   B. D. E. Ketelaere, M. I. A. Hubert, and E. Schmitt, "Overview of PCA-Based Statistical

Process-Monitoring Methods for Time-Dependent , High-Dimensional Data," *J. Qual. Technol.*, vol. 47, no. 4, pp. 318–335, 2015.

[69] U. Kruger and L. Xie, *Statistical Monitoring of Complex Multivariate Processes.* Chichester, UK: John Wiley & Sons, Ltd, 2012.

[70] T. Kourti and J. F. J. F. MacGregor, "Process analysis, monitoring and diagnosis, using multivariate projection methods," *Chemom. Intell. Lab. Syst.*, vol. 28, pp. 3–21, 1995.

[71] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, Third. John Wiley & Sons, Inc., 2003.

[72] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1–3, pp. 1–6, 1998.

[73] F. Wu, T. Wang, and J. Lee, "An online adaptive condition-based maintenance method for mechanical systems," *Mech. Syst. Signal Process.*, vol. 24, no. 8, pp. 2985–2995, 2010.

[74] L. F. Gonçalves, J. L. Bosa, T. R. Balen, M. S. Lubaszewski, E. L. Schneider, and R. V. Henriques, "Fault detection, diagnosis and prediction in electrical valves using self-organizing maps," *J. Electron. Test. Theory Appl.*, vol. 27, no. 4, pp. 551–564, 2011.

[75] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, "SOM Toolbox for Matlab 5," 2000.

[76] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[77] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, no. 1, pp. 59–69, 1982.

[78] T. Kohonen, E. Oja, O. Simula, a. Visa, and J. Kangas, "Engineering applications of the self-organizing map," *Proc. IEEE*, vol. 84, no. 10, pp. 1358–1384, 1996.

[79] R. Xu and D. WunschII, "Survey of Clustering Algorithms," *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, May 2005.

[80] G. Niu, F. Qian, and B.-K. Choi, "Bearing life prognosis based on monotonic feature selection and similarity modeling," *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.*, vol. 0, no. 0, pp. 1–11, Sep. 2015.

[81] M. J. Mcghee, G. Galloway, V. M. Catterson, B. Brown, and E. Harrison, "Prognostic Modelling of Valve Degradation within Power Stations," in *PHM 2014 - Proceedings of the Annual Conference of the Prognostics and Health Management Society 2014*, 2013, pp. 70–75.

[82] Fang Qian and Gang Niu, "Remaining useful life prediction using ranking mutual information based monotonic health indicator," in *2015 Prognostics and System Health Management Conference (PHM)*, 2015, pp. 1–5.

[83] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression.* New York, NY: Springer New York, 2002.

[84] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *Am. Stat.*, vol. 46, no. 3, pp. 175–185, Aug. 1992.

[85] J. M. Ver Hoef and H. Temesgen, "A Comparison of the Spatial Linear Model to Nearest Neighbor (k-NN) Methods for Forestry Applications," *PLoS One*, vol. 8, no. 3, p. e59129, Mar. 2013.

[86] X. Tian, Z. Su, E. Chen, Z. Li, C. van der Tol, J. Guo, and Q. He, "Estimation of forest above-ground biomass using multi-parameter remote sensing data over a cold and arid area," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 14, no. 1, pp. 160–168, Feb. 2012.

[87] H. Gu, L. Dai, G. Wu, D. Xu, S. Wang, and H. Wang, "Estimation of forest volumes by

integrating Landsat TM imagery and forest inventory data," *Sci. China Ser. E Technol. Sci.*, vol. 49, no. S1, pp. 54–62, Jun. 2006.

[88] H. Sun, H. Liu, H. Xiao, R. He, and B. Ran, "Short term traffic forecasting using the local linear regression model," *Transp. Res. Rec.*, no. 1836, pp. 143–150, 2003.

[89] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Stat. Surv.*, vol. 4, pp. 40–79, 2010.

[90] L. Ljung, *System Identification Toolbox ^{TM} User ß€^{TM} s Guide*. The MathWorks, Inc., 2015.

[91] A. Saxena, J. Celaya, E. Balaban, K. Goebel, B. Saha, S. Saha, and M. Schwabacher, "Metrics for evaluating performance of prognostic techniques," in *2008 International Conference on Prognostics and Health Management*, 2008, pp. 1–17.

[92] *Statistics and Machine Learning Toolbox User's Guide*. The MathWorks, Inc., 2016.

[93] "SOM Toolbox 2.0." [Online]. Available: http://www.cis.hut.fi/projects/somtoolbox/.

[94] *Bioinformatics Toolbox ^{TM} User ß€^{TM} s Guide*. The MathWorks, Inc., 2015.