



HAL
open science

Unsupervised clustering of vibration signals for identifying anomalous conditions in a nuclear turbine

Piero Baraldi, Francesco Di Maio, Marco Rigamonti, Enrico Zio, Redouane Seraoui

► **To cite this version:**

Piero Baraldi, Francesco Di Maio, Marco Rigamonti, Enrico Zio, Redouane Seraoui. Unsupervised clustering of vibration signals for identifying anomalous conditions in a nuclear turbine. *Journal of Intelligent and Fuzzy Systems*, 2015, 28 (4), pp.1723-1731. 10.3233/ifs-141459 . hal-01265647

HAL Id: hal-01265647

<https://hal.science/hal-01265647>

Submitted on 1 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Unsupervised clustering of vibration signals for identifying anomalous conditions in a nuclear turbine

Piero Baraldi¹, Francesco Di Maio¹, Marco Rigamonti¹, Enrico Zio^{1,2}, and Redouane Seraoui³

¹Energy Department, Politecnico di Milano, Via Ponzio 34/3, 20133 Milan, Italy

piero.baraldi@polimi.it
francesco.dimaio@polimi.it
marcomichael.rigamonti@polimi.it

²Chair on Systems Science and the Energetic Challenge, Ecole Centrale Paris and Supelec, France

enrico.zio@ecp.fr

³EDF-R&D\STEP Simulation et Traitement de l'information pour l'exploitation des systèmes de production,

6 quai Watier, Chatou Cedex, France

redouane.seraoui@edf.fr

ABSTRACT

We consider a real industrial case concerning 148 shut-down multidimensional transients of a nuclear power plant (NPP) turbine. The objective is to identify groups of transients with similar functional behaviors, and distinguish transients with peculiar behaviors which can be representative of anomalous conditions in the turbine. This objective is pursued by analyzing 7 vibration signals referred to the turbine shaft. The novelty of the work consists in transforming the signals into the “turbine speed-domain” for aligning them according to the turbine speed, so as to easily recognize outlier transients and then performing a fuzzy similarity analysis based on pointwise differences. Spectral analysis and Fuzzy C-Means (FMC) clustering are applied to identify the turbine anomalous conditions.

1. CONTEXT OF THE WORK AND OBJECTIVE

In general terms, the methods for assessing the health condition of equipment, detect the onset of abnormal events and classify them in different types can be divided into two categories: model-based and pattern recognition methods (Venkatasubramanian, Rengaswamy, Yin & Kavuri, 2003; Sheng & Rovnyak, 2004; Sandidzadeh & Dehghani, 2013). Model-based methods need expert specific domain knowledge for building the models, while pattern recognition methods are empirically built to fit data (Zio, 2007).

In several practical cases the complexity of the phenomena involved and the highly non-linear interrelationships between the causes that determine the equipment behavior and the signal evolutions are such that it is difficult to

develop model-based methods. For this reason, this work focuses on the development of a new pattern recognition method that consists in identifying different peculiar signal evolution behaviors due to different types of faults, which are “a priori” unknown. Therefore, the problem is formulated as one of unsupervised classification aimed at partitioning the vectors of measured signal values into homogeneous clusters so that those belonging to the same cluster are very similar to each other, and then assigning the clusters of similar behavior to the different fault types (Bocaniala, Sa Da Costa & Palade, 2004).

The development of the method is carried out within the context of fault diagnosis of the steam turbine of a Nuclear Power Plant (NPP) and using real data. The objective is to identify groups of signals with similar functional behaviors from 148 shut-down transients. Once the groups are identified, we can distinguish anomalous transients, representative of anomalous conditions in the turbine.

The application considers 7 signals regarding the turbine shaft vibrations. Vibration monitoring has been widely recognized as an effective tool for the detection and diagnosis of incipient failures of turbines, because a vibration signature contains a good amount of information which, if properly interpreted, can reveal the running condition of the machine (Betta, Liguori, Paolillo & Pietrosanto, 2002). Usually, transient data are misaligned, i.e., amplified and/or delayed, thus challenging the comparison between transients (Secchi, Sangalli, Vantini & Vitelli, 2007).

Therefore, we originally propose to preprocess data for signal alignment and outlier identification by changing the vibration signals domain from the “time” to the “turbine speed” domain. By so doing, every vibration measurement

corresponds to only one turbine speed value and, thus, the vibration values can be easily compared for the identification of anomalous turbine behavior at a given speed. A fuzzy similarity measure is computed for comparing the transients by considering the pointwise difference between the vibration signals values (Baraldi, Di Maio & Zio, 2013). Based on the obtained similarity matrix, a similarity graph (von Luxburg, 2007) is built, in which each vertex represents a trajectory and the weight associated to the edge connecting two vertices is the value of (fuzzy) similarity between the two corresponding trajectories. Spectral analysis techniques are finally applied in order to find an optimal partition of the graph (von Luxburg, 2007), i.e., groups of transients.

The rest of the paper is organized as follows: Section 2 shows the preliminary signal alignment and outliers identification; Section 3 illustrates the proposed operative procedure for transient clustering; Section 4 presents and discusses the obtained results; finally, in Section 5 some conclusions and remarks are drawn.

2. DATA PREPROCESSING

2.1. Transients alignment

The dataset consists of $N=148$ shut-down multidimensional transients of $Z=7$ turbine shaft vibration signals. In order to overcome problems connected with the possibility of misalignment between transients, the considered signals have been transformed from the “time domain” to the “turbine speed domain”. The rationale behind this choice is that a particular magnitude of vibration is normally directly associated to a precise turbine speed value and, in case of abnormal conditions, a different vibration value at a given speed should allow the analysts identifying a signature of an anomalous functional behavior (Roemer & Kacprzyński, 2000).

A preliminary analysis of the available data has shown that the dataset is composed by misaligned transients with different initial values of the turbine speed and oscillations around the steady-state operational conditions. Figure 1 shows three transients characterized by different behaviors: the turbine speed rises from $t=500$ to $t=600$ minutes in the transient represented by the continuous line, whereas it is stationary until $t=2000$ minutes in the transient represented by the dotted line, and it oscillates between $t=2000$ and $t=3000$ minutes in that represented by the dash-dotted line.

In order to accommodate for transients with different initial values and to avoid problems due to the operative oscillations, the transformation of the vibration signals from the “time domain” to the “turbine speed domain” has been performed by considering only the time interval where the turbine speed decreases from 1300 rpm to 500 rpm (e.g., in Figure 1, the part of the signals that are shown in bold line).

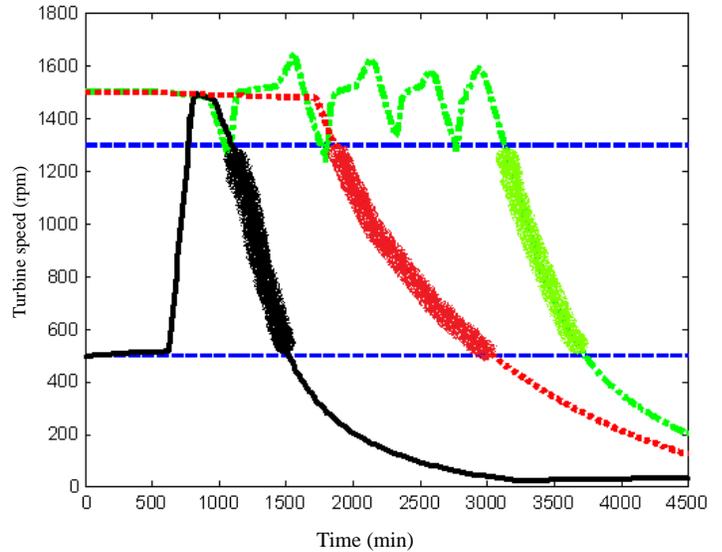


Figure 1. Time evolution of the signal “turbine speed” for three different transients

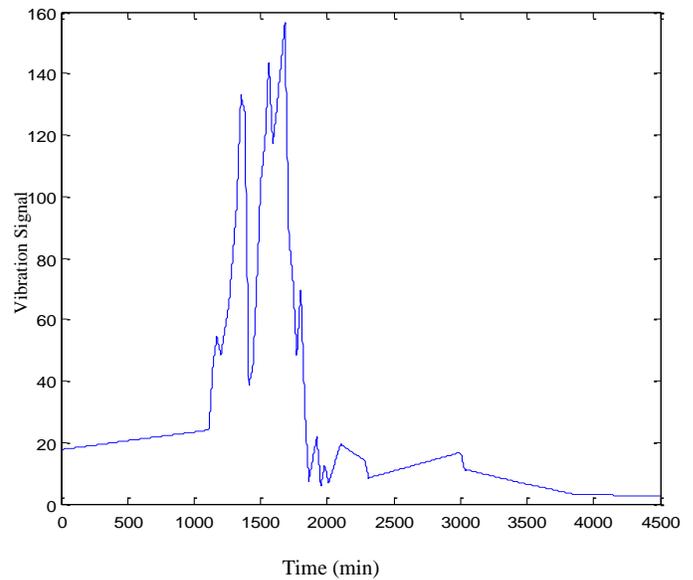


Figure 2. Time evolution of the vibration signal “shaft vibration level 3” in transient 58

The number of measurements for each signal in the “turbine speed domain” is therefore equal to 800. Now, we associate a vibration signal value (Figure 2) to the corresponding turbine speed value and map the considered signals into the “turbine speed domain”. Figure 3 sketches the proposed alignment methodology.

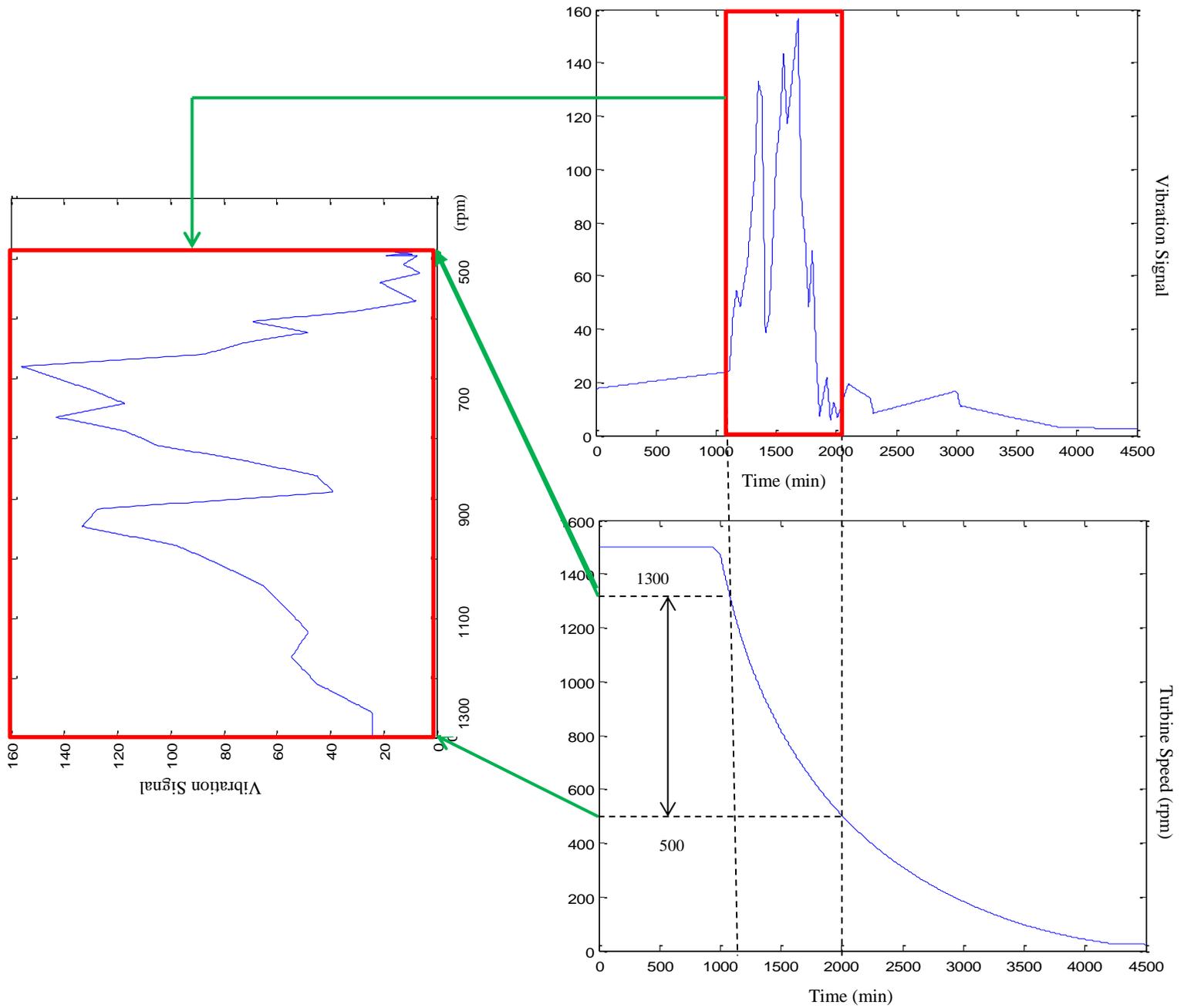


Figure 3. Sketch of the proposed transformation of the signals from the “time domain” to the “turbine speed domain”

After mapping the vibration signals into the “turbine speed domain”, all the 7 vibrational signals have been normalized. The rationale behind the normalization process is to avoid that the computation of the pointwise similarity is affected by different magnitude and scale of the signals (Angstenberger, 2001).

The normalization of a generic signal y is performed by identifying the minimum and maximum values, y^m and y^M , of the signal in all the 148 transients and then by applying:

$$y^n = 0.2 + \frac{y - y^m}{y^M - y^m} \cdot 0.6 \quad (1)$$

Notice that the obtained normalized values of a signal are in the range [0.2 0.8] which has been preferred to the range [0 1] in order to avoid possible numerical problems caused by the presence of 0 values.

Figure 4 shows a normalized vibration signal for transient 58: it has to be pointed out that its maximum normalized value is 0.5 since the original raw signal reaches its maximum in another transient.

By so doing, the pointwise similarity that will be considered in the following will not be affected by different magnitudes/scales of the considered signals (Angstenberger, 2001).

2.2. Outlier identification

The similarity between trajectories may be hidden by the presence of transients whose behavior greatly differs from the others, i.e., the outliers and endangers the discriminating capability (Frigui & Krishnapuram, 1996).

To identify the outlier transients, we consider a similarity measure which determines the degree of closeness of two trajectories i and j with reference to the pointwise difference between the values of the matrices \bar{Y}^i [7, 800] and \bar{Y}^j [7, 800], representative of transient i and transient j , where $y_{zt}^{i/j}$ is the t -th measurement, $t=1, \dots, 800$ (from 1300 to 500 rpm), of the z -th signal, $z=1, \dots, 7$, of matrix \bar{Y}^i and \bar{Y}^j , respectively (Zio & Di Maio, 2013). The pointwise difference δ_{ij} between the $7 \times 800 = 5600$ values of trajectories \bar{Y}^i and \bar{Y}^j is defined by:

$$\delta_{ij} = \sqrt{\sum_{z=1}^7 \sum_{t=1}^{800} (y_{zt}^i - y_{zt}^j)^2} \quad (2)$$

To evaluate the similarity, we use the bell-shaped, “approximately zero” fuzzy set (FS) (Zio & Di Maio, 2010):

$$\mu_{ij} = e^{-\frac{\delta_{ij}^2}{\sigma^2}} \quad (3)$$

where σ can be set by the analyst to shape the desired interpretation of similarity into the fuzzy set: the smaller the value of σ , the narrower the fuzzy set and the stronger the definition of similarity (Zio & Di Maio, 2013). Values of μ_{ij} close to 0 indicate that the signal evolutions in the two transients i and j are very different, whereas values close to 1 indicate high similarity (Angstenberger, 2001). In what follows, the similarity measures have been obtained with an optimized value of the bell-shaped function parameter $\sigma=1.5$.

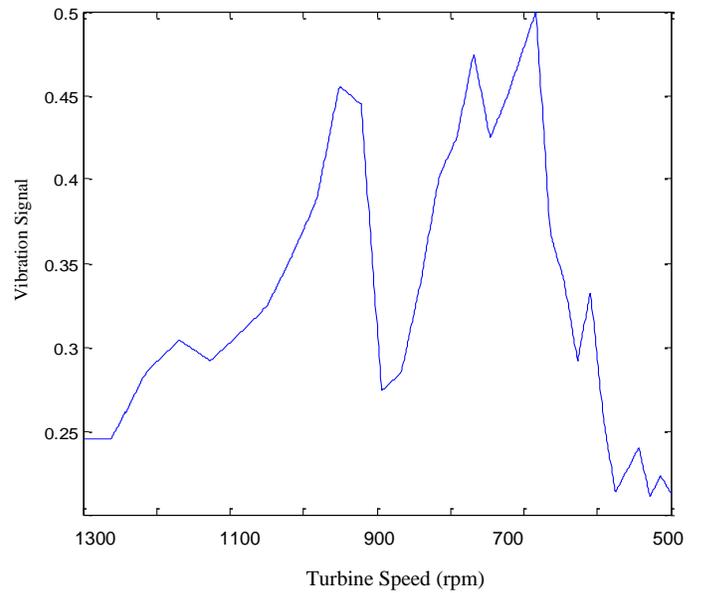


Figure 4. Representation of the evolution of Vibration signal “shaft vibration level 3” in transient 58

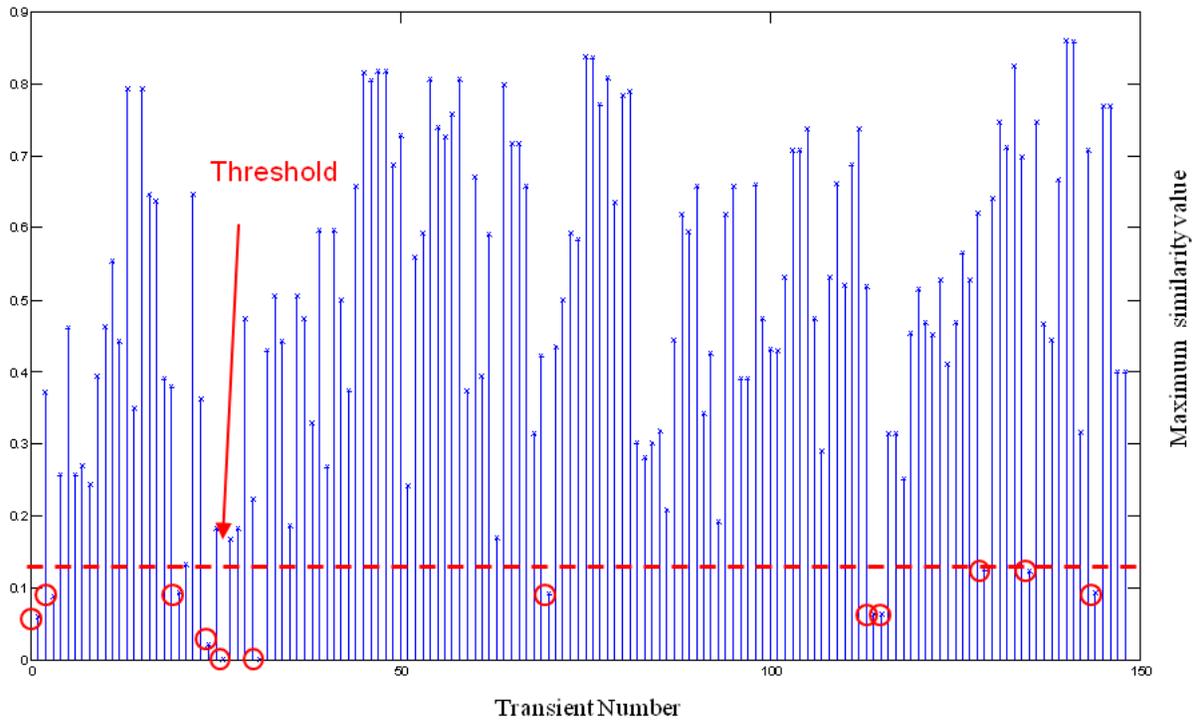


Figure 5. Maximum value of similarity for each transient

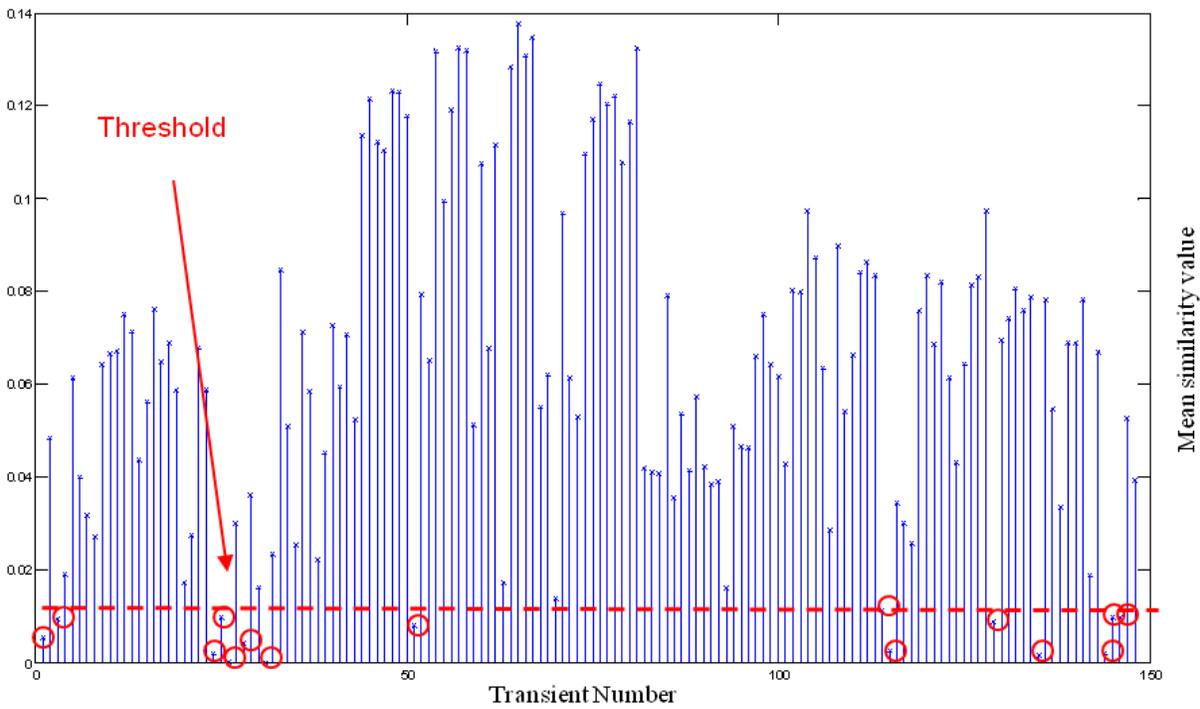


Figure 6. Mean value of similarity for each transient

For outliers identification, we have considered the distribution of μ_i^M , that is the maximum value among the similarities between the i -th transient and all the others, $\mu_i^M = \max(\mu_{ij}, j=1, \dots, 148)$, and the distribution of μ_i^m , that is the mean value of the similarities of the i -th transient with all the other, $\mu_i^m = \text{mean}(\mu_{ij}, j=1, \dots, 148)$. By exploiting a trial and error procedure, we have identified and set the 8th percentile of both distributions as threshold level to be used for outliers identification: if a transient has either μ_i^M or μ_i^m lower than the 8th percentile of the respective distribution, it is considered an outlier and eliminated from the dataset (circles in Figure 5 and Figure 6). The rationale behind this choice is that if a transient is either different from all the others or similar to only one transient but different from all the others, i.e. an outlier, in the first case its maximum value of similarity has to be low, whereas in the second case its maximum value of similarity could be large, but its mean value has to remain low.

It is worth pointing out that the indicator used for driving the trial and error procedure in setting the threshold is the discriminating capability of the successive clustering (presented in Section 3) between different transient behaviors; in particular, the presence of the outliers in the similarity matrix hid the differences between the eigenvectors of transients belonging to different clusters that have been fed to the spectral clustering approach, resulting in several clusters composed only by one or two transients, (i.e., outliers) and in one cluster containing all the other transients (von Luxburg, 2007). Although intuitive and reasonable from the engineering point of view, future work will focus on overcoming this latter empirical setting of the threshold levels for the outlier identification, by developing a more robust decision making technique.

The analysis of μ_i^M and μ_i^m has highlighted the presence of 17 outliers transients; 10 out of these 17 transients have been identified as outliers by both analyses; 2 out of 17 result as outliers only by the analysis of μ_i^M , which means that they are different from all the other transients but are characterized by higher values of similarity to the others transients than the 10 outliers previously identified (i.e., their μ_i^m is higher than the 8th percentile), while 5 out of 17 result as outliers only by the analysis of μ_i^m , which means that they are similar to a very low number of transients (i.e., 1 or 2), but completely different from all the other transients. The 17 identified outlier transients will not be considered in the rest of the analysis, which will regard only the remaining 131 transients.

3. GROUPING TRANSIENTS

Applying Eq. 2 and Eq. 3 to the 131 transients, we have built a similarity matrix \bar{W} of size [131, 131], where each element of the matrix is the similarity value μ_{ij} between the i -th transient, represented by the i -th row, and the j -th transient, represented by the j -th column of the considered element.

The diagonal components μ_{ii} are set to 1 and the matrix is symmetric $\mu_{ij} = \mu_{ji}$.

From the matrix \bar{W} a similarity graph $G = (V, E)$ is constructed, where each vertex v_i represents the i -th trajectory and the weight associated to the edge p_{ij} connecting the two vertices i and j is the similarity value μ_{ij} (von Luxburg, 2007). The original problem of identifying groups of similar trajectories can be reformulated in that of finding a partition of the similarity graph such that the edges connecting elements of different groups have small weights and the edges connecting elements within a group have large weights (Alpert, Kahng & Yao, 1999)

Figure 7 shows the 131 eigenvalues obtained by applying the spectral analysis method to matrix \bar{W} as described in Appendix A.

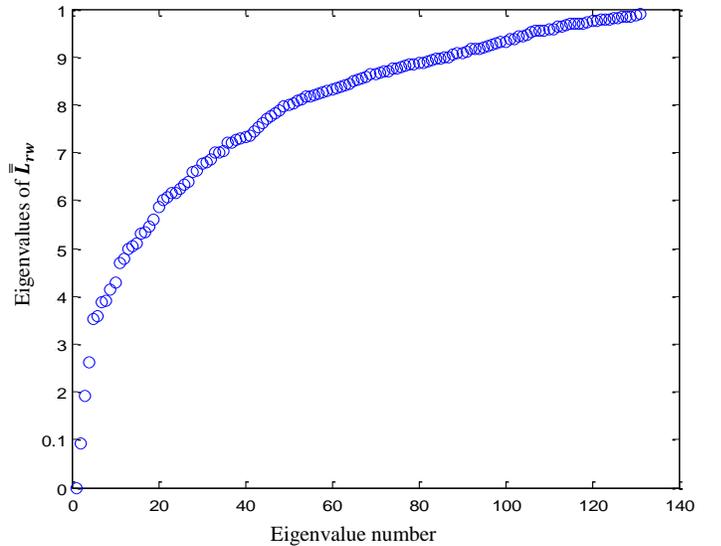


Figure 7. Representation of the 131 eigenvalues of \bar{L}_{rw}

The first four eigenvalues of the laplacian matrix \bar{L}_{rw} are close to zero, with the gap between the fourth and the fifth being the largest among all the successive gaps; for this reason, the number of clusters C is set equal to 4, according to the eigengap theory (von Luxburg, 2007).

The relevant information on the structure of the matrix \bar{W} is obtained by considering the eigenvectors $\bar{u}_1, \bar{u}_2, \bar{u}_3, \bar{u}_4$ associated with the 4 smallest eigenvalues of its laplacian matrix \bar{L}_{rw} . The square matrix \bar{W} is transformed into a reduced matrix \bar{U} of size [131, 4], in which the 4 columns of \bar{U} are the eigenvectors; it is important to underline that the first eigenvector is constant and consequently does not carry any information about the clustering of the data. The obtained eigenvectors $\bar{u}_1, \bar{u}_2, \bar{u}_3, \bar{u}_4$ are finally fed to the FCM spectral clustering algorithm (Alata, Molhim & Ramini, 2008; Karayiannis, 2000) in order to obtain groups of transients with similar characteristics.

4. RESULTS

The FCM partitioning of the obtained vectors \bar{u}_i provides the memberships m_{ic} of the i -th transient, $i=1,2,3,\dots,131$, to the c -th cluster, $c=1,2,3,4$; as mentioned in Section 3, 4th paragraph, the number of clusters $C=4$ has been set according to the eigengap theory (von Luxburg, 2007). The transient belonging to a cluster with the largest value of membership is named prototypical trajectory, and its functional behavior can be taken as most characteristic of the cluster.

The matrix of values m_{ic} containing the degree of membership of each transient, $i=1,\dots,131$ to each cluster $c=1,\dots,4$ has been firstly analyzed by considering for each transient the largest membership value. In practice, two categories of transients have been distinguished based on the degree of confidence of their classification:

- Transients assigned to a cluster with high confidence, hereafter labelled “assigned”. For these transients, the largest membership value is larger than 0.7 (106 transients).
- Transients not assigned to any cluster with enough confidence, hereafter labelled “not assigned”. These transients are characterized by membership values all lower than 0.7 (25 transients).

A threshold membership value equal to 0.7 has been chosen to distinguish between “assigned” and “not assigned” transients in order to obtain a satisfactory trade-off between the maximization of the fraction of transient assigned to a cluster that are correctly classified and the minimization of the fraction of transient that are not assigned (Zio & Di Maio, 2009).

4.1. “Assigned” transients

The most representative signals (i.e., signals 3, 4, 6) of the four identified clusters (i.e., rows 1 to 4) of the 106 “assigned” transients are represented in Figure 8, where each row refers to a different cluster.

For the ease of the comprehension, let us focus on the first column of Figure 8: cluster 2 is characterized by a delayed peak for signal 3, which has also a lower magnitude with respect to the peaks characteristics of the others identified clusters; cluster 1 is characterized by peaks of largest magnitudes, with respect to clusters 3 and 4.

Clusters 1, 3 and 4 can be distinguished by looking at the behavior of signal 4 (Figure 8, 2nd column): cluster 1 shows a significant peak around the turbine speed value 400, whereas the other two clusters are characterized by peaks of significantly lower magnitude (in particular cluster 4) and by a flatter behavior (cluster 3). Finally, clusters 3 and 4 can be distinguished looking at signal 6 (Figure 8, 3rd column); cluster 3 is characterized by a flat behavior, while transients belonging to cluster 4 show vibration peaks around the turbine speed value 600.

The obtained clusters clearly differ from each other and may be representative of different operational conditions. An in-depth analysis which has been performed on the identified clusters has shown that transients belonging to the same clusters are referred to the same chronological period. This consideration may lead to the conjecture that the functional behaviors identified by the clusters have some physical explanation that can be due to the degraded behavior of some components, as confirmed by *Electricité de France* (EDF) experts.

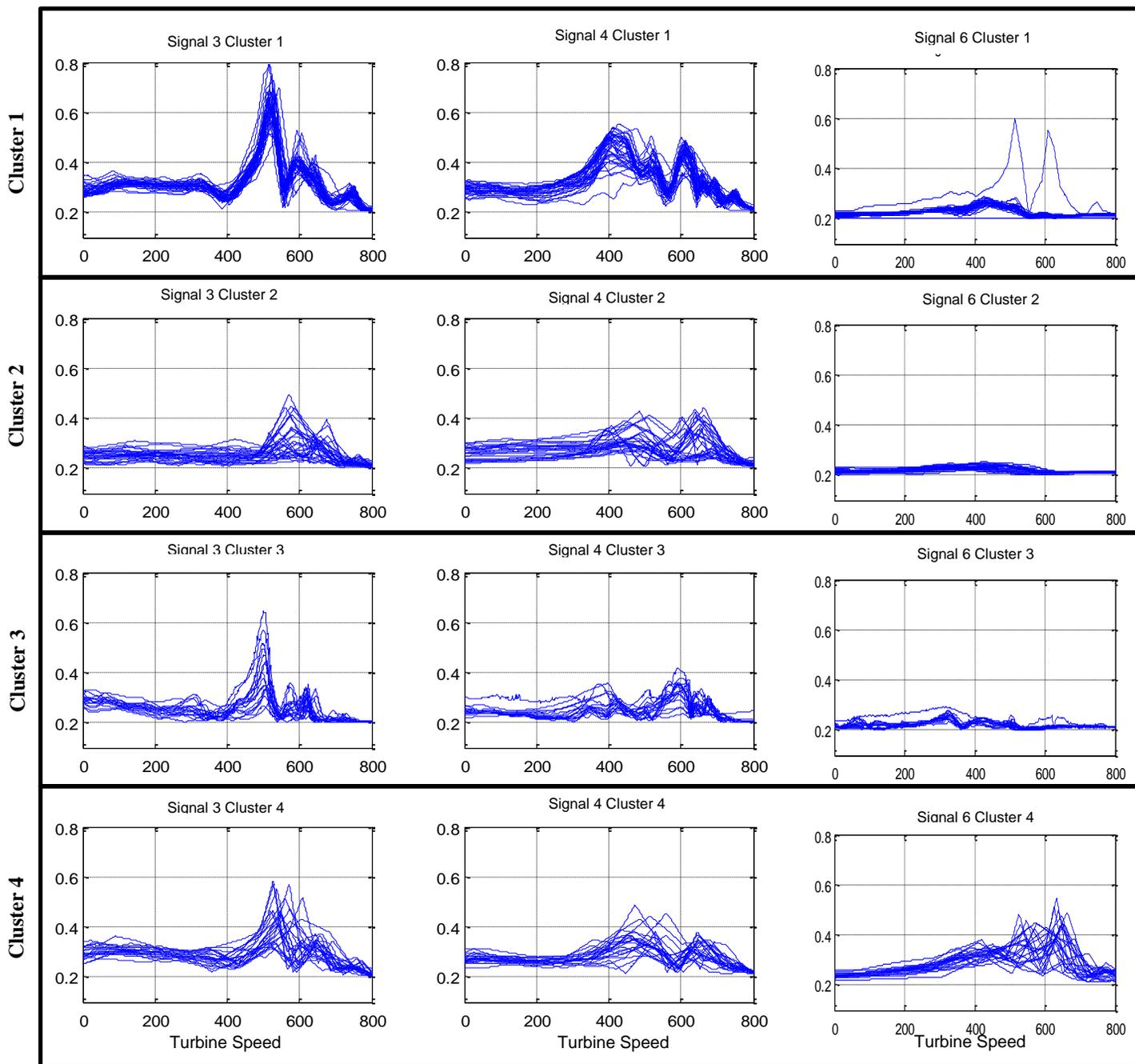


Figure 8. Representation of the behavior of the signals in the 4 identified clusters: each column represents a different signal, each row a different cluster

- (1st column) representation of the 4 clusters of signal 3 for “assigned” transients in “turbine speed” domain.
- (2nd column) representation of the 4 clusters of signal 4 for “assigned” transients in “turbine speed” domain.
- (3rd column) representation of the 4 clusters of signal 6 for “assigned” transients in “turbine speed” domain.

4.2. “Not Assigned” transients

By the analysis of the matrix of values m_{ic} containing the degree of membership of each transient, $i=1,\dots,131$ to each cluster $c=1,\dots,4$, we have identified 25 “not assigned” transients, which are defined as transients not assigned to any cluster with enough confidence, whose membership values are all lower than 0.7.

Figure 9 shows the behavior of signal 2 for the “not assigned” transients and the transient belonging to each cluster with the largest value of membership (black dashed lines), which will be referred to as prototypical trajectory since its functional behavior can be taken as most characteristic of the cluster

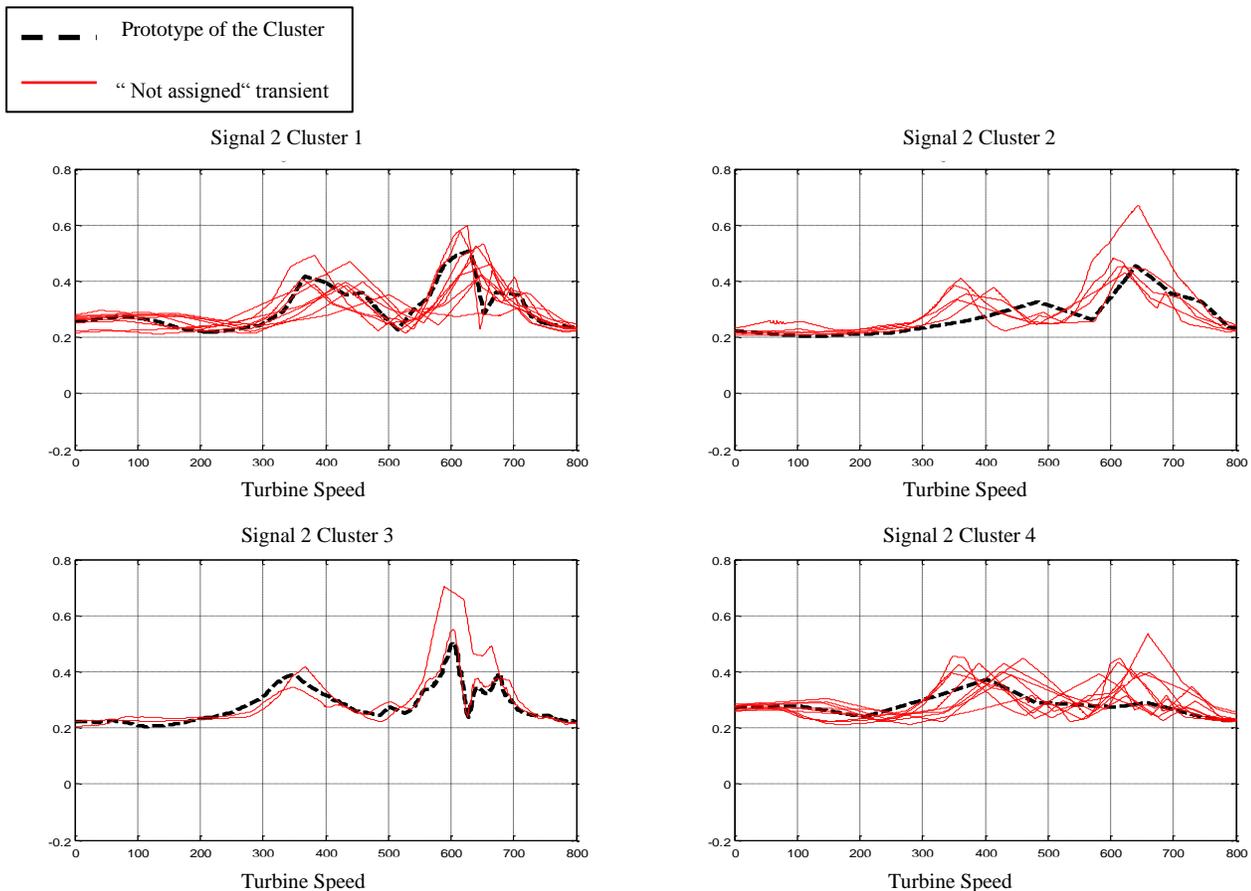


Figure 9. Representation of the “not assigned” transients for the 4 identified clusters. The prototypes of each cluster are shown by using a black dashed line.

“Not assigned” transients of cluster 1 are characterized by magnitudes of the peaks larger or a smaller than their prototypical trajectory (Figure 9, upper-left). “Not assigned” transients of cluster 2 (Figure 9, upper-right) are characterized by an earlier and larger peak in the signal around the turbine speed value 350, and a larger peak than the prototypical trajectory around 650. “Not assigned” transients of cluster 3 (Figure 9, bottom-left) have a larger magnitude of the second peak, while “not assigned” transients of cluster 4 (Figure 9, bottom-right) have a larger magnitude of the first peak, which seems also delayed or

early with respect to the prototype, and presents a second peak around the turbine speed 630 which is not characteristic of the prototypical trajectory of cluster 4 (black dashed line).

4.3. Outlier transients

We have analyzed the “assigned” and the “not assigned” transients and noticed that the identified clusters seem to represent different classes of operational conditions.

We now focus on the functional behavior of the 17 “outlier” transients identified in Section 2.2.

Indeed, if the identified clusters represent different classes of operational conditions, transients identified as outliers (that greatly differ from all the other transients) may be representative of anomalous conditions.

Figure 10 shows the behavior of signal 4 for the 17 identified outlier transients (see Section 2.2), plotted together with the prototypical trajectories of each identified cluster. It is easy to notice that the outliers are very different from any of the prototypical trajectories: in particular, they are characterized by larger magnitudes of the signals values.

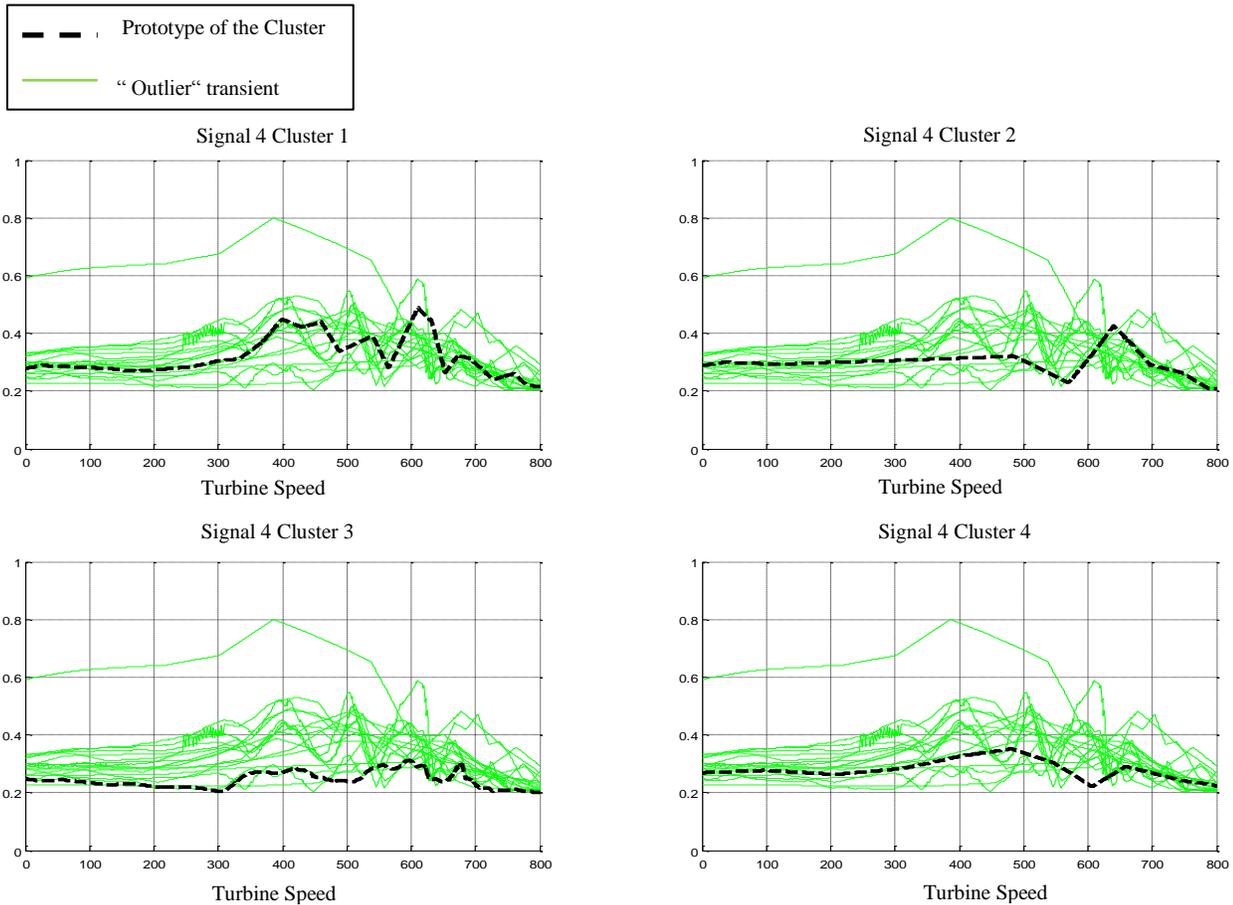


Figure 10. Representation of all the 17 identified “Outlier” transients (continuous line) with respect to the prototype of each cluster (black dashed line)

The only prototypical trajectory which shows a behavior which is not completely different from the outlier functional behavior is the prototype of cluster 1 (Figure 10, upper-left); however, even in this case the signals magnitude of the outlier transients is sensibly larger than the magnitude of the prototype. It is necessary to say that due to confidentiality reasons the causes of the identified outlier/anomalous behaviors cannot be described in details.

5. CONCLUSION AND FUTURE WORK

We have considered a real case study concerning 148 shutdown transients of a nuclear power plant (NPP) turbine. We have proposed a methodology for unsupervised classification of transients with similar behavior, for distinguishing among different signal functional behaviors. In order to overcome misalignment problems, the considered signals have been transformed from the “time domain” to the “turbine speed domain”, based on the fact that a particular magnitude of vibration is associated to a turbine speed value.

A fuzzy similarity measure capable of treating vibration signals has been constructed by computing the pointwise differences between signal transient values. Outlier transients (17 transients), with no similar functional behavior, have been discarded.

The clustering of the remaining 131 transients has then been obtained by computing the fuzzy similarity among the transients. A similarity graph is built, in which each vertex represents a transient and the weight associated to the edge connecting two vertices is the value of the fuzzy similarity between the two corresponding transients. An FCM-based spectral clustering algorithm is applied in order to find an optimal partition of the graph that leads to the identification of different groups of transients of different similarity.

Four different groups of transients representing different operational conditions have been identified: the difference between these clusters is mainly due to both different magnitudes of signal values (in particular this allows distinguishing between cluster 1, which has the largest magnitude of the vibration signals, and the others) and different functional behaviors (peaks at different “turbine speed”, which allows distinguishing between clusters 2, 3 and 4; for more details see Section 4.1). Furthermore, 25 transients have not been assigned to any cluster with enough confidence. Finally, the behavior of the 17 outlier transients has been analyzed under the conjecture that they may be representative of anomalous conditions in the turbine plant.

The proposed analysis procedure can be further developed. In particular it is necessary to identify a robust decision-making technique for setting the threshold for the identification of the outliers. Furthermore, feature extraction techniques could allow a better description of the functional characteristics of the vibration signals, such as Wavelet transforms (Baraldi, Di Maio, Pappaglione, Zio & Seraoui, 2011) and FFT (Fast Fourier Transform) (Mc Inerny & Dai, 2003), that could permit to obtain a similarity measure that improves the separation of the groups of similar transients. It is also possible to embed the spectral clustering algorithm within a possibilistic classifier (Zio, Podofillini, Mercurio & Dang, 2010) that could permit not eliminating the outliers before doing the clustering (Frigui et al., 1996).

Finally, the results obtained with the proposed procedure can be used for the development of informed empirical classification models for the identification of anomalous conditions in NPP nuclear turbine. In this context, the results obtained by the application of the unsupervised methodology here illustrated for the grouping of the transients could help in the early labelling of the transients, which were previously unknown.

REFERENCES

- Alata, M., Molhim, M., & Ramini, A., (2008). Optimizing of Fuzzy C-Means Clustering Algorithm Using GA. *World Academy of Science, Engineering and Technology*, p.224-229.
- Alpert, C., Kahng, A., & Yao, S., (1999). “Spectral partitioning: the more eigenvectors, the better”, *Discrete Applied Math*, Volume 90, pages 3-26.
- Angstenberger, L., (2001). Dynamic fuzzy pattern recognition. *International series in intelligent technologies*, vol. 17. Dordrecht, Kluwer Academic Publishers.
- Baraldi, P., Di Maio, F., Pappaglione, L., Zio, E., & Seraoui, R., (2011). A Wavelet-Based Approach for Condition Monitoring of a Gas Turbine During Start-Up Transients. *Proceedings of the European Safety and Reliability Conference*, Troyes, 24-29 September, 2011.
- Baraldi, P., Di Maio, F., Pappaglione, L., Zio, E., & Seraoui, R., (2012). Condition Monitoring of Electrical Power Plant Components During Operational Transients. *Proceedings of the Institution of Mechanical Engineers*, Part O, Journal of Risk and Reliability, 226(6) 568–583.
- Baraldi, P., Di Maio, F., & Zio, E., (2012). Unsupervised Clustering for Fault Diagnosis. *Proceedings of Prognostics and System Health Management Conference (PHM-2012)*, Beijing, China, 23-25 May 2012.
- Baraldi, P., Di Maio, F., & Zio, E., (2013). Unsupervised Clustering for Fault Diagnosis in Nuclear Power Plant Components. *International Journal of Computational Intelligence Systems*, Vol. 6, No. 4, pp. 764-777.
- Betta, G., Liguori, C., Paolillo, A., & Pietrosanto, A., (2002). A DSP-based FFT-Analyzer for the fault diagnosis of rotating machine based on vibration analysis. *IEEE Transactions on instrumentation and measurements*, Vol. 51, No. 6, pp. 1316-1322.
- Bezdek, J.C., (1981). “Pattern Recognition with Fuzzy Objective Function Algorithms”, Plenum, New York.
- Bocaniala, C.D., Sa Da Costa, J., Palade, V., (2004). A novel fuzzy classification solution for fault diagnosis. *Journal of Intelligent and Fuzzy Systems*, 15 (3-4), pp. 195-205.
- Frigui, H., & Krishnapuram, R., (1996). Robust clustering algorithm based on competitive agglomeration and soft rejection of outliers. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 550-555.
- Karayiannis, N.B., (2000). Generalized fuzzy c-means algorithms. *Journal of Intelligent and Fuzzy Systems*, 8 (1), pp. 63-81.
- Klir, G., & Yuan B., (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Application*. Prentice Hall, Englewood Cliffs, NJ.

Leguizamon, S., & Azzali, S., (1996). "Unsupervised Fuzzy C-Means Classification for the Determination of Dynamically Homogeneous Areas", p.851-856.

McInerney, S. A., & Dai, Y., (2003). Basic Vibration Signal Processing for Bearing Fault Detection. *IEEE Transactions on Education* 46 (1) , pp. 149-156.

Mohar, B., (1997). "Some Applications of Laplace Eigenvalues of Graphs", in G. Hahm and G. Sabidussi (Eds.), *Graph Symmetry: Algebraic Methods and Applications* (Vol. NATO ASI Ser. C 497), pages 225-275.

Roemer, M. J., & Kacprzyński, G. J., (2000). Advanced diagnostics and prognostics for gas turbine engine risk assessment. *IEEE Aerospace Conference Proceedings* 6, pp. 345-354.

Sandizadeh, M.A., Dehghani, M., (2013). Intelligent condition monitoring of railway signaling in train detection subsystems. *Journal of Intelligent and Fuzzy Systems*, 24 (4), pp. 859-869.

Secchi, P., Sangalli, L. M., Vantini, S., & Vitelli, V., (2008). K-mean alignment for curve clustering. *Communications in Applied and Industrial Mathematics*, ISSN 2038-0909, 1, 1, 204-224, Milano, Italy.

Sheng, Y., & Rovnyak, S.M., (2004). Decision Tree-Based Methodology for High Impedance Fault Detection, *IEEE Transactions on Power Delivery*, Vol. 19, No. 2.

Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S.N., (2003). A Review of Process Fault Detection and Diagnosis: Part I: Quantitative Model-based Methods, *Computers & Chemical Engineering*, Vol. 27, pp. 293-311.

von Luxburg, U., (2007). A Tutorial on Spectral Clustering. *Statistics and Computing*, Volume 17(4), pages 395-416.

Zhao, Z., & Liu, H., (2007). "Spectral feature selection for supervised and unsupervised learning", *Proceedings of the 24th international conference on Machine learning*, p.1151-1157, June 20-24, Corvallis, Oregon.

Zio, E., Soft Computing Methods Applied to Condition Monitoring and Fault Diagnosis for Maintenance, *Reliability: Theory and Applications*, Vol. 3, 2007.

Zio, E., & Di Maio, F., (2009). Processing dynamic scenarios from a reliability analysis of a nuclear power plant digital instrumentation and control system. *Annals of Nuclear Energy* 36, pp. 1386–1399.

Zio, E., & Di Maio, F., (2010). A Data-Driven Fuzzy Approach for Predicting the Remaining Useful Life in Dynamic Failure Scenarios of a Nuclear System. *Reliability Engineering and System Safety*, RESS, Volume 95(1), Pages 49-57.

Zio, E., Podofillini, L., Mercurio, D., & Dang, V. N., (2010). Dynamic safety assessment: Scenario identification via a possibilistic clustering approach. *Reliability Engineering and System Safety* 95 (5) , pp. 534-549.

Zio, E., & Di Maio, F., (2013). Failure Prognostics by a Data-Driven Similarity Based Approach. *International Journal of Reliability Quality and Safety Engineering*, Vol 20 (2), No.1, pp.1-17.

Appendix A: Unsupervised spectral Clustering

- Step 1: normalized Graph Laplacian Matrix

Compute:

- the degree matrix $\overline{\overline{D}}$ which is a diagonal matrix with diagonal entries d_1, d_2, \dots, d_N defined by

$$d_i = \sum_{j=1}^N \mu_{ij}, \quad i = 1, 2, \dots, N \quad (1a)$$

where $j=1, \dots, N$ indicates the j -th elements of the i -th row of the similarity matrix whose elements are represented by μ_{ij} .

- the normalized graph Laplacian matrix:

$$\overline{\overline{L}}_{rw} = \overline{\overline{D}}^{-1} \overline{\overline{L}} = \overline{\overline{I}} - \overline{\overline{D}}^{-1} \overline{\overline{W}} \quad (2a)$$

where $\overline{\overline{L}} = \overline{\overline{D}} - \overline{\overline{W}}$ and $\overline{\overline{I}}$ is the identity matrix of size $[N, N]$.

- Step 2: eigenvalues and eigenvectors of $\overline{\overline{L}}_{rw}$

The information on the structure of a graph can be obtained from its spectrum (Zhao & Liu, 2007). Given $\overline{\overline{L}}_{rw}$, compute the first C eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_C$ and corresponding eigenvectors $\overline{\overline{u}}_1, \overline{\overline{u}}_2, \dots, \overline{\overline{u}}_C$. The first C eigenvalues are such that they are very small whereas λ_{C+1} is relatively large [von Luxburg, 2007].

- Step 3: Number of clusters

The number of clusters is set equal to C , according to the eigengap heuristic theory (Mohar, 1997).

- Step 4: Feature extraction

The relevant information on the structure of the matrix $\overline{\overline{W}}$ is obtained by considering the eigenvectors $\overline{\overline{u}}_1, \overline{\overline{u}}_2, \dots, \overline{\overline{u}}_C$ associated to the C smallest eigenvalues of its laplacian matrix $\overline{\overline{L}}_{rw}$. The square matrix $\overline{\overline{W}}$ is transformed into a reduced matrix $\overline{\overline{U}}$ of size $[N, C]$, in which the C columns of $\overline{\overline{U}}$ are the eigenvectors $\overline{\overline{u}}_1, \overline{\overline{u}}_2, \dots, \overline{\overline{u}}_C$. Thus, the i -th transient similarity with

other transients is captured in the C -dimensional vector \bar{u}_i corresponding to the i -th row of the matrix \bar{U} . It has been shown that this change of representation enhances the cluster properties in the data, so that clusters can be more easily identified [von Luxburg, 2007].

- ***Step 5: Unsupervised clustering***

In this work, we resort to the Fuzzy C-Means (FCM) algorithm to partition the data into C clusters (Bezdek, 1981; Leguizamon & Azzali, 1996; Alata, Molhim & Ramini, 2008). FCM originates from hard C-Means clustering: the difference is that it allows elements (transients, in our case) to belong to two or more clusters (Klir & Yuan, 1995)

. For each i -th element, the algorithm provides its membership m_{ic} to all clusters, $c = 1, 2, \dots, C$. If needed, crisp assignment can be obtained, e.g., by considering the cluster to whom the element belongs with the largest membership value. A prototypical transient can be identified for each cluster by considering the transient with the largest membership value to the cluster.