CHEMICAL ENGINEERING TRANSACTIONS
VOL. 33, 2013
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at: www.aidic.it/cet
Guest Editors: Enrico Zio, Piero Baraldi

ISBN 978-88-95608-24-2; **ISSN** 1974-9791
DOI: 10.3303/CET1333019

Maturation of Detection Functions by Performances Benchmark. Application to a PHM Algorithm

Ouadie Hmad*[,a,c], Jean-Rémi Massé[b], Edith Grall-Maës[c], Pierre Beauseroy[c], Agnès Mathevet[a]

[a]Safran Engineering Services, rond point René Ravaud 77550, Moissy Cramayel, France
[b]Safran Snecma
[c]UMR STMR – LM2S : ICD – Université de Technologie de Troyes, 12 rue Marie Curie CS 42060 10004, Troyes, France
ouadie.hmad@safran-engineering.com

To monitor the start sequence of turbofan engines, a PHM algorithm has been developed. It is called ESC for Engine Start Capability. The anomaly detection phase is performed using a global abnormality criterion. This criterion is obtained by computing the Mahalanobis-distance on standardized residuals (z-scores). In the maturation (performance assessment) perspective of the ESC PHM algorithm, two detection approaches are compared. The first approach (approach 1) is based on the crossing of a threshold determined from the moving average over n starts of the global abnormality criterion. The second approach (approach 2) is based on a minimum number (k) of the global abnormality criterion observations that cross a threshold out of a fixed one (n). The two thresholds are respectively determined in accordance with the two detection approaches. The performances of each approach have been estimated, in terms of probability of detection, P(Detection|Degradation), and No Fault Found ratio (NFF), P(Healthy|Detection). A benchmark of these performances is presented in this paper.

1. Introduction

One of the aims of Prognostics and Health Management (PHM) is to monitor the health state of a system to improve its availability and maintenance schedule. In that perspective, several PHM algorithms have been developed by Safran. The aim of these PHM algorithms is to provide defects detection and isolation and to predict degradations for avoiding failures. It is important to note that the development step is not the ultimate step in the perspective of their implementation on board. The maturation (performance assessment) step follows the development step to make the PHM algorithms operational. The maturation step consists in assessing PHM algorithms performances. It is not possible to know if our PHM algorithms reached some required performances without this step.
A PHM algorithm that monitors the start sequence of a turbofan engine, Engine Start Capability (ESC), is considered in this communication. This algorithm is built as shown in figure 1. A first step of the ESC PHM algorithm consists in taking measurements from turbofan engines (Ausloos et al., 2009). Then eight degradation indicators (Flandrois et al., 2010), are extracted from these measurements for each start. Outliers, those correspond to abnormal starts, are removed thanks to a pre-processing step, to create a healthy dataset. Models are built (thanks to this healthy dataset) for each of the eight extracted degradation indicators. Then, z-scores, those correspond to the normalized residuals, are calculated. A global abnormality criterion is finally computed by applying a Mahalanobis distance on the z-scores (Lacaille, 2009). This is the health status indicator produced by the ESC PHM algorithm. It is used as an input for the detection, localization and prognostic phases to take a global decision about the health status of the start sequence.
In this communication, only the detection phase is considered. It consists in applying a decision threshold to the global abnormality criterion to detect an abnormality or not. Two detection approaches are considered. The first one consists in determining the moving average of the global abnormality criterion

Please cite this article as: Hmad O., Masse J.R., Grall-Maes E., Beauseroy P., Mathevet A., 2013, Maturation of detection functions by performances benchmark. application to a phm algorithm, Chemical Engineering Transactions, 33, 109-114 DOI: 10.3303/CET1333019

over n engine starts and to compare it to a threshold. The second one consists in determining the decision threshold to detect at least k out of n crossing by the global abnormality criterion when a problem occurs.
The aim of this communication is to compare these two approaches thanks to performance benchmark. This study has been realized using operational values for k and n. The impact of each of these parameters on the performances has been studied. Considered performance metrics are described in section 2. Section 3 describes the two detection approaches. Results are compared in section 4 and finally, section 5 concludes the study.

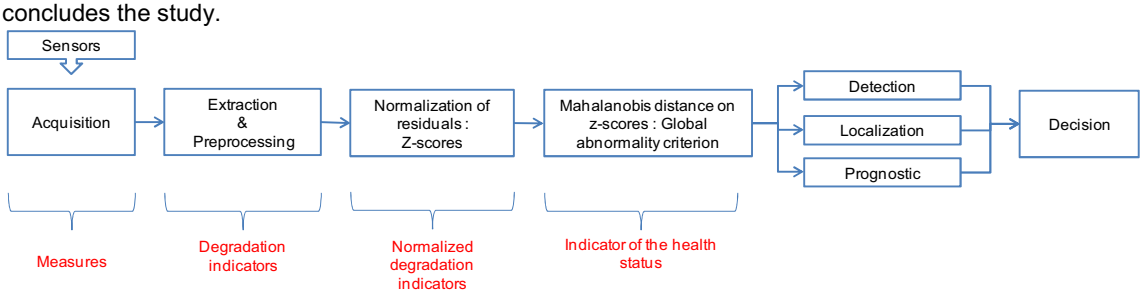

Figure 1: Engine Start Capability (ESC) PHM algorithm description

2. Performance Indicators or metrics

Airlines requirements for PHM algorithms rely on two performance indicators (Hmad et al, 2012). The NFF, No Fault Found ratio, *P(Healthy|Detection)*, and the Probability Of Detection POD, *P(Detection|Degradation)*.
POD is the power of the detector. It is the probability to detect an abnormality given that there is a degradation. Using labeled data, its empirical estimates can be easily obtained. Labels are obtained thanks to generation of simulated degradations. The generation of simulated degradations is based on the analysis of bench test data. The effects of degradations have been identified, quantified and modeled. The model is then used to inject degradation into indicators obtained while the system was running under normal condition (Hmad et al, 2012).
NFF represents the probability that the system is healthy when an abnormality is detected. Applying Bayes theorem, its analytical expression is given by equation (1).

$$NFF = \frac{P(Detection|Healthy)*P(Healthy)}{P(Detection|Healthy)*P(Healthy)+ P(Detection|Degradation)*P(Degradation)} \quad (1)$$

with *P(Detection|Healthy)* the first order error.
Airline companies give the required values of NFF and also specify that the power of the detector have to be close to 1. Thanks to this information it is possible to determine $P(Detection|Healthy)_{target}$ (targeted first order error) corresponding to the required NFF (NFF_{target}) and POD (POD_{target}). $P(Detection|Healthy)_{target}$ can be determined by equation (2) thanks to airline companies requirements.

$$P(Detection|Healthy)_{target} = \frac{NFF_{target}}{1-NFF_{target}} * POD_{target} * \frac{P(Degradation)}{1-P(Degradation)} \quad (2)$$

P(Degradation) may be known through Failure Mode and Effect Analysis (FMEA) or field experience. In our case, $P(Degradation)=10.10^{-6}$.
The aim of the maturation is to evaluate and verify if the required performances are reached. This is done thanks to decision thresholds tuning to respond the airline requirements. If the required performances cannot be reached, the algorithm is not receivable and has to be improved before being used.
To have a more precise analysis of the performances, it has been decided to estimate it for different intensities of premise degradations. Two degradation intensities are considered. Medium (respectively high) degradation intensities which represent medium (respectively short) remaining time before failure.

3. Decision threshold determination

Detection performances of the ESC PHM algorithm depend on the considered decision statistic and a threshold applied to it. To determine the threshold, we chose to estimate the distribution of the decision statistic and then to determine the decision threshold as the value that corresponds to the targeted P(Detection|Healthy) quantile on this distribution. Two decision statistics are considered in 3.1 and 3.2 respectively.

3.1 Approach 1 : detection based on moving average score over n starts

This approach consists in detecting an anomaly if the global abnormality criterion averaged over a rectangular window of size n is higher than a threshold. Computing the moving average over n observations enables to reduce the variance by factor $n^{1/2}$ of the statistic used for decision and thus improve the power of detection.

To estimate the performances of this approach, it is necessary to determine the detection threshold that responds to the airline requirements. To do this, the calculation of $P(Detection|Healthy)_{target}$ according to the airline requirements is the starting point of the detection threshold estimation. This is done thanks to equation (2) with NFF_{target} = 1 %, POD_{target} ≈ 100 % and P (degradation) = 10.10^{-6}.

As the distribution of the decision statistic (averaged global abnormality criterion without degradation) is unknown, a Parzen window (Silverman, 1986) is applied on the decision statistic to determine the detection threshold thanks to equation 3. The choice of the Parzen estimatior has been realised thanks to the study presented in (Hmad et al, 2011).

$$\text{Threshold\_approach\_1} = F^{-1}(1 - P(\text{Detection}|\text{Healthy})_{\text{target}}) \quad (3)$$

with F^{-1}: the inverse cumulative density function of the Parzen estimator applied to the average global abnormality criterion without degradation.

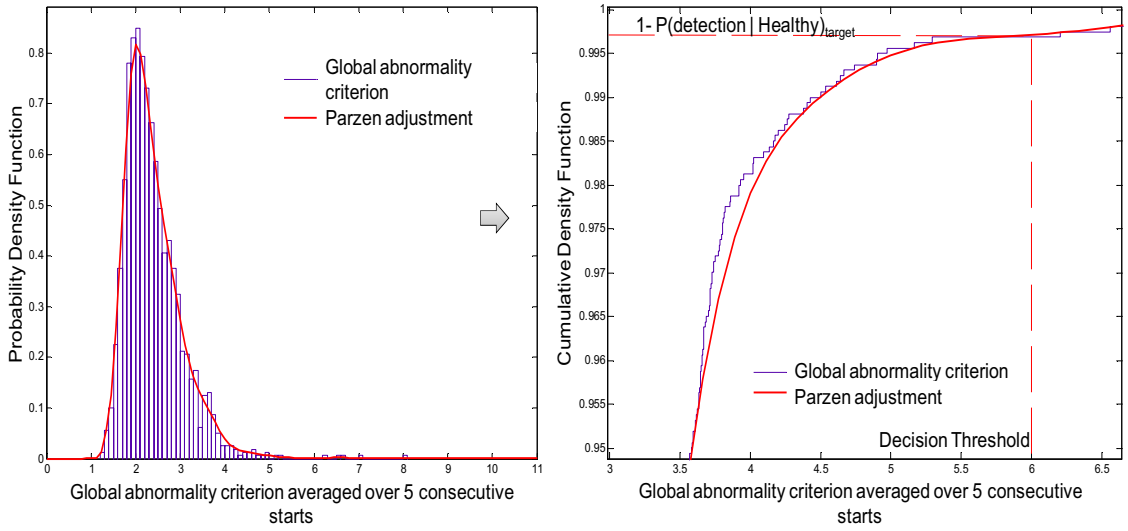

Figure 2: Detection threshold corresponding to the $P(Detection|Healthy)_{target}$ quantile after a Parzen window adjustment on the averaged global abnormality criterion (without degradation) over n=5 starts

3.2 Approach 2 : detection k out of n starts

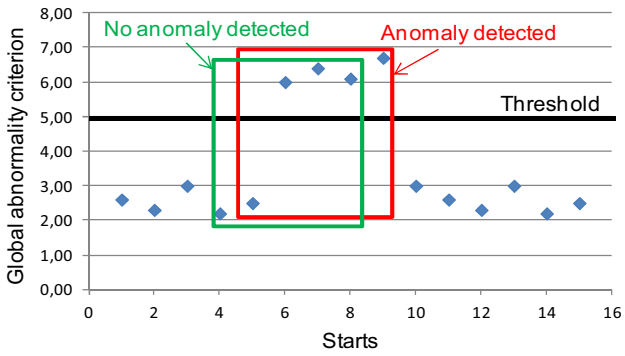

Figure 3: Example of detection k=4 out of n=5 starts. An anomaly is detected if at least k=4 values of the global abnormality criterion cross the threshold in a window of size n=5.

Detection k out of n must inform the user about the persistence of a detection signal. This is done by observing if a predefined proportion (e.g. k=4 and n=5: k/n=80 %) of the observations are detected as abnormal over a sliding window of size n (figure 3). Once again, n is chosen according to operational

constraints. This function detects an anomaly if at least k observations of the global abnormality criterion out of n are higher than a threshold. This threshold is determined after estimating the distribution of the considered decision statistic.
For each observation, the criterion value has a probability p to cross the detection threshold while the system is working fine. One can consider the decision variable as a random variable from a Bernoulli distribution with parameter p – Bernoulli(p). The distribution of n repetition of Bernoulli experiences is a Binomial distribution with parameters n and p – Binomial(n,p). The probability of detection k out of n observations is therefore given by Binomial(n,p).
To estimate the performances of this second approach, it is necessary to determine the first order error that meets the airline requirements. As in previous case, it is done thanks to equation (2) with NFF_{target} = 1%, $POD_{target} \approx 100$ % and P (degradation) = 10.10^{-6}.
Once the targeted first order error $(P(Detection|Healthy)_{target})$ is estimated, we have to determine the p parameter of the Binomial(n,p) under H_0 as the probability that at least k out of n observations cross a threshold with confidence 1-p is inferior to $P(Detection|Healthy)_{target}$. To do this, we must solve the polynomial of degree k of the Binomial distribution function such as $P(Binomiale_{(n,p)} \geq k) \leq P(Detection|Healthy)_{target}$. However, as $P(Binomiale_{(n,p)} \geq k) = Beta_{(k,n-k+1)}(p)$ (Nikulin et al., 2006), p can be estimated thanks to equation (4).

$$p = Beta^{-1}_{(k,n-k+1)}(P(detection|Healthy)_{target}) \quad (4)$$

with n the size of the observation window, k the minimum number of detected observations to confirm the persistence of the detected signal.
Once the parameter p is estimated, we choose the threshold that corresponds to the p quantile on the Parzen window adjustment apply to the global abnormality criterion without degradation.

$$Threshold\_approach\_2 = F^{-1}(1-p) \quad (5)$$

with F^{-1}: the inverse cumulative density function of the Parzen estimator applied to the global abnormality criterion without degradation.

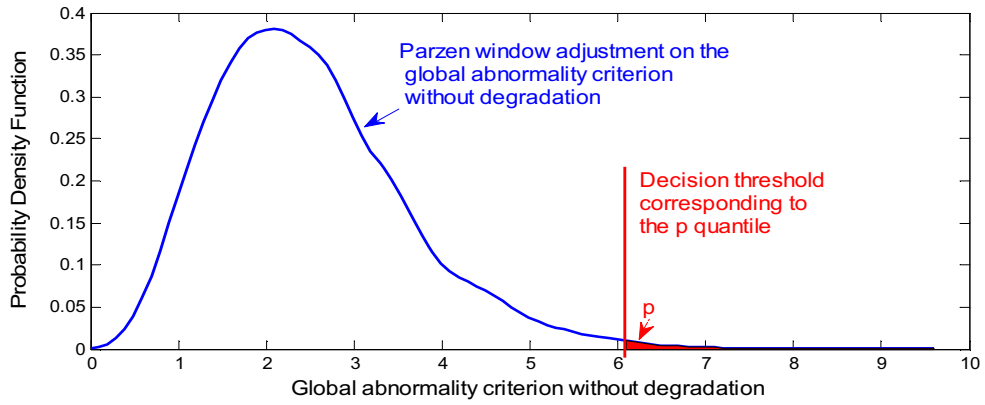

Figure 4: Detection threshold corresponding to the p quantile after a Parzen window adjustment applied to the global abnormality criterion without degradation

4. Result

4.1 Dataset and experimentation design

The performances of both detection approaches have been estimated for the same dataset. This dataset is obtained thanks to measurements extracted from ten turbofan engines.
As the available data are free from degradation, we have generated degradation by simulation. The degradation models have been identified thanks to bench test data with degradations. Two degradation intensities are considered. Medium (respectively high) degradation intensities which represent medium (respectively short) remaining time before failure.
The considered turbofan engines are used for medium range flight. About 5 flights are operating per day. So, the considered window size for the both detection approaches is n = 5 which represents one operation day. Then the performances are estimated for n = 10 which represents two operation days. Several values of k are tested to show its impact on the performances.

4.2 Comparison of the two detection approaches results

The estimated performances for both approaches are shown in this section. The two next paragraphs shows the estimated NFF and POD when the window size is n=5 and n=10 starts.

Concerning detection based on moving average, the threshold is determined tanks to the methodology describe in subsection 3.1.

Concerning detection k out of n. Several confirmation ratios (k/n) have been tested. Then thresholds are determined thanks to the methodology described in subsection 3.2 for each k.

Case of n = 5

Figure 5 shows the estimated NFF (on the right) and POD (on the left) averaged over the 10 engines for each degradation intensity (medium and high). It appears that, whichever the degradation intensity, the performances obtained by approach 1 are better than approach 2 for all k values. This can be explained by the presence of extreme values on the global abnormality criterion that impacts the determined decision threshold values. In fact, we first estimate the probability density function of the decision statistic and then determine the threshold that corresponds to the targeted first order error ($P(Detection|Healthy)_{target}$). As the probability density function of the decision statistic for approach 1 is obtained after average, extreme values are smoothed and have a minor impact on the density estimation and the threshold determination. This is unfortunately not the case of approach 2.

It appears that NFF decreases when k increases. This is an expected result, the alarms which are false decrease when the needed number of the global abnormality criterion values that cross the threshold increases. The increasing in k helps deciding an anomaly just when it is true. This explains the increase of POD when k increases.

The required performances are reached for the medium and high degradation intensities for approach 1. It is also reached by approach 2 for k=5 for high degradation intensities.

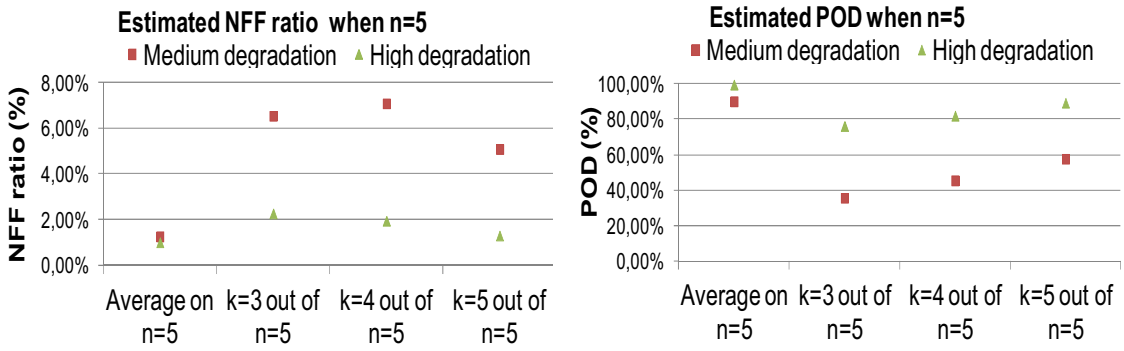

Figure 5: estimated NFF and POD for the both detection approaches for the case of n=5

Case of n = 10

Figure 6 shows the estimated NFF (on the right) and POD (on the left) averaged over the 10 engines for each degradation intensity (medium and high).

As when n=5, the performances with approach 1 are better than with approach 2. The required performances are reached by approach 1 whichever the degradation intensity.

Concerning approach 2, the performances are reached for all k values for high degradation intensities. However, the required performances are reached for k=7 to k=9 for the medium degradation intensities. This can be explained as follow. On the one hand, when k is low, the threshold value is important. So the probability of non detection increases leading to important NFF (according to equation (1)). On the other hand, when k is high (important), the threshold value is low but we want more threshold crossing to confirm the detection. As the needed number of threshold crossing k is important, the NFF decreases.

There are some limits, when k increases, the corresponding threshold decreases. If the threshold value is too low, the probability of false alarm increases. It becomes easy to decide a healthy observation as faulty. It is so necessary to choose carefully an optimal k. This discussion is illustrated by figure 6. It appears, for the considered dataset, that the optimal k value is k=8 when n=10. It is not k=10 because the corresponding threshold value is low and induce some false alarms. Indeed, some healthy observations have been detected as abnormal because of the low threshold value.

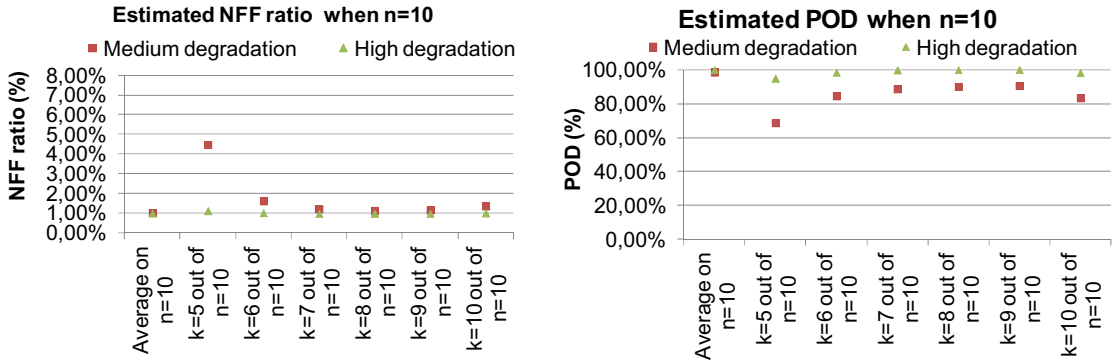

Figure 6: estimated NFF and POD for the both detection approaches for the case of n=10.

5. Conclusion and perspectives

Two detection approaches have been compared thanks to performances benchmark. The first one consists in determining the moving average of the global abnormality criterion over n engine starts and to compare it to a threshold. The second one consists in determining the decision threshold to detect at least k out of n crossing by the global abnormality criterion when a problem occurs.

The considered performance criteria are the NFF ratio (P(Healthy|Detection)) and POD (P(Detection|Degradation)). These criteria correspond to airline requirements. The choice of the performance criteria is important and may have significant impact on the study. If we choose irrelevant performance criteria, we think that the performances meet the requirements but it is not true.

This study shows that, whichever the size of the observation window, approach 1 gives better performance than approach 2. It can be explain by the use of a moving average that reduces the impact of extreme values when estimating the probability density function of the decision statistic to determine a threshold.

In both cases, the performance can be improved by increasing the window size. However, we have to note that increasing in the window size has a direct impact on the reactivity of the detector.

Different values for k have been used. The study shows that there is a tradeoff to be found between NFF and POD to choose the optimal k value for a fixed window size.

A perspective of this study can be to automatically choose the optimal k and n values.

References

Ausloos A., Grall E., Beauseroy P., Grall A., Massé J.R., 2009, Estimation of Monitoring Indicators Using Regression Methods - Application to Turbofan Starting Phase, ESREL conference, September 7-10, 1, 193-200, Prague, Czech Republic.

Coullet J., 1988, Calculation method of the cumulative probability distribution functions of the usual probability laws by direct calculation of integrals (Méthode de calcul des fonctions de répartition des lois usuelles de probabilité par le calcul direct des intégrales), Revue de Statistique Appliquée, 36, 5-18.

Flandrois X., Massé J.R., Ausloos A., Mouton P., Aurousseau C., 2010, Method for monitoring the health status of devices that affect the starting capability of a jet engine, Patent W02010092080 A1 (International).

Hmad O., Grall E., Beauseroy P., Massé J.R., Mathevet A., 2011, A comparison of distribution estimators used to determine a degradation decision threshold for very low first order error, ESREL conference, September 18-22, 345-352, Troyes, France.

Hmad O., Grall E., Beauseroy P., Massé J.R., Mathevet A., 2012, A Maturation Procedure For Prognosis and Health Monitoring Algorithms, PSAM11 & ESREL12 conference, June 25-29, 1-10, Helsinki, Finland.

Lacaille J., 2009, Standardized Failure Signature for a Turbofan Engine, IEEE Aerospace Conference, 1-8, Big Sky, Montana, USA.

Nikulin M., Bagdonavicius V., Huber C., Nikouline V., 2006, Cours de Statistique Mathématique, Université Victor Segalen, Bordeaux 2, 416-417.

Silverman, B. W., 1986, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, UK.