On the Monitoring of Noisy Data as a Multidimensional Shell

Martin GAGNON [1], François LEONARD [1], James MERLEAU [1], Dominique TAPSOBA [1]

[1] Institut de recherche d'Hydro-Québec (IREQ), 1800 boulevard Lionel-Boulet, Varennes, QC, Canada J3X 1S1
gagnon.martin@ireq.ca

Abstract

Based on the idea that multidimensional data is better summarized as a shell rather than a cloud, we have developed a surveillance approach that can detect with high sensitivity behaviour changes in a monitored process and alert the operator. Our methodology uses the time series of a high number of monitored indicators which we cluster together dynamically as a function of operating conditions. These clusters represent groups of similar realizations used to characterize a multidimensional manifold that can be interpolated to assess each new realization of the process behaviour. We evaluated the methodology on the data from a hydroelectric turbine. The event of interest was the loss of the turbine propeller runner cone. The results are good and the approach is able to detect the abnormal behaviour months before the event happened. We are currently looking at larger scale deployment to benchmark the approach's performance.

1 Introduction

One of the primary objectives of monitoring is the early detection of changes in a monitored system or process. Some of these changes can stem from modifications with time of operating conditions (i.e. system input) or changes in the behaviour (system response or output). Usually, changes in the monitored system inputs are intentional hence already known. Generally, we are interested in detecting changes in the system response. The objective of this paper is to account at the same time for changes in the expected behaviour and associated dispersions for any number of monitored inputs in order to detect significant changes while being able to explain in detail the contribution from each of these inputs.

The basis for the proposed approach has been put forward by Léonard and Gauvin, 2013 [1]. They studied the sphere-hardening phenomenon in multidimensional signal projection problems. In fact, this is not a new concept and was first proposed by Shannon, 1949 [2]. While common in communication theory, it seems relatively unknown in the field of equipment and process monitoring. By looking at the cumulative combined random response and measurement noise of a given process over a high enough number of variables in an experiment $\boldsymbol{R}$ repeated many times $(\boldsymbol{R}_i, i = 1, \dots, M)$, a shell will be formed at a given distance $\mu_{\perp S}$ from the expected value $\boldsymbol{S}$ as shown in Figure 1 for the two-dimensional case. This means that looking at a deviation from the shell $(d_i - \mu_{\perp S})$ rather than the deviation from the expected value (d_i) of the noisy process in multidimensional space is more efficient.

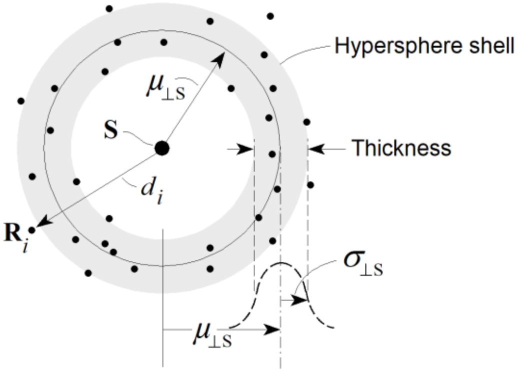


Figure 1: Multidimensional shell from noisy data

However, we cannot do statistics with only one realization of a given event and when monitoring equipment in operation, a difficulty arises since the exact same conditions usually never repeat themselves. The operating conditions of the equipment are always changing hence the need to group similar events together in order to use the sphere hardening concept. Monitored events in similar operating conditions need to be clustered together. Our approach uses a dynamic clustering approach [3] similar to the k-mean methodology [4]. Furthermore, since monitoring cannot be restrained to events that are members of a limited number of known clusters, we need to interpolate in-between the clusters and properly account for the uncertainty induced to prevent false alarms that would lead to unwanted downtime and maintenance costs.

The concepts of shell hardening, clustering and interpolation are used to build the monitoring methodology put forward in this paper. To our knowledge, the approach is novel for situations where many channels or indicators are considered simultaneously. Similar to other monitoring approaches, our methodology starts by modelling the equipment response, then estimates the response under the current operating conditions and finally determines the deviation of the current observed response. Dynamic clustering is used to first model the response while also modelling the dispersion. Then, we use kriging to obtain the behaviour across all possible operating conditions. Finally, we assess the deviation in the obtained multidimensional subspace.

Our paper is structured as follows. We start with the concept of a multidimensional shell resulting from noisy data. Next, the full methodology proposed is explained. Then a study case is presented to illustrate the capability of the proposed methodology. Finally, we discuss some of the limitations of the proposed approach.

2 Noisy data and multidimensional shell

At the root of the proposed methodology is the concept that noisy data over a large number of monitored dimensions generate a multidimensional shell with relative thickness that is inversely proportional to the number of dimensions as proposed by Shannon, 1949 [2], see also [5]. If we consider the information in the form of an equipment signature $\boldsymbol{S}$ that we transmit over N dimensions contaminated by noise, the received signal is given by:

$$\boldsymbol{R} = [S_1 + \varepsilon_1,\ S_2 + \varepsilon_2,\ \dots,\ S_N + \varepsilon_N\ ] = \boldsymbol{S} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{R}$ is the received signal and $\boldsymbol{\varepsilon}$ the random noise vector. However, notice that in the case where the monitored signature is unknown it needs to be estimated using a sample of received signals; relying on the mean as an estimate, one has:

$$\widehat{\boldsymbol{S}} = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{R}_i \tag{2}$$

Furthermore, since the signature transmitted is constantly changing with the operating conditions of the equipment, the signature $\boldsymbol{S}$ is a manifold rather than a single location as shown in Figure 2. This point is discussed further in the following sections.

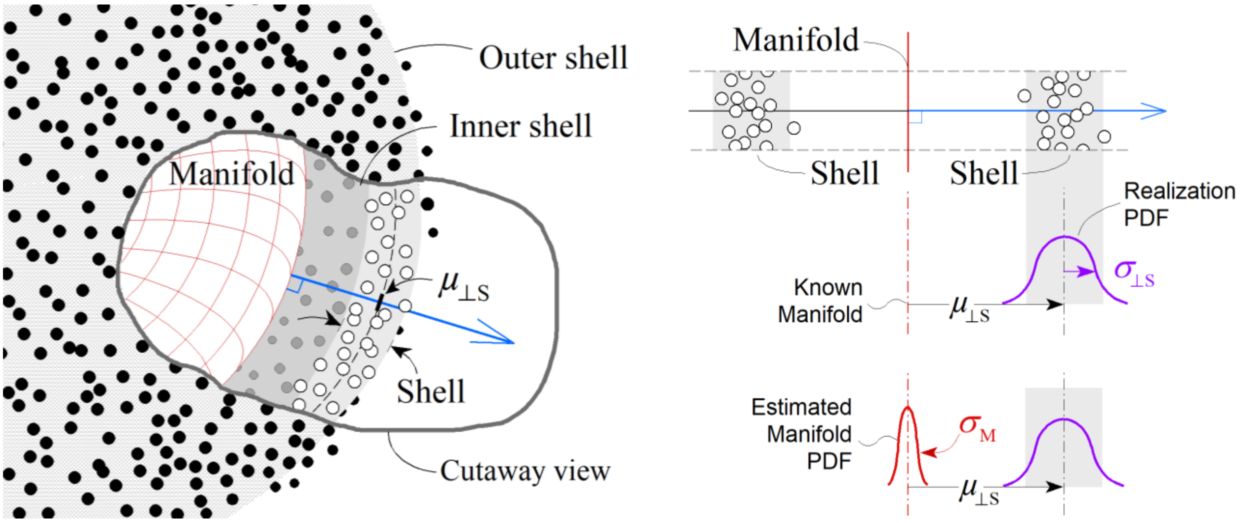


Figure 2: Multidimensional shell and manifold

In this multidimensional space, the distance $\mu_{\perp S}$ between the received signal $\boldsymbol{R}$ and the transmitted signature $\boldsymbol{S}$ at any given location on the manifold can be defined as the average of the realizations at this location as follows:

$$\hat{\mu}_{\perp S} = \frac{1}{M}\sum_{i=1}^{M}\left\|\boldsymbol{R}_i - \hat{\boldsymbol{S}}\right\| = \frac{1}{M}\sum_{i=1}^{M} d_i \qquad (3)$$

where $\hat{\boldsymbol{S}}$ is the estimate of the transmitted signature of interest $\boldsymbol{S}$, a location on the manifold. In a similar manner, the dispersion can be obtained with:

$$\hat{\sigma}^2_{\perp S} = \frac{1}{M}\sum_{i=1}^{M}(d_i - \hat{\mu}_{\perp S})^2 \qquad (4)$$

where $\sigma_{\perp S}$ represents the standard deviation or half-shell thickness. In the present case, where the manifold is also estimated, as shown in Figure 2, the quadratic sum of the manifold dispersion σ^2_{M} and shell dispersion $\sigma^2_{\perp S}$ can be used to assess the likeliness of a given data point $\boldsymbol{R}_i$. Note that the shell wraps around the manifold when N is greater than the number of operating condition indicators.

3 Methodology

In applications, the use of the multidimensional shell concept is not that simple. As shown in Figure 3, realizations need to first be assembled in clusters. Then, we need to interpolate and extrapolate in the hyperspace between clusters. Finally, a fast estimate of the likeliness of a given new realization needs to be made in order for the information about an alert to be relevant in the context of equipment monitoring.

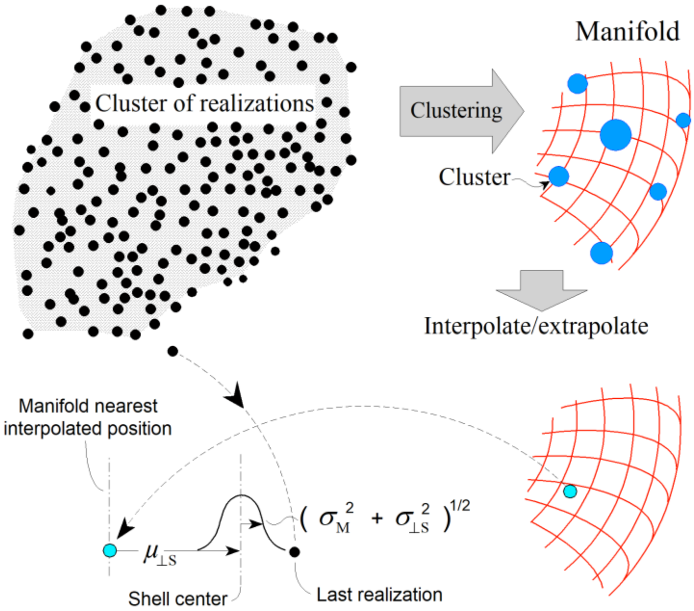


Figure 3: Monitoring methodology

3.1 Data acquisition

The first step is to ensure the data quality before using our algorithm because any error in the data might trigger unwanted alarms. Such preliminary processing is highly dependent on the context of the study and will not be discussed here. For the purpose of this study, let us define the input data as a time series of snapshots $\boldsymbol{X}_i \equiv \boldsymbol{R}_i \cup \boldsymbol{O}_i$ where $\boldsymbol{R}_i$ is a set of response indicators and $\boldsymbol{O}_i$ is a set of operating condition indicators. These $\boldsymbol{X}_i$ cannot be used directly and first need to be formatted and filtered properly to remove unwanted input operating conditions and/or output values. Then, each $\boldsymbol{R}_i$ needs to be normalized in order to ensure that all indicators are represented on similar scales.

3.2 Clustering

Having filtered and normalized the $\boldsymbol{X}_i$ vectors, our goal is to generate clusters of similar $\boldsymbol{O}_i$ to estimate the multidimensional manifold $\boldsymbol{S}$. Initially, for the creation of the clusters, it is important to have a reference dataset of validated history of $\boldsymbol{X}_i$ that cover most operating conditions $\boldsymbol{O}_i$ with corresponding responses $\boldsymbol{R}_i$. Afterwards, with each new $\boldsymbol{X}_i$, it is the dynamic clustering methodology that will determine if a new $\boldsymbol{X}_i$ should be included in the clustering data history. The clustering is dynamic in the sense that the centroid locations are updated every time a new $\boldsymbol{X}_i$ enters the data history. Figure 4 shows the typical process every new $\boldsymbol{X}_i$ is subjected to.

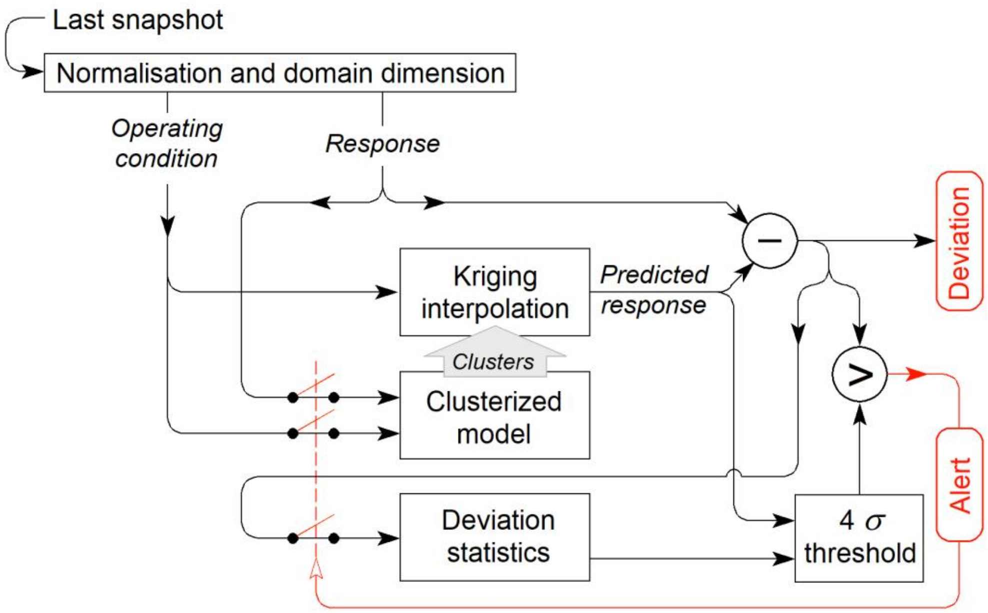


Figure 4: Data snapshots processing flowchart

The clustering process for a new snapshot $\boldsymbol{X}_i$ looks like this:

- If the number of clusters $k < k_{max}$ then the centroid location is $\boldsymbol{Centroid}_k = \boldsymbol{O}_i$ and the population $p_k = 1$ which processes the new snapshot $\boldsymbol{X}_i$
- If the number of cluster $k \geq k_{max}$ then find the smallest distance to an existing cluster c_{min}

$$c_{min} = min_k \|\boldsymbol{Centroid}_k - \boldsymbol{O}_i\| \tag{5}$$

- If $c_{min} > c_{max}$ then merge the two closest clusters together, $\boldsymbol{Centroid}_k = \boldsymbol{O}_i$, $p_k = 1$ and process the new snapshot $\boldsymbol{X}_i$.

$$c_{max} = c\frac{1}{K}\sum_{k=1}^{K} \|\boldsymbol{Centroid}_k - \boldsymbol{O}_i\| \tag{6}$$

- If $c_{min} \leq c_{max}$ then merge snapshot $\boldsymbol{X}_i$ to the nearest cluster, update the population p_k and location $\boldsymbol{Centroid}_k$

Here, in equation 6, c is an arbitrary decision level usually set between 1 and 3. Furthermore, note that any abnormal $\boldsymbol{X}_i$ above the alert thresholds will be rejected and not used for the clustering data history (see section 3.4 for the alert threshold definition). More details about the clustering algorithm used in this study can be found in [5].

3.3 Interpolation

Having clusters of similar data over a large set of different operating conditions $\boldsymbol{O}_k$ with estimated expected response vectors $\boldsymbol{E}(\boldsymbol{O}_k)$ and standard deviation vectors $\boldsymbol{C}(\boldsymbol{O}_k)$ enables us to use a multidimensional interpolator to estimate the response vector for any new operating condition $\boldsymbol{O}_i$. The interpolated $\boldsymbol{E}(\boldsymbol{O}_i)$ with corresponding $\boldsymbol{C}(\boldsymbol{O}_i)$ can then be used to set an alert threshold and assess abnormal behaviour of the monitored system. For simplicity and to limit the computational cost of this interpolation step, the dual kriging formulation was chosen [6]. Kriging is a well-known and extensively used interpolation method. The two traditional formulations which assume a wide sense stationary field, known expected value and variance are the simple kriging formulation and the dual kriging formulation. Implementation of the simple kriging formulation can be either a $O(MN^3) + O(MN)$ or $O(N^3) + O(MN^2)$ process depending on the implementation compared to the dual kriging formulation which is a $O(N^3) + O(MN)$ process for an N

positions over M dimensions problem [6]. Furthermore, because the number of clusters is limited to k_{max} during clustering, we ensure that the numerical cost of the interpolation does not become unmanageable.

3.4 Comparison metric

With the estimated vectors of the expected response $\boldsymbol{E}(\boldsymbol{O}_i) = \boldsymbol{\hat{S}_i}$ (in the notation of section 2) and standard deviation $\boldsymbol{C}(\boldsymbol{O}_i)$ at any new operating condition $\boldsymbol{O}_i$, it is possible to establish an alert threshold above which the behaviour of the monitored process is considered different from past typical responses. Our metric is based on the distance d_i between the snapshot response $\boldsymbol{R}_i$ and the expected response at the operating condition $\boldsymbol{E}(\boldsymbol{O}_i)$:

$$d_i = \|\boldsymbol{R_i} - \boldsymbol{E}(\boldsymbol{O}_i)\| = \sqrt{\sum_{n=1}^{N}\left(r_{i,n} - e(\boldsymbol{O_i})_n\right)^{\boldsymbol{2}}} \tag{7}$$

More precisely, the alert threshold defines the acceptable relative deviation w_i with regards to the expected value of a given ensemble of similar operating condition as shown in Figure 5. However, to have a faster algorithm, we recommend initially using a single average deviation for all the operating conditions. This can be refined as needed. The average distance $\bar{d}$ and relative deviation w_i are expressed as follow:

$$\hat{\mu}_{\perp S} = \bar{d} = \frac{1}{\sum_{j \in \boldsymbol{A}} 1} \sum_{j \in \boldsymbol{A}} d_j \text{ with } j \in \boldsymbol{A} \text{ if } \boldsymbol{O_j} \approx \boldsymbol{O_k} \tag{8}$$

$$w_i = d_i - \bar{d} \tag{9}$$

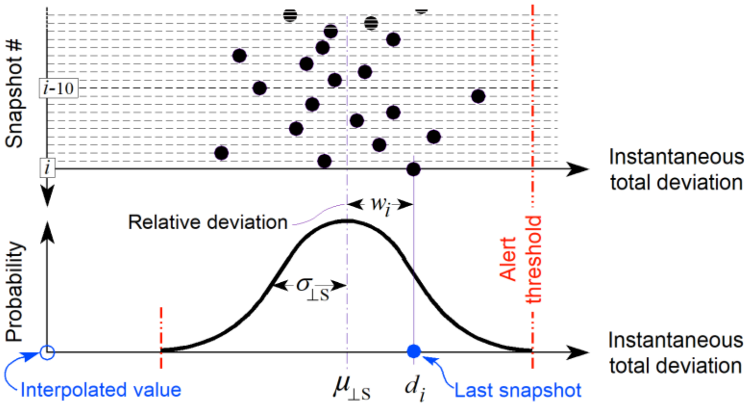


Figure 5: Illustration of the comparison metric

The alert threshold is a function of the global standard deviation of the ensemble of snapshots d_i which accounts for the interpolation standard deviation $\hat{\sigma}_{\mathrm{M}} = \|\boldsymbol{C}(\boldsymbol{O}_i)\|$ and standard deviation of the relative deviation w_i as follows:

$$\sigma' = \sqrt{\hat{\sigma}_{\perp S}^2 + \hat{\sigma}_{\mathrm{M}}^2} \tag{10}$$

$$\hat{\sigma}_{\perp S}^2 = \frac{1}{\sum_{j \in \boldsymbol{A}} 1} \sum_{j \in \boldsymbol{A}} {w_j}^2 \text{ with } j \in \boldsymbol{A} \text{ if } \boldsymbol{O_j} \approx \boldsymbol{O_k} \tag{11}$$

For the case study given next, we have used an alert threshold of $4\sigma'$.

4 Case study

In this paper, we focus on a case study which is the loss of a hydroelectric turbine propeller runner cone. Figure 6 shows a view of the runner before and after the loss of the cone. The cone structure minimizes hydraulic losses and improves efficiency. Without the cone, we should expect reduced efficiency (around 0.6%) and increased vibration due to the vortex rope which is normally dampened by the cone's presence.

Here, the question is not if we can detect the loss of the cone but rather how early we can alert the operator that something is happening to the runner. The earlier we can detect a problem related to the cone, the more time is available for maintenance outage planning which reduces the unexpected downtime. In this case, we have used approximately two years of snapshots history prior to the event and limited the study to the following indicators:

- Mean spiral case pressure ($\in \boldsymbol{R}$)
- Water temperature
- Peak to peak thrust bearing axial acceleration ($\in \boldsymbol{R}$)
- RMS thrust bearing axial acceleration ($\in \boldsymbol{R}$)
- RMS generator guide bearing radial displacement X ($\in \boldsymbol{R}$)
- RMS generator guide bearing radial displacement Y ($\in \boldsymbol{R}$)
- Mean turbine guide bearing radial displacement X ($\in \boldsymbol{R}$)
- RMS turbine guide bearing radial displacement X ($\in \boldsymbol{R}$)
- Peak to peak turbine guide bearing radial displacement Y b($\in \boldsymbol{R}$)
- Mean turbine guide bearing radial displacement Y ($\in \boldsymbol{R}$)
- RMS turbine guide bearing radial displacement Y ($\in \boldsymbol{R}$)
- Excitation tension
- Wicket gates opening ($\in \boldsymbol{O}$)
- Mean power output ($\in \boldsymbol{R}$)

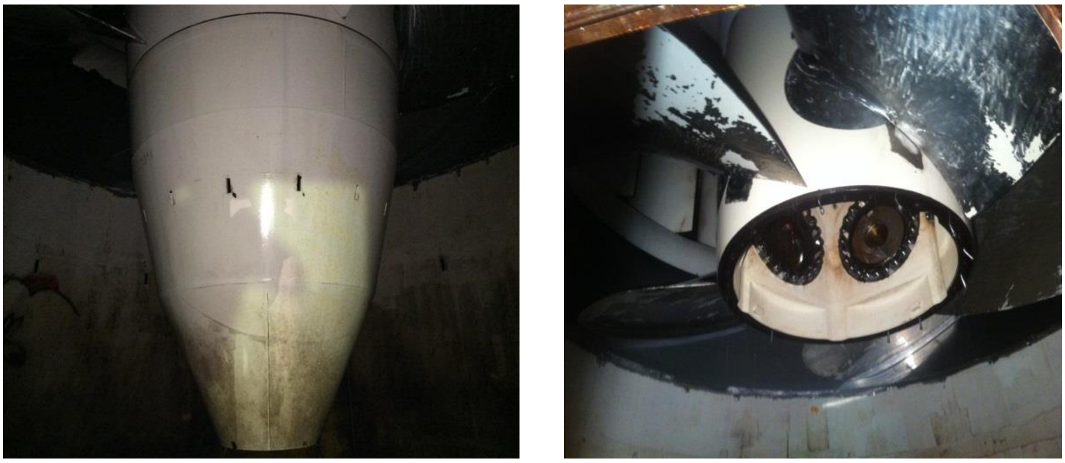


Figure 6: View of the propeller runner cone, before (left) and after (right) the loss

5 Results

With our methodology, we observe seven different phases in the behaviour of the hydroelectric turbine and two types of transient events (see Figure 7). In phase 1, the snapshots serve as reference data for the algorithm to dynamically define the clusters' centroid and dispersion. We observe that the uncertainty bands gradually stabilize. In phase 2, the method is ready to be used to alert the user of unexpected behavior. Notice that the sudden increase in dispersion after phase 1 is artificial and helps highlight the transition between the learning and monitoring regimes. In phase 3, we systematically observe deviations above the alert level. The deviations increase gradually at each subsequent phase until phase 7 is reached and the cone is lost at the end of the snapshots' time history. One can notice some holes in the time history because some snapshots were unsuitable for the methodology and automatically removed during the data acquisition step. Furthermore, two types of transient events are clearly visible in Figure 7. The first, event 8, is the largest of a family of such events that are due to a cooldown period where the monitored unit was stopped. When the unit is restarted, the generator temperature needs to first stabilize then the surrounding structure temperature also needs to stabilize. This generates a transient state that is not a real alert in the sense that the unit is working as expected; this type of event could easily be filtered out if needed. The second, event 9, is simply due to the initialization of the methodology and one can see that the alert bands rapidly stabilize after a sufficient number of data points have been processed.

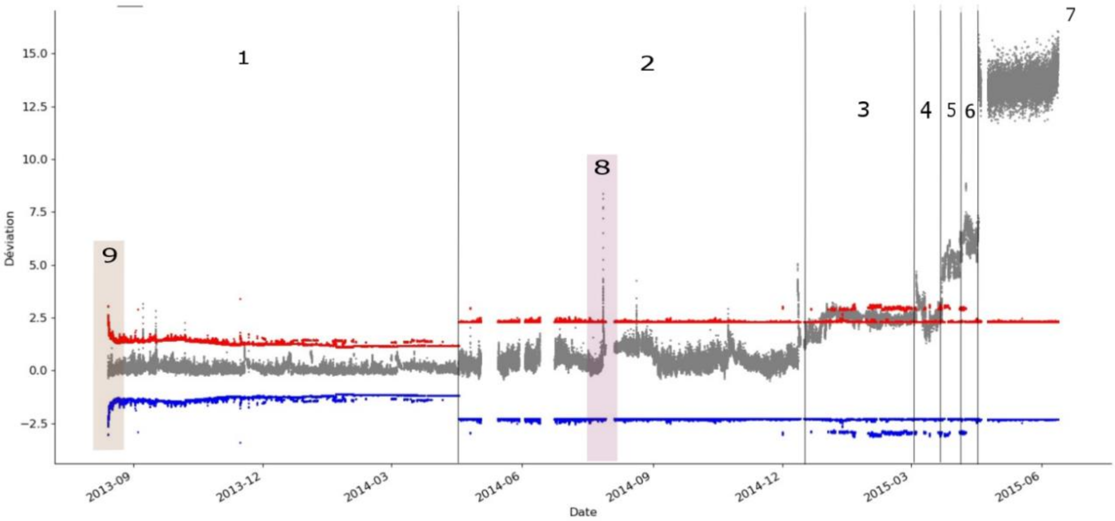
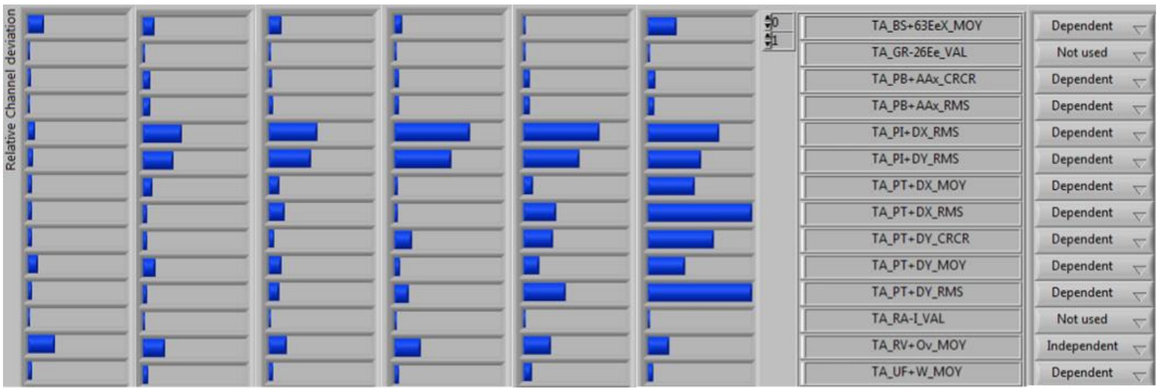


Figure 7: Multidimensional deviation

An advantage of the proposed methodology is that we have access to the contribution of each snapshot response $\boldsymbol{R}_i$ for a given relative deviation w_i. This is of high importance to do a diagnostic of the alert and justify appropriate maintenance outage. In Figure 8, we present an excerpt of the evolution of the individual response contributions for timestamps in each phase from 2 to 7. At first, in phase 3, we observe a highly localised contribution with a slow but gradual increase in contribution from the other response indicators as we move towards phase 5. In subsequent phases, a sudden increase across many of the indicators becomes manifest.



TA_BS+63EeX_MOY	Dependent
TA_GR-26Ee_VAL	Not used
TA_PB+AAx_CRCR	Dependent
TA_PB+AAx_RMS	Dependent
TA_PI+DX_RMS	Dependent
TA_PI+DY_RMS	Dependent
TA_PT+DX_MOY	Dependent
TA_PT+DX_RMS	Dependent
TA_PT+DY_CRCR	Dependent
TA_PT+DY_MOY	Dependent
TA_PT+DY_RMS	Dependent
TA_RA-I_VAL	Not used
TA_RV+Ov_MOY	Independent
TA_UF+W_MOY	Dependent

Figure 8: Screenshots for a given timestamp of the relative multidimensional deviation for each channel. From left to right: phase 2, 3, 4, 5, 6 and 7

In comparison, if we look at the time series of certain selected response indicators, by for example intuitively selecting the ones related to the guide bearing which are the closest to the propeller cone, we get the results shown in Figure 9. The problem becomes noticeable only at the end of March 2015 in phase 5, even if the selected indicators are the closest to the propeller cone. By using a larger ensemble of indicators, the approach proposed in this paper is able to alert the operator of an abnormal behaviour more than three months beforehand in phases 3 and 4.

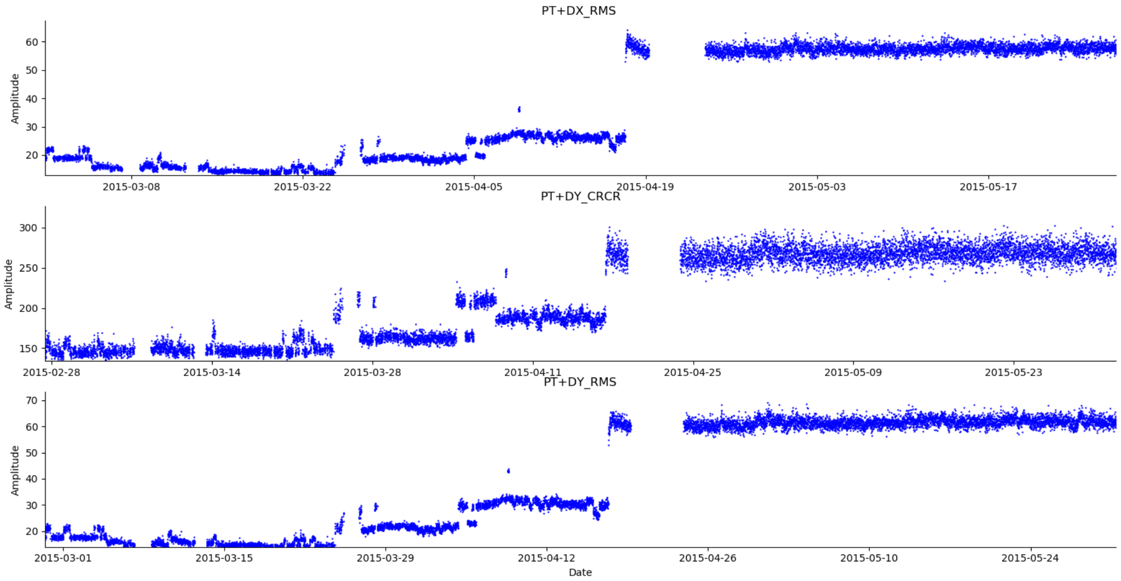


Figure 9: Selected individual responses prior to the propeller runner cone loss

6 Discussions

The case study shows that the sphere hardening principle is applicable to equipment monitoring and performs well in practice. For this application, the method alerts the user months before the actual failure was detected. However, no comparison was made with other mathematical approaches to detect abnormal operational behaviours. Moreover, we simplified the problem by limiting the number of data input to reduce the validation burden. In the proposed approach no difference is made between a change in the behaviour of the sensor and a change in the monitored system. The high sensitivity of the approach to deviations from previous behaviour relies on having data of good quality to avoid false alarms.

Furthermore, no effort was made to optimize the methodology; our initial goal was speed and ease of implementation. In fact, we might be able to optimize the clustering approach to improve kriging performance. Even then, numerical performance might not be the right criteria. The numerical cost of the interpolation might not be a limitation if computational possibilities such as parallelization are considered, given that the necessary infrastructure is becoming more easily available. The same might be true for the alert threshold that could be assessed using conditional numerical simulations.

7 Conclusions

We demonstrated that the data gathered over a group of indicators can be reduced to a single global metric that can be used to monitor equipment behaviour and alert an operator of abnormal equipment behaviour months before an actual failure. The approach is statistically based and is highly sensitive to any deviation from normal past behaviour. An important advantage of the approach is that we can easily track the contribution of each individual indicator and thus explain an alert to establish a diagnostic. We are currently looking at implementing the proposed approach on our hydroelectric turbine fleet in order to benchmark its performance.

References

[1]F. Léonard and M. Gauvin, Trending and pattern recognition using the sphere hardening phenomenon of signal multidimensional projection, Surveillance 7 conference, IUT, Chartres 2013.

[2]C. Shannon, *Communication in the Presence of Noise*, Proceeding of the IRE (1949), Vol. 37, No.1, pp. 10–21.

[3]F. Léonard, *Dynamic clustering of transient signals*, patent US 2014/0100821, priority 2011.

[4]J. B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, in Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability , no. 1, p. 281–297, 1967.

[5]F. Léonard, *Noisy data clusters are hollow*, Joint Statistical Meetings 2015, Aug 2015,
Seattle, United States. JSM 2015 Proceedings, 2015.

[6]F. G. Horowitz, P. Hornby, D. Bone and M. Craig, *Fast Multidimensional Interpolations* Chapter 9 (pp. 53-56) of the 26th Proceedings of the Application of Computers and Operations Research in the Mineral Industry (APCOM 26), R.V. Ramani (ed.), Soc. Mining, Metall., and Explor. (SME), Littleton, Colorado, U.S.A., 538 p., 1996.