Applying multiple regression analysis to adjust operational limits in condition-based maintenance

Ana Cristina Garcia Bicharra[1], Inhaúma Neves Ferraz[1], José Viterbo[1] e Daniel Costa de Paiva[2]

[1]Active Documentation and Design Laboratory (ADDLabs)
Instituto de Computação - Universidade Federal Fluminense (UFF)
{cristina,ferraz,viterbo}@addlabs.uff.br

[2]Departamento de Informática (DAINF)
Universidade Tecnológica Federal do Paraná, campus Ponta Grossa (UTFPR)
professordanielpaiva@gmail.com

Abstract. Condition-based maintenance (CBM) seeks to implement a policy wherein maintenance management decisions are based on the identification of the current condition of monitored machinery. It involves not only collecting data but also comparing them with reference values and, if necessary generating alerts based on preset operational limits. This approach is adopted by a system responsible for monitoring turbomachinery plants in oil platforms, to identify when a machine deserves special attention. With the purpose of extending the functionalities of such system for dynamically adjusting the detection limits and thus improving the precision in setting the appropriate time for maintenance, we proposed an approach based on the identification of clusters of correlated variables and multiple regression analysis. In this paper, we describe our approach and discuss our experience in implementing such functionalities.

Keywords: Condition-based maintenance; Multiple regression analysis; Variable selection

1 Introduction

Maintenance activities fall into two broad categories, which are corrective maintenance and preventive maintenance. Corrective maintenance, also known as breakdown maintenance, is performed as an action to restore the functional capabilities of failed or malfunctioned equipment or systems. It is a reactive approach triggered by the unscheduled abnormal event of an equipment failure. This kind of maintenance policy usually imposes elevated costs due to the high cost of restoring equipment to an operable condition under emergency. There may be extra costs due to secondary damage and safety/health hazards inflicted by the failure and to penalties associated with lost production. On the other hand, preventive maintenance is the approach developed to avoid this kind of waste (Tsang, 1995).

In oil platforms, the equipment used for oil extraction and exploration operates under severe conditions. High pressure, high temperatures, aggressive working conditions, high throughput and long shifts can have a critical effect on any component. In this scenario, the turbomachinery systems are the most sensitive equipment, since any interruption causes total shutdown of platform activity, resulting in extremely high financial cost (Ferraz and Garcia, 2014). As such, to avoid interruption in turbomachinery operation, it is very important to carry out preventive maintenance that is scheduled to occur during circumstances and at times when there is a high degree of control.

There are different approaches to preventive maintenance. Time-based maintenance (TBM), also known as periodic-based maintenance, is the most common approach. Following this approach, maintenance is performed to prevent or retard failures at hard time intervals regardless of other information that may be available when the preset time occurs. Such task also requires an intrusion into the equipment, thereby rendering it out of service until the task is completed (Tsang, 1995). An inadequate TBM strategy, however, may lead to unnecessarily high downtime – if it is carried out earlier than needed –, or accidental breakdown of a machine – if it is performed too late –, in both cases causing loss of money (Pham e Yang, 2009). Hence, maintenance optimization is a topic of great interest to researchers for its significant appeal to the safety and financial aspects involved (Marseguerra *et al*., 2002).

When the system condition can be continuously monitored, a Condition-Based Maintenance (CBM) strategy can be implemented, according to which the decision of maintaining the system is taken dynamically, based on the observed condition of the system (Marseguerra *et al*., 2002). CBM is assessed as the most effective technology that can identify incipient faults before they become critical. While other approaches such as corrective and time-based maintenance have shown to be costly in many applications, CBM enables more accurate planning of maintenance (Pham e Yang, 2009).

CBM seeks to implement a policy wherein maintenance management decisions are based on the identification of the current condition of monitored machinery (Emmanouilidis *et al*., 2006). Nowadays, advanced equipment and sensor technologies provide a great variety of timely data to reveal a machine's condition. Such data has to be analyzed in real-time to allow observing and evaluating the equipment condition in a more timely fashion (Chen and Wu, 2007). Condition monitoring data are very versatile. It can be vibration data, acoustic data, oil analysis data, temperature, pressure, moisture, humidity, weather or environment data, etc (Jardine *et al.*, 2006).

Condition monitoring involves not only collecting such data but also comparing it with reference values and, if necessary generating alerts based on preset operational limits (Niu and Yang, 2010). In general, such operational limits are established based on manufacturers' recommendations, and utility and industry operating experience. However, as machines may operate in different environments, and thus, subject to different work conditions, in some situations such nominal values may not be the most adequate references. Therefore, artificial intelligence techniques may be useful to adjust such operational limits to the specific conditions of a machinery apparatus.

In this work we discuss our experience in a project for implementing new functionalities for a system responsible for monitoring turbomachinery plants in oil platforms. This system collects data from different sensors distributed all over a plant, each representing a variable. The collected signals are compared with a set of lower and upper reference values and when they cross pre-defined limits, the respective equipment may go under maintenance.

Our project aimed at developing new computational algorithms to be aggregated as new functionalities to the original system, introducing a degree of techniques based on artificial intelligence in the analysis of the turbomachine variables. We proposed several approaches with the purpose of dynamically adjusting the detection limits and thus improving the precision in setting the appropriate time for maintenance. In this work we focus on an approach based in the identification of clusters of correlated variables. The correlation coefficients are used to assess the influence of the variables in one another, thus allowing to calculate narrower limits based on this influence. In this paper, we discuss the fundaments for this novel approach and how it was implemented. In the next section, we present the scenario of our project. In Section 3, we present some fundamental concepts involved in our solution. In Section 4, we explain how the approach was implemented. In Section 5, we describe a case study. Finally, in Section 6 we draw our conclusions.

2 Scenario

In this work we discuss our experience in a project for implementing new functionalities for a system that consists in a center for monitoring the health of turbo machinery plants in oil platforms operated by a large energy corporation. The overall monitoring is performed via the remote monitoring of a set of variables collected from several sensors distributed along the plants, applying a Condition-Based Maintenance (CBM) strategy.

In the referred system, abnormal conditions of the machines are detected when any of the variables related with a given machine go above or below some predefined values. Such values define a set of alarms, in which there are 3 levels above (high, very high and critically high), and 3 levels below (low, very low and critically low). Only when the value of a monitored variable crosses one of the defined levels, the operator will check the variable. Because there is no continuous monitoring, many transients are not perceived at the time to issue a warning. The large number of variables that must be visually monitored by the operator causes the scan time to become too long.

As usual, at beginning of operation, such operational limits were set based on manufacturers' recommendations. As the machines operate subject to different work conditions in each plant of each platform, such limits were gradually and manually adjusted along time, based on the operators experience. This option reflects believes of the operators, the thresholds could be adjusted to alert a long time or a little time before the effective operator interference.

The purpose of our project was to develop new computational algorithms to be aggregated as new functionalities to the original system, introducing a degree of techniques based on Artificial Intelligence in the analysis of the turbomachine variables. Such functionalities should make the system capable of making more accurate and faster problem identification, contributing to increase the availability of the platforms main compression and generation systems, thus reducing losses in the oil and gas production. The main purpose was propose algorithms to automatically and dynamically adjust the reference limits for each variable.

In the specification of the new functionalities, a set of functional requirements were proposed by the client. Among them, the following are tackled in this paper:

- R1: The project should provide a functionality to calculate new monitoring limits of a variable, based on the historical values of such variable.
  - R1.1 The user should choose the variables for having new limits calculated;
  - R1.2 The user should decide for the adoption or not of the new limit calculated.
- R2: The project should provide a functionality to calculate new monitoring limits of a variable based on the influence other variables may have on this one.
  - R2.1: The user should define, based on their experience. a set of variables that he believes are strongly correlated. This option is related to time of reaction, some of the users prefer alerts with a comfortable time and others on a limit that they have to interfere;
  - R2.2: Along the process, the user should be capable of defining (or redefining default values) each parameter involved in such calculation.

In order to achieve such purposes, the project team had to surpass a number of challenges. As a first task, it was necessary to understand the behavior of the huge number of variables. In a set of approximately 150 machines, on average 100 variables were monitored by machine, totaling about 15,000 variables. Such large number of variables comprises different plants in different platforms and also different types (such, as external temperature, oil temperature, vibration, etc). Besides, the frequency of data acquisition and data acquisition time-window were two extra parameters to be studied in the pursue of the most adequate data sample conditions. After overcoming such challenges, the implementation of the first requirement was straightforward.

For meeting the second requirement, however, it was necessary to propose an innovative algorithm to identify groups of correlated variables. In Section 4, we describe the technique we proposed to dynamically calculate a more precise set of alarms for each monitored variable, based on the natural interdependency among some of the variables in a plant. As this proposal involves multiple regression analysis and this topic is discussed in the next section.

3 Fundamental concepts

In this section, we discuss some concepts on which this work was based and which are necessary for a clear understanding of the proposed approach, such as multiple regression analysis and variable selection.

3.1 Multiple regression analysis

According to Aiken *et al.* (2003), multiple regression analysis is a general system for examining the relationship of a collection of independent variables to a single dependent variable. Multiple regression analysis provides an assessment of how well a set of independent variables taken as a whole account for the dependent variable. Multiple regression (MR) analysis involves the estimation of a multiple regression equation that summarizes the relationship of a set of predictors to the observed criterion.

According to Møller *et al*. (2005), in multiple linear regression, the response variable Y_i is regressed on k explanatory variables ($X_{i1},...,X_{ik}$) in the model:

$$Y_i = b_0 + b_1X_{i1} + \ldots + b_kX_{ik} + \varepsilon_i$$

where Y is a dependent variable, X_j are the independent variables, b_0 is the regression intercept, b_j are partial regression coefficients (or regression weights) and ε is the residual error. Considering $\widehat{Y}$ the predicted value for the dependent variable, and Y the observed value. In multiple regression, the values of each regression weight b_0, b_1, ... , b_k are chosen so as to minimize the sum of the squared residuals across the participants. That is, the regression weights b_0, b_1, ... , b_k are chosen so that

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 \text{ is minimum.}$$

This criterion is termed ordinary least squares. Multiple regression computed using ordinary least squares produces optimal estimates of the regression weights in that the correlation between the predicted score $\widehat{Y}$ and the observed dependent variable Y is maximized.

Variable Selection.

A crucial problem in building a multiple regression model is the selection of the independent variables that will form the best model (George and McCulloch, 1993). Stepwise regression is a standard procedure for variable selection, which is based on the procedure of sequentially introducing the predictors into the model one at a time. The stepwise regression is classified into three methods: forward selection, backward elimination and stepwise method. The forward selection adds predictors to the model one at a time. In contrast to the forward selection, the backward elimination begins with the full model and successively eliminates one predictor at a time. The stepwise method starts as the forward selection, but at each stage the possibility of deleting a predictor, as backward elimination, is considered (Chong and Jun, 2005).

In these methods, the number of variables retained in the final model is determined by some criteria assumed for inclusion (or exclusion) of variables in a model, such as the covariance index (Edwards, 1985) or the level of significance. Thus, in our approach we based our decision of including or not a variable in a group for regression on the covariance index, as will be further explained in the next section.

4 The proposed approach

Our purpose was to propose a technique to dynamically calculate a more precise set of alarms (variable monitoring limits) for each monitored variable, based on the natural interdependency among some of the variables in a plant. The rationale behind this approach goes as follows:

> *If the value of a given variable Y can be estimated by applying the values of a set of variables X_j in a multiple regression model, the set of limits of Y can be recalculated using the same model and taking as input the set of limits of each variables X_j, thus allowing the definition of tighter limits.*

This is true because, given a multiple regression model that represents this system, if we admit that each variable X_j assumes exactly the same value corresponding to a limit associated with a specific alarm level, the expected value of Y for this situation can be estimated as $\overline{Y}$. As this situation deserves the level of attention associated with that given alarm level, if Y reaches the value $\overline{Y}$, it may indicate a malfunctioning deserving the same level of attention. Thus, if this value is tighter then the previous limit of Y for that alarm level, the value calculated by regression $\overline{Y}$ may be considered as a tighter limit for alarming. We call this the "multivariate limit".

4.1 Multivariate Limit Algorithm

The algorithm for calculating the multivariate limit for a given dependent variable was implemented meeting some functional requirements defined by the users during the specification phase:

- R2.1: At the beginning of the process, the users should define, based on their experience. a set of variables that she or he believes are strongly correlated;
- R2.2: Along the process, the users should be capable of defining (or redefining default values) each parameter of the algorithm.

Therefore, in the process of calculating the multivariate limits, the user has to interacts with the system in several steps to set or change each parameter of the algorithm. Figure 1 illustrates this process, enumerating the eight steps executed and indicating if they are performed by the user or by the system. Each step is explained ahead.

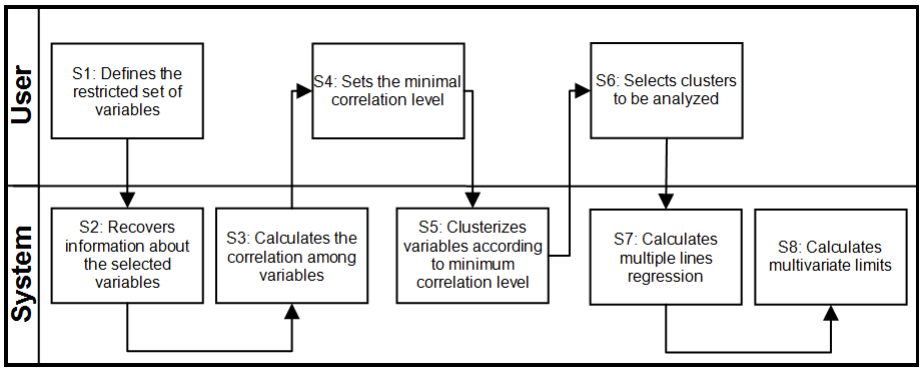

Fig. 1. Steps for calculating the multivariate limits.

Step 1: Defining the restricted set of variables and time period.

At first the user must define a restricted set of variables to be analyzed. This decision is based on the user's own experience, i.e., the user will select a set of variables that he thinks are strongly correlated, so that they will be the ones considered in the further steps for the calculation of the multivariate limits. The user must also define the time window for data samples that will be used in the calculation. This step was implemented to meet functional requirements elicited in the specification phase.

Step 2: Recovering information about the selected variables.

Once the variables were selected and the time period to be considered in the calculation was defined by the user in the previous step, it is necessary to retrieve the respective data sample from the database. The calculation of the correlation between all the variables is carried out using such data set.

Step 3: Calculating the correlation among variables.

As we discussed in Section 2.2, it is necessary to identify in the restricted set of variables which ones should be included in the final model. To assess the correlation among variables, we opted to calculate the covariance between each of the variables of the set. We generate the covariance matrix, a matrix whose element C_{ij} in the position i, j is the covariance between the i^{th} and j^{th} variables. Thus, the covariance matrix gives a hint on how each pair of variables is correlated.

Step 4: Setting the minimal correlation level.

Once again abiding to the elicited requirements, this steps allows the user to set the minimal correlation level that will be used to identify variables that belong to the same cluster, i.e., selecting variables to compose a model. As such, the covariance matrix is presented to the user, so that he can define which is the minimum correlation value for determining the clusters of variables.

Step 5: Clusterizing variables according to minimum correlation level.

To generate the clusters we used an agglomerative algorithm, which takes the correlation value as the measure of similarity. The algorithm gathers the variables in a same cluster until the highest value of this correlation in the matrix is less than the minimum correlation value defined by the user. The generated clusters are exclusive and partial, whereas a variable can only belong to one group and not all variables are grouped together due to the defined minimum level of correlation. As our purpose is to identify a set of strongly correlated variables, the covariance is adopted as criteria to separate clusters of correlated variables that can be represented by a multivariate regression model, according with Algorithm 1:

Algorithm 1. Clustering strongly correlated variables

```
1 Calculate the covariance matriz
2 While there is a Cij > Cmin and matrix dimension > 1
3     Select the biggest Cij
4     Group i^th and j^th variables
5     Recalculate the covariance matrix
```

In this algorithm, C_{min} is the minimal covariance for which we consider that two variables are strongly correlated, and as such, should be included in the same set for defining a multivariate regression model. In each step of the algorithm (line 4) the most correlated variable are linearly combined, forming a new derived variable. Given *n* variables, at the end of the algorithm, we may have 1 to *n* groups (clusters) of variables.

Step 6: Selecting clusters to be analyzed.

The algorithm gathers the variables in a same cluster until the highest value of this correlation in the matrix is less than the minimum correlation value defined by the user. The generated clusters are exclusive and partial, whereas a variable can only belong to one group and not all variables are grouped together due to the defined minimum level of correlation.

Step 7: Calculating the multiple linear model.

Once the clusters are defined, the multiple linear regression is calculated for each group in several iterations, each assuming one variable as the dependent (or target) variable (as discussed in Section 3).

Step 8: Calculating multivariate limits.

Finally, the multivariable restrictive limits are calculated.

5 Case Study

For this case study, we selected a set of ten variables named *Var₁* to *Var₁₀*, for each of which we have collected a hundred sample values. The graph presented in Figure 2 shows the behavior of such variables along the time.

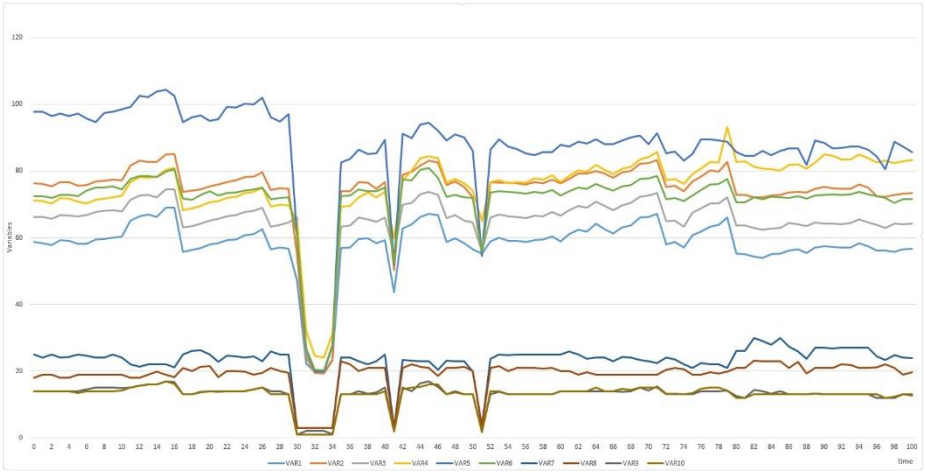

Fig. 2. Graph showing the behavior one of each the ten selected variables.

In this same figure, it is possible to identify that in general there is some interdependency among these variables, as they vary similarly along the time. In fact, applying the clustering algorithm, according to minimum correlation level, previously discussed, the following clusters were formed:

- Cluster 1: Var_1, Var_2, Var_3, Var_6, Var_9, Var_{10}
- Cluster 2: Var_7, Var_8
- Cluster 3: Var_4
- Cluster 4: Var_5

Variables Var_4 and Var_5, formed individual clusters, what means in fact that these variables weren't groupped in any of the clusters, i.e., they don't influence or are influenced by any other in the group.

After the clusters are formed, it is possible to calculate the multivariate limits by applying the multiple linear regression. Figure 3 shows the visualization that is presented for the user indicating the limits for Var_3. The green stripe is limited by the high and low limits, the yellow stripe is limited below by the high limit and above by the very high limit, the orange stripe is limited bellow by the very high limited and above by the critically high limit. Then there is the red stripe above all the other stripes.

The multivariate limits can also be seen in the same figure as two black lines, one above and the other below the time series of the variable. Their respective values are 70.83 and 63.67, which are limits very much tighter than the previous limits indicated in the figure, which are 80 and 13, the borders of the green stripe.

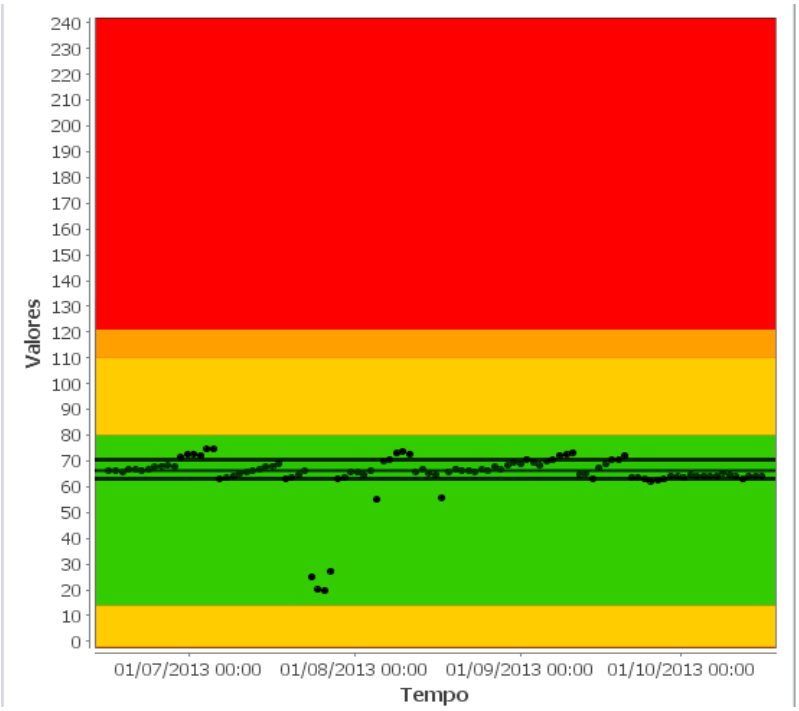

Fig. 3. Visualization of the limits and behavior of variable Var_3

6 Conclusions

In this work, we discussed our experience in a project for implementing new functionalities for a system responsible for monitoring turbomachinery plants in oil platforms. This system collects data from different sensors distributed all over a plant, each representing a variable. Abnormal conditions of the machines are detected when any of the variables related with a given machine go above or below some predefined values. At beginning of operation, such operational limits were set based on manufacturers' recommendations. As the machines operate subject to different work conditions in each plant of each platform, such limits were gradually and manually adjusted along time, based on the operators experience.

The purpose of our project was to develop new computational algorithms to be aggregated as new functionalities to the original system, introducing a degree of techniques based on artificial intelligence in the analysis of the turbomachine variables. The main purpose was propose algorithms to automatically and dynamically adjust the reference limits for each variable reducing the dependence of the operators experiences.

In this work we focused on an approach based in identification of clusters of correlated variables. The correlation coefficients are used to assess the influence of the variables in one another, thus allowing the calculation of narrower limits based on this

influence. We explained the techniques applied and how such functionality was implemented. With a case study we showed how the multivariate limits are calculated and presented to the users, which may use this information to update the monitoring limits.

The implementation of such functionalities met the users requirements and received a satisfactory evaluation from the client, which in fact adopted the new algorithm in the original system. However, the evaluation if the new limits will improve the overall performance of the plant maintenance could not be performed immediately. In fact, it requires a long time of data collection and analysis on the turbomachines' maintenance cycles and will presented in the future.

References

1. Aiken, Leona S., West, Stephen G., and Pitts, Steven C. "Multiple Linear Regression", in Weiner I.B., et al. (eds.) Handbook of psychology. V.02. Research methods in psychology, 483-507. Wiley. (2003)
2. Chen, Argon, and G. S. Wu. "Real-time health prognosis and dynamic preventive maintenance policy for equipment under aging Markovian deterioration". International Journal of Production Research 45.15: 3351-3379. (2007)
3. Chong, I. G., & Jun, C. H. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems, 78(1), 103-112. (2005)
4. Edwards, A. L.. Multiple regression and the analysis of variance and covariance. WH Freeman/Times Books/Henry Holt & Co. (1985)
5. Emmanouilidis, Christos, Erkki Jantunen, and John MacIntyre. "Flexible software for condition monitoring, incorporating novelty detection and diagnostics". Computers in Industry 57.6: 516-527. (2006)
6. Ferraz, Inhaúma Neves, and Ana Cristina Bicharra Garcia. "Turbo machinery failure prognostics". Proceedings of the IEA-AIE The 27th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems. (2014)
7. George, E. I., & McCulloch, R. E.. Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423), 881-889. (1993)
8. Jardine, Andrew KS, Daming Lin, and Dragan Banjevic. "A review on machinery diagnostics and prognostics implementing condition-based maintenance". Mechanical systems and signal processing 20.7: 1483-1510. (2006)
9. Kuo, L., & Mallick, B.. Variable selection for regression models. Sankhyā: The Indian Journal of Statistics, Series B, 65-81. (1998)
10. Marseguerra, Marzio, Enrico Zio, and Luca Podofillini. "Condition-based maintenance optimization by means of genetic algorithms and Monte Carlo simulation". Reliability Engineering & System Safety 77.2: 151-165. (2002)
11. Møller, S. Frosch, Jürgen von Frese, and Rasmus Bro. "Robust methods for multivariate data analysis." Journal of Chemometrics 19.10: 549-563. (2005)
12. Niu, Gang, and Bo-Suk Yang. "Intelligent condition monitoring and prognostics system based on data-fusion strategy". Expert Systems with Applications 37.12: 8831-8840. (2010)

13. Pham, Hong Thom, and Bo-Suk Yang. “Estimation and forecasting of machine health condition using ARMA/GARCH model”. Mechanical Systems and Signal Processing 24.2: 546-558. (2010)
14. Tsang, Albert HC. “Condition-based maintenance: tools and decision making”. Journal of Quality in Maintenance Engineering 1.3: 3-17. (1995)