Clustering acoustic emission signals by mixing two stages dimension reduction and nonparametric approaches

O I Traore, Paul Cristini, Nathalie Favretto-Cristini, L Pantera, Philippe Vieu, Sylvie Viguier-Pla

Noname manuscript No.
(will be inserted by the editor)

Clustering acoustic emission signals by mixing two stages dimension reduction and nonparametric approaches

O. I. Traore[1], P. Cristini[1], N. Favretto-Cristini[1], L. Pantera[2], P. Vieu[3], S. Viguier-Pla[4&3]

Received: date / Accepted: date

Abstract In the context of nuclear safety experiments, we consider curves issued from acoustic emission. The aim of their analysis is the forecast of the physical phenomena associated with the behavior of the nuclear fuel. In order to cope with the complexity of the signals and the diversity of the potential source mechanisms, we experiment innovative clustering strategies which creates new curves, the envelope and the spectrum, from each raw hits, and combine spline smoothing methods with nonparametric functional and dimension reduction methods. The application of these strategies prove that in nuclear context, adapted functional methods are effective for data clustering.

Keywords Functional clustering · curve smoothing · hierarchical clustering · semi-metric · Functional Principal Component Analysis

1 Introduction and context of the study

Reactivity Initiated Accident (RIA) is a nuclear reactor accident which involves an unexpected and very fast increase in fission rate and reactor power due to the ejection of a control rod. The power increase may damage the fuel and reactor core, and in severe cases, create pressure pulses in the reactor coolant [Rudling et al., 2016]. Historically, the worst RIA accident took place on April 1986 in reactor 4 of the Chernobyl power plant in Ukraine [Jernkvist and Massih, 2010]. This triggered experimental programs in the early of the 1990s in France, Japan and Russia. In the specific case of France, since 1993 fourteen experiments have been operated in the French Alternative Energies and Atomic Energy Commission (CEA) research center of Cadarache in southern France, with the objective to study the behavior of the nuclear fuel in RIA conditions.

In order to cope with the difficulty of access and the hostility of the nuclear environment, the experimentalists of the CEA rely among other on the Acoustic Emission (AE) technique through the recordings of AE sensors installed at the top and the bottom of the test device hosting the fuel to be tested. The experimental AE signals obtained through these sensors are composed of numerous transients (called the hits) resulting from wave propagation generated by unknown source mechanisms of interest, and corrupted by several sources among which the choice and the settings of the AE acquisition system [Roget, 1988, Gautschi, 2002], the environmental noise and the wave propagation

[1]Aix-Marseille Univ., CNRS, Centrale Marseille, L.M.A., France, E-mail: toumarissiaka@gmail.com · [2]CEA, DEN, DER/SRES, Cadarache,F13108 Saint-Paul-Lez-Durance, France, E-mail: laurent.pantera@cea.fr · [3]Equipe de Stat. et Proba., Institut de Mathématiques, UMR5219, Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France, E-mail: vieu@math.univ-toulouse.fr · [4] Université de Perpignan via Domitia, LAMPS, 52 av. Paul Alduy, 66860 Perpignan Cedex 9, France, E-mail: viguier@univ-perp.fr

path in the test device hosting the nuclear fuel. After each experiment, the raw experimental signals have to be preprocessed in order to cope with their corruption and then to detect the hits.

Once the preprocessing is realized, the remaining of the work-flow consists in identifying the specific mechanism (clad failure, gas ejection...) corresponding to each hit. In general, this is done by a characterization of each hit by computing several variables called AE parameters like the amplitude, the rise time, the duration... (Figure 1).

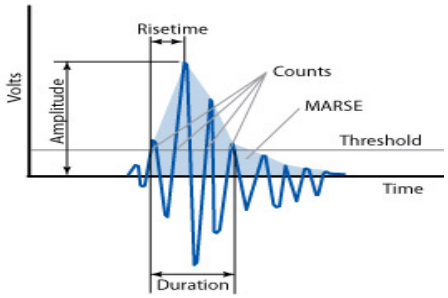


Fig. 1: Illustration of the computation of some typical AE parameters (`https://www.nde-ed.org/index_flash.htm`)

.

When the number of experiments to process and then of hits to characterize gets large, clustering methods based on these AE parameters are used to classify them. Depending on the diversity of the source mechanisms associated with the hits (cracks, fractures, delaminations ...) and the type of material (nuclear fuel, zircaloy, inox...) very different types of AE parameters can be discriminant [Favretto-Cristini et al., 2016, Anastassopoulos and Philippidis, 1995, Ai et al., 2010, Keyvan and Nagaraj, 1996]. This is the case for the nuclear environment of interest here. A very attractive alternative to the classical approach is then to avoid the steps of AE parameters computation and selection by considering each hit as a functional variable and use the tools dedicated to Functional Data Analysis (FDA) [Ferraty and Vieu, 2006, Ramsay and Silverman, 2005] to classify them (Figure 2). Thus, we move from the finite dimensional problem in which the random vector $\mathbf{x}$ taking values in $\mathbb{R}^p$ such that $\mathbf{x} = \{\mathbf{x}^1, \ldots, \mathbf{x}^p\}$, $p \geq 1$ to the infinite dimensional problem associating each hit with a functional random variable $\boldsymbol{\mathcal{X}}$ taking values in an infinite dimension space E such that $\boldsymbol{\mathcal{X}} = \{\boldsymbol{\mathcal{X}}(t);\ t \in \mathbf{T}\}$, where $\mathbf{T}$ is a random time interval.

Since the end of the 1990s, facing the increasing number of situations as ours in which data are continuous (functions, curves, images, surfaces...), the interest of the scientific community has consistently grown for FDA [Goia and Vieu, 2016, Cuevas, 2014]. In particular, the field has been popularized by the means of the book's of [Ramsay and Silverman, 2002, Ramsay and Silverman, 2005] and [Ferraty and Vieu, 2006]. From linear regression to classification, the contributions in FDA cover most of the problems conventionally encountered in multivariate data analysis [Cardot et al., 1999, Chiou et al., 2004, Aguilera et al., 2006, Rossi and Villa, 2006, Preda et al., 2007, Cuesta-Albertos and Febrero-Bande, Ferraty and Vieu, 2003]. The recent developments concern as well mathematical backgrounds [Huckemann and Eltzner, Boudou and Viguier-Pla, 2017], methodological aspects [Jiang et al., 2017, Febrero-Bande et al., 2017], as applications [Walders and Liebl, 2017, Lila et al., 2017].

If the concept is very attractive, FDA is known to be a difficult task because of the infinite dimensional space data belong to. The definition of a probability density of a functional random variable and

the definition of a metric (or a semi-metric) to compare the variables are two examples of such difficulties [Jacques and Preda, 2014]. Along the years, various methods have been proposed in the litterature to manage functional data. In the particular case of data clustering which is of interest in this work, the two most popular of them are the dimension reduction method and the functional nonparametric method. The first method consists in reducing the infinite dimensional problem intrinsic to functional data into a finite one by approximating data with elements from some finite dimensional space [Ramsay and Silverman, 2005, Abraham et al., 2003, Ullah and Finch, 2013, Rossi et al., 2004], allowing thus to use all the known classical clustering methods (k-means, unsupervised neural networks, hierarchical clustering, Self-Organizing Map, ...) [Friedman et al., 2001]. In the case of the functional nonparametric method, the intuitive idea is about defining specific metrics (or semi-metrics) to compare directly the curves [Ferraty and Vieu, 2006, Ieva et al., 2013], the two main clustering methods being then the k-means and the hierarchical clustering [Jacques and Preda, 2014].

Whatever the objective of the use of FDA, in general, a preliminary treatment of data before applying FDA techniques is required. This is motivated by potential registration problems as reconstruction of functional form of the curves from discrete observations, smoothing, handling of horizontal shifts and handling of differences in curves supports. In order to cope with these data processing problems, several tools are proposed in [Ramsay and Silverman, 2005]. Furthermore, in the specific case of the hits presented in Figure 2, in addition to all these cited problems and despite the precautions of the hit detection strategy, it is difficult to ensure an optimal detection of all the hits with a global setting. This raises a question of reliability of the exploited hit signals to correctly characterize the associated physical source mechanisms. For example, knowing that the start and the end of a hit are respectively triggered by an exponential rise or fall of its energy, it is clear that the hits 40, 149 and 52 show a start detection bias, while the hits 109 and 126 show a end detection one. Moreover, one can observe that the hit 84 shows a local maximum.

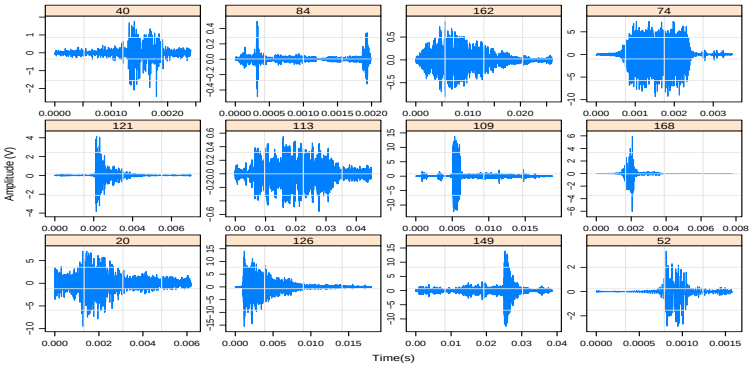


Fig. 2: Sample of 12 hits (sensor M1) chosen among 168 detected during the experiments operated since 1993.

The first objective of this work is to propose an effective registration strategy for this type of complex curves, which could automatically manage the hit detection biases in addition to the classical problems of horizontal shifts and difference of supports. Secondly, we want to cluster them as efficiently as possible, making out with their roughness and the associated difficulty to build effective semi-metrics for their comparison.

In Section 2, we propose a complete approach leading from the raw experimental signals to the processed envelopes and spectra. Section 3 is devoted to the presentation of three different strategies for the clustering of the AE hits relying on the two curves (the spectrum and the envelope) created from each hit. In Section 4, we resume the proposed preprocessing and clustering strategies, then we present the criteria for clusters validation. Section 5 is devoted to compare and discuss the obtained results and we conclude in the Section 6.

Note that all the computations and results showed in this work have been realized with the open-source R software [R Development Core Team, 2008].

2 Preprocessing data

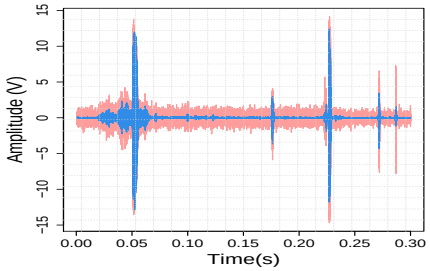
2.1 Registration of the Raw experimental curves

From a mathematical point of view, if the environmental noise n is assumed to be additive, the received experimental signal y can be expressed as:

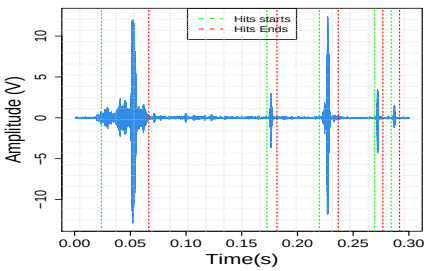
$$y(t) = h_1 * h_2 * x(t) + n(t), t \in \{1, \ldots, T\} \tag{1}$$

where h_1 and h_2 correspond to the impact of the structure and the acquisition system, respectively, and x is the signal associated with the physical source mechanism of interest and t the time.

A detailed presentation of the part of the preprocessing consisting in the estimation of x and the detection of the hits is beyond the scope of this article. For more details, the reader can refer to our works [Pantera and Traore, 2015, Traore et al., 2017d, Traore et al., 2017c, Traore et al., 2017b, Traore et al., 2017a]. We note that here, a default treatment of y based on a spectral subtraction denoising and a hit detection using a moving variance algorithm (Figure 3) have been chosen. The application of this treatment to the fourteen experiments of interest leads to a sample of 168 hits like those presented in Figure 2.



(a) Raw experimental signal (red) and associated denoised signal (blue) obtained after the use of the spectral subtraction method [Traore et al., 2017c]



(b) Illustration of the hits detection procedure by using a moving variance algorithm [Pantera and Traore, 2015]. The hits starts and ends are indicated in green and red, respectively.

Fig. 3: Illustration of the processing of a raw experimental signal. Case of an experiment operated in 1998 (sensor M1). Note that the codes developed for the denoising of the raw experimental signals and the hits detection use mainly the packages *signal, seewave and qcc* [signal developers, 2014, Sueur et al., 2008, Scrucca, 2004].

In the following we assume that we have a sample 168 hits potentially associated with physical sources mechanisms of interest and for which we want to identify the best clustering strategy based on FDA tools.

2.2 Correction of the hits detection biases

As mentioned in the introduction, the potential hit detection biases which could occur after the processing of the raw experimental signals are of three types: the hit start detection bias, the hit end detection bias and the existence of local maxima. In order to handle them, we consider the envelope of each hit.

The analytical signal $\mathcal{Z}_i$ associated with a given hit χ_i is expressed as:

$$\mathcal{Z}_i(t) = \chi_i(t) + j\mathcal{Y}_i(t) = \mathcal{E}_i(t)\mathrm{e}^{j\phi(t)} \quad ,$$

where $j^2 = -1$ and $\mathcal{Y}_i(t)$ is the Hilbert transform of $\chi_i(t)$:

$$\mathcal{Y}_i(t) = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{\chi_i(t)}{t-\tau}\mathrm{d}\tau,$$

$\phi(t)$ and $\mathcal{E}_i(t)$ are the instantaneous phase and the envelope of $\chi_i(t)$:

$$\mathcal{E}_i(t) = \sqrt{\chi_i^2(t) + \mathcal{Y}_i^2(t)} \tag{2}$$

Let us denote by I_i the interval defined as:

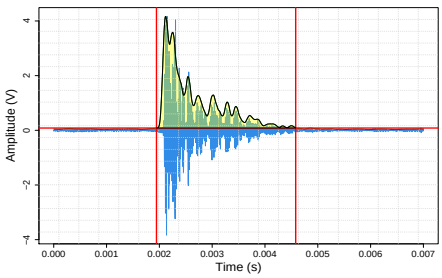
$$I_i = \{t \in \mathcal{T}_i, \mathcal{E}_i(t) > \alpha_1 \max(\mathcal{E}_i(t))\}, \quad \alpha_1 \in (0,1) \quad . \tag{3}$$

If $I_i = \mathcal{T}_i$, we consider that there is no detection bias. Otherwise, if $I_i \neq \mathcal{T}_i$, we suppose that there is a hit start and/or a hit end detection bias and the support $\mathcal{T}_i$ of χ_i is replaced by I_i, such that $\chi_i^* = \{\chi_i(t), t \in I_i\}$. Then, if the support I_i of χ_i^* is continuous, we stop the processing and χ_i is replaced by χ_i^* (Figure 4a). Otherwise, if the support I_i of χ_i^* is discontinuous, we conclude that there is a local maxima to process (Figure 4b). In this last case, I_i can be written as the union of at least two disjoint and not related intervals:

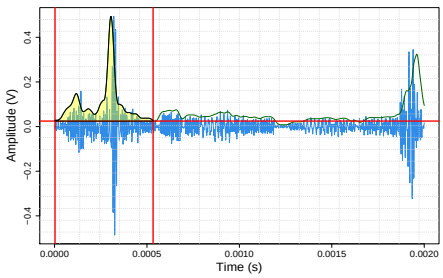
$$I_i = \bigcup_{j=1}^{J} I_{ij} \quad , J \geq 2 \;\; \textit{such that} \;\; \forall j \neq j' \in \{1,\ldots,J\} \; \textit{we have} \; I_{ij} \cap I_{ij'} = \emptyset. \tag{4}$$

Then, we keep only one interval I_i^* with a continuous support such that:

$$I_i^* = \{I_{ij}, \max(\mathcal{E}_i(t)) \in I_{ij}\} \;\; \textit{with} \; j \in \{1,\ldots,J\}. \tag{5}$$



(a) Correction of the detection biases for hit 121. The support I_i of χ_i^* is indicated by the red vertical lines. $\alpha_1 = 0.002$.



(b) Correction of the detection biases for hit 84. The support I_i^* of χ_i^* is indicated by the red vertical lines. $\alpha_1 = 0.12$.

Fig. 4: Illustration of the proposed method for handling of the hit detection biases.

Note that the main difficulty related with this method is the choice of the parameter α_1. Indeed, as highlighted in the Figure 4, the optimal value of α_1 depends on the characteristics of the hit. Then, an optimization of the value α_1 according to the characteristics of each hit is desirable. This remains an open problem for us and is beyond the scope of this work. However, for the results presented in the following, a default value of $\alpha_1 = 0.1$ has led to good classification results, suggesting a good robustness to a global choice of the value of α_1.

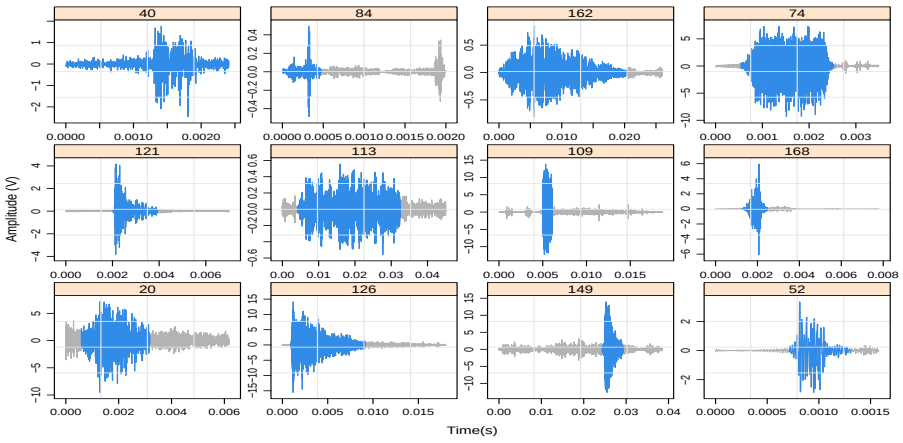


Fig. 5: Results of the application of the pre-treatment process for the hits presented in Figure 2.

2.3 Decomposition of the raw curves into envelopes and spectra

2.3.1 Computation of the envelopes and the spectra

In order to make easiest the comparison of the hits, for example by using $\mathcal{L}_2$ semi-metrics (which we will use by default in the following), it is advisable to ensure them a kind of smoothness. However, in the specific case of interest here, applying a smoothing is equivalent to use a low-pass filter. This could lead to the loss of the information associated with certain high frequency components. Then,

we make the assumption that information contained in the raw hit is summarized by those contained in its envelope and spectra:

$$\mathcal{I}(\chi_i) = \mathcal{I}(\mathcal{E}_i) + \mathcal{I}(\mathcal{S}_i) \quad . \tag{6}$$

This choice is justified by the fact that smoothing is a very classical processing for the envelope or the spectrum of a signal. Furthermore, from a physical point of view, these two curves allow to compute most of the AE parameters associated with the waveform or the frequency content of a hit.

2.3.2 Processing the envelopes and the spectra

The decomposition of the hits into two curves (envelope and spectra) paves the way for the use of smoothing methods and then for curves comparison. In general, the same tools can be used for reconstruction of the functional form of the curves and for smoothing. For a given hit χ_i, let us consider the discrete observations $(t_{ij}, \mathcal{E}_{ij})$ associated with the envelope (respectively, $(f_{ij}, \mathcal{S}_{ij})$ associated with the spectra), we have:

$$\mathcal{E}_{ij} = \mathcal{E}_i(t_j) = \mathcal{E}_i^*(t_j) + \epsilon_i(t_j), \quad with \;\; t_j \in \{0, \ldots, t_{n_i}\} \quad , \tag{7}$$
$$\mathcal{S}_{ij} = \mathcal{S}_i(f_j) = \mathcal{S}_i^*(f_j) + \xi_i(f_j), \quad with \;\; f_j \in \{0, \ldots, f_i/2\} \quad , \tag{8}$$

we want to estimate $\mathcal{E}_i^*(t)$ and $\mathcal{S}_i^*(f)$ such that:

$$\mathcal{E}_i(t) = \mathcal{E}_i^*(t) + \epsilon_i(t), \quad with \;\; t \in \mathcal{T}_i \quad , \tag{9}$$
$$\mathcal{S}_i(f) = \mathcal{S}_i^*(f) + \xi_i(f), \quad with \;\; f \in [0, f_i/2] \quad . \tag{10}$$

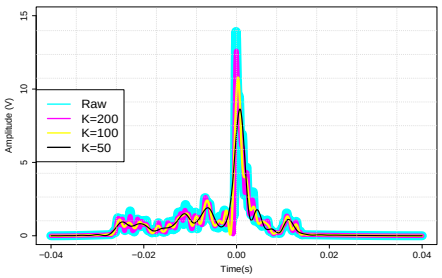
To do this, the classical approach is to use basis functions ϕ_{ik} and ψ_{ik} such that:

$$\mathcal{E}_i^*(t) = \sum_{k=1}^{K_i} c_{ik}\phi_{ik}(t), \quad with \;\; t \in \mathcal{T}_i, \tag{11}$$

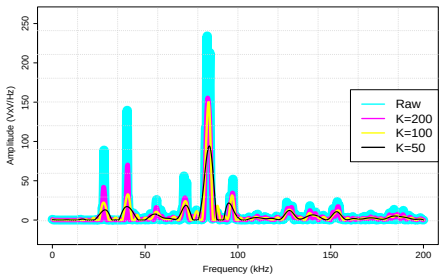
$$\mathcal{S}_i^*(f) = \sum_{k=1}^{L_i} d_{ik}\psi_{ik}(f), \quad with \;\; f \in [0, f_i/2], \tag{12}$$

where K_i and L_i are the number of basis functions ϕ_{ik} and ψ_{ik}, respectively $c_i = (c_{i1}, \ldots, c_{iK_i})^T$ and $d_i = (d_{i1}, \ldots, d_{iL_i})^T$ are the vectors of coefficients associated with the different basis functions generally estimated by the minimization of the mean square error.

According to the nature of the curves, several types of basis functions could be convenient (exponential, wavelets, splines...). In our case, as the envelopes and the spectra are non periodic curves, we consider spline bases. Figure 6 allows to observe that a relatively small number of basis functions is enough to get good smoothing. However, one has to manage the reduction of energy of the envelopes or spectra, by multiplying for example the smoothed curves by the ratio between the maximum amplitude of the raw curves and the maximum amplitude of the smoothed curve.



(a) Smoothing an evelope (hit 149)



(b) Smoothing an spectra (hit 126)

Fig. 6: Illustration of the smoothing of the envelopes and the spectra by cubic splines according to the number of basis functions.

Once the smoothing and the reconstruction functional form made, the remaining part of the process is more or less complex depending on the curve type. In the case of the spectra $\mathcal{S}_i$, there is no horizontal shift to take into account, the only remaining problem is then to cope with the differences of supports. The natural solution to this problem is to upsample all the hits in order to uniformize the sampling frequency to f, such that all the spectra have the same support $[0, f/2]$ with:

$$f = \max(f_i). \tag{13}$$

In the case of the envelopes, in addition to the handling of the differences of supports, it is necessary to deal with the horizontal shifts. For this, we suggest to adapt each envelope $\mathcal{E}_i$ with the following transformations.

Consider $\tilde{\mathcal{E}}_i(t)$, $t \in [-(2t_{n_i} - \delta_i), 2t_{n_i} - \delta_i]$, such that:

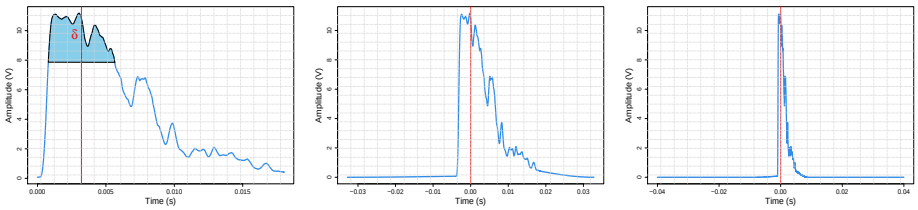
$$\tilde{\mathcal{E}}_i(t) = \begin{cases} \mathcal{E}_i(t + \delta_i) & \text{if } t \in [0, t_{n_i}] \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where $\delta_i = median\{t, \mathcal{E}_i(t) \geq \alpha_2 \ \max(\mathcal{E}_i(t))\}$, with $\alpha_2 \in (0, 1)$. Thus, $\forall i$, we have $\tilde{\mathcal{E}}_i(0) = \mathcal{E}_i(\delta_i)$. This is equivalent to the correction of the horizontal shift such that all the envelopes have their peak at the same landmark. Note that as in the case of the parameter α_1 (see section 2.2), an optimization of the value α_2 according to the characteristics of each hit is beyond the scope of this work.

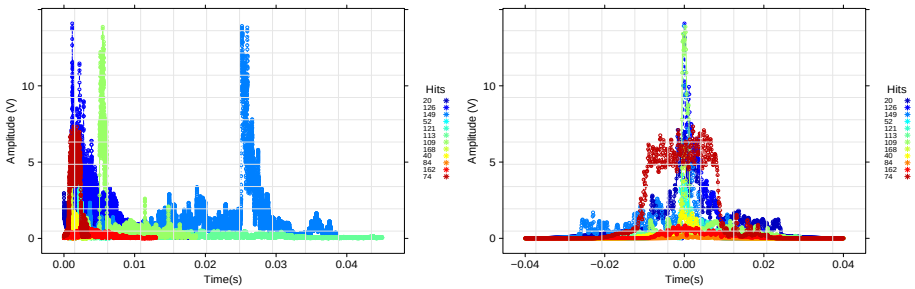
Consider $D \in \mathbb{R}_+$, $D_i = \dfrac{2t_{ni} - \delta_i}{D}$, and:

$$\tilde{\tilde{\mathcal{E}}}_i(t) = \tilde{\mathcal{E}}_i\left(\frac{t}{2D_i}\right) \quad with \quad t \in [-(2t_{n_i} - \delta_i), 2t_{n_i} - \delta_i]. \tag{15}$$

This last operation ensures the same support $[-D/2, D/2]$ to all the envelopes.

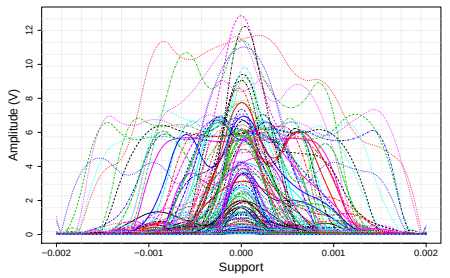


(a) Correction of the horizontal shift, handling of the support difference and contraction. Case of the envelope 126

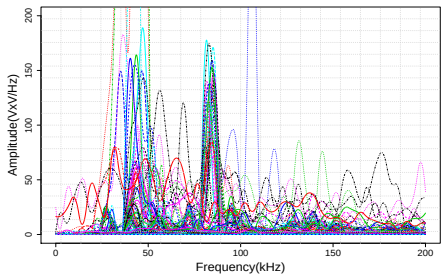


(b) Correction of the horizontal shift, handling of the supports differences and contraction. Case of the 12 hits presented in Figure 2

Fig. 7: Illustration of the handling of the differences of supports and of horizontal shifts for the envelopes



(a) Complete sample of processed envelopes. Spline of order 3 and K=15



(b) Complete sample of processed spectra. Spline of order 3 and K=50

Fig. 8: Sample of envelopes and spectra obtained after the application of the proposed preprocessing strategies.

3 Clustering

3.1 Clustering strategies using the envelope and the spectra

In the introduction, we have noticed that the two most used approaches for functional data clustering are the dimension reduction and the nonparametric functional approach. As each hit is now repre-

sented by two processed curves (the envelope and the spectra), several clustering strategies combining these two curves and the two functional data clustering approaches can be experimented. Here we consider three of them.

3.1.1 Strategy 1: Exclusive dimension reduction

The first strategy consists in creating a matrix of data which gathers the variables issued from a reduction of dimension based, on one hand, on the spectra, and on another hand, on the envelopes. Then we would have:

$$X = \left(\begin{array}{ccc|ccc} x_1^1 & \dots & x_1^{j^*} & x_1^{j^*+1} & \dots & x_1^p \\ x_2^1 & \dots & x_2^{j^*} & x_2^{j^*+1} & \dots & x_2^p \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ x_N^1 & \dots & x_N^{j^*} & x_N^{j^*+1} & \dots & x_N^p \end{array} \right) \tag{16}$$

where $\{x^1, \dots, x^{j^*}\}$ and $\{x^{j^*+1}, \dots, x^p\}$ are two sets of variables associated with the envelope and spectra, respectively, and N is the number of considered hits.

One natural tool to make the dimension reduction and then to get the x^js is the Functional Principal Component Analysis (FPCA). For example, in the case of the envelopes, this consists in representing the individuals in an orthonormal basis $\{f_k\}_{k\geq 1}$, such that c_i^j is the score associated with the i^{th} individual. We have:

$$c_i^j = \int_{-D/2}^{D/2} f_j(t)\bar{\mathcal{E}}_i(t)dt \ , \tag{17}$$

where $[\frac{-D}{2}, \frac{D}{2}]$ is the support of $\bar{\mathcal{E}}_i(t)$, with $\bar{\mathcal{E}}_i(t) = \mathcal{E}_i(t) - \mu(t)$ and $\mu(t) = \dfrac{1}{N}\sum_{k=1}^{N} \mathcal{E}_k(t)$.

If we denote by

$$\mathcal{V}(s,t) = \frac{1}{N-1}\sum_{k=1}^{N}(\mathcal{E}_k(t) - \mu(t))(\mathcal{E}_k(s) - \mu(s)) \ ,$$

the variance-covariance operator associated with the sample of envelopes, the weighting functions $f_k(t)$ are then the eigenfunctions of $\mathcal{V}$.

We notice here that in the case of the spectra, an alternative to the FPCA is to make a Correspondence Analysis (CA). Indeed, the spectra of a signal can be defined as the repartition of its energy according to the different frequency ranges.

3.1.2 Strategy 2: functional nonparametric tools

The second one is based on the use of semi-metrics which lets us compare two hits χ_i and $\chi_{i'}$, by the way of a combination of the results for the comparison of their envelopes ($\mathcal{E}_i$ and $\mathcal{E}_{i'}$) and of their spectra ($\mathcal{S}_i$ and $\mathcal{S}_{i'}$):

$$d(\chi_i(t), \chi_{i'}(t)) = \alpha\frac{d_1(\mathcal{E}_i(t), \mathcal{E}_{i'}(t))}{\beta_1} + (1-\alpha)\frac{d_2(\mathcal{S}_i(f), \mathcal{S}_{i'}(f))}{\beta_2}, \tag{18}$$

where $\alpha \in [0,1]$ lets us define the weight attributed to each curve. For example, $\alpha = 0$ (respectively, $\alpha = 1$), means that the analysis is based only on the spectra (respectively, on the envelopes). As

for the parameters β_k, they are devoted to the correction of possible scale effects between the semi-metrics d_1 and d_2. Drawing the inspiration from the normalization of the data matrices for Principal Components, we can choose

$$\beta_k = \sqrt{var(d_k)}, \tag{19}$$

where

$$var(d_1) = \frac{\sum_{i=1}^{N}\sum_{i'=1}^{N}\big(d_1(\mathcal{E}_i(t),\mathcal{E}_{i'}(t)\big) - \bar{d_1})^2}{2N^2} \quad with \quad \bar{d_1} = \frac{\sum_{i=1}^{N}\sum_{i'=1}^{N} d_1(\mathcal{E}_i(t),\mathcal{E}_{i'}(t))}{2N^2}. \tag{20}$$

Note that the optimal value of α depends on the context and the studied AE signals. In the case of a clustering problem considered here, we have chosen optimize α by maximizing the silhouette index.

3.1.3 Strategy 3: Mix of the strategies 1 and 2

Combining these two strategies, we build a third approach, which uses jointly the dimension reduction and non parametric methods. The method consists in processing a first clustering from the information contained in the spectra, and then in refining these first results with the aid of the envelopes. This order in the use of the two types of curves raises the important question of why clustering first on the spectra and then on the envelopes ? This is motivated the by fact that in nuclear safety experiment context, the envelopes are less robust to the experiment conditions than the spectra. In our context:

1. two signals which are associated with the same source mechanism cannot have very different spectra;
2. if two source mechanisms have the same spectra, their envelopes let us make difference between them.

3.2 Choice of a clustering method

According to the state of the art realized by [Ullah and Finch, 2013] the hierarchical clustering is an efficient method for functional data clustering. Here we adopt it as the principal method whatever the clustering strategy.

3.2.1 Metrics and semi-metrics

When a reduction of dimension is used, the comparison of the individuals (the hits) relies on a data matrix of dimension $N \times p$ (cf. Equation 16), such that the distance between two hits x_i and $x_{i'}$ is generally expressed as:

$$d(x_i, x_{i'}) = \left(\sum_{j\in J}(x_i^j - x_{i'}^j)^n\right)^{1/n}, \tag{21}$$

where $J = \{1,\dots,p\}$ (resp. $J = \{j^*+1,\dots,p\}$) in the case of the first strategy (resp. the third clustering strategy).

In the case of a functional nonparametric approach, [Ferraty and Vieu, 2006] propose, among other semi-metrics, the one based on derivatives of the curves, which is well adapted after a smoothing of the curves. We have:

$$d_1(\mathcal{E}_i, \mathcal{E}_{i'}) = \left(\int_{-D/2}^{D/2}(\mathcal{E}_i^{(l)}(t) - \mathcal{E}_{i'}^{(l)}(t))^2 \mathrm{d}t\right)^{1/2}$$

and

$$d_2(\mathcal{S}_i, \mathcal{S}_{i'}) = \left(\int_0^{f/2} (\mathcal{S}_i^{(l)}(f) - \mathcal{S}_{i'}^{(l)}(f))^2 \mathrm{d}f \right)^{1/2},$$

where $\mathcal{E}_i^{(l)}(t)$ and $\mathcal{S}_i^{(l)}(f)$ are the derivatives of order l of $\mathcal{E}_i(t)$ and $\mathcal{S}_i(f)$, respectively.

3.2.2 Agglomeration

In hierarchical clustering, agglomeration consists in an irreversible algorithm merging at each of its $N-1$ steps a pair of clusters, including singletons [Murtagh and Contreras, 2017]. It starts with the fine partition consisting of N clusters and ends with the trivial partition consisting in just one class.

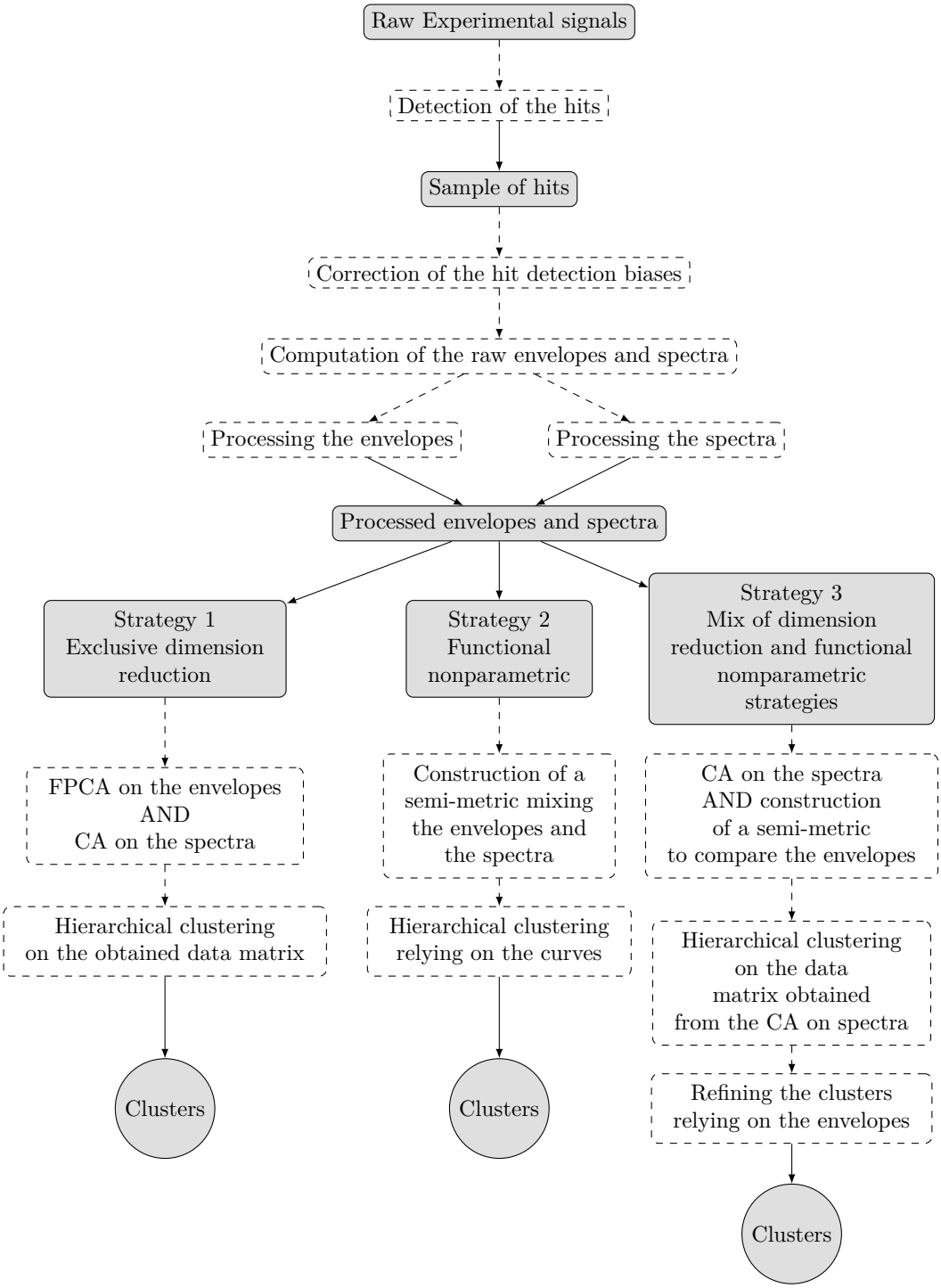
In the case of Ward's criterium which we have used in this work, the dissimilarity between two clusters A and B indicates how much the sum of squares will increase when we merge them:

$$\Delta(A, B) = \frac{N_A N_B}{N_A + N_B} d^2(\mu_A, \mu_B),$$

where N_A and N_B are the number of hits belonging to the clusters A and B, respectively, and μ_A and μ_B their respective centroïds. In usual $N-$dimensional data reduction, d is the usual euclidean distance. For functional data, d is a semi-metric which measures the proximity between appropriated curves, as presented in the previous section.

4 From raw hit signals processing to clusters: a summary

4.1 Synoptic diagram



4.2 Statistical and *a priori* criteria for clusters validation

In order to compare the different clustering strategies, we consider the criteria of the mean silhouette [Desgraupes, 2013,Ieva et al., 2013] and *a priori* information associated with two already identified clusters.
For each hit x_i, the silhouette value s_i is defined as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \ , \tag{22}$$

where a_i is the average distance according to the chosen strategy and (semi-)metric of x_i to all the other hits which are in the same cluster $\mathcal{G}_k$. And :

$$b_i = \min_{k' \neq k} \mathcal{D}(x_i, \mathcal{G}_{k'}) \ . \tag{23}$$

where $\mathcal{D}(x_i, \mathcal{G}_{k'})$ is the mean distance between i and the clusters $\mathcal{G}_{k'}$.

s_i is a quantity between -1 and 1: a value near 1 indicates that the point x_i is affected to the right cluster whereas a value near -1 indicates that the point should be affected to another cluster.

The hits corresponding to two physical mechanisms have been already identified by the experimentalists. They correspond to the failure of the fuel clad (12 hits) and the drop of the control rods (13 hits). In order to use this *a priori* information, we introduce a notion of «loss». For example, we know that 8 of 168 hits are potentially associated with clad failures. If only 5 clad failures are in the same group for a given method, we then conclude that there are 3 losses.

5 Results and discussions

5.1 Illustration with simulated data

Before considering the results of the application of the three strategies in the case of the nuclear safety experiment data, let use use first perfectly known simulated hits. For this purpose, we consider the model proposed in [Wotzka, 2014]:

$$s^2(t) = \frac{a}{1 + e^{-b(t-\mu_1)}} e^{-c(t-\mu_2)} \sum_{i=1}^{n} A_i cos(2\pi f_i t) \tag{24}$$

This model is composed of two parts. The first part is a combination of sigmoid- and exponential-type functions and describes the envelope of the signal. The second one is composed of a sum of cosine functions, each with its own magnitude and frequency, and represents the frequency component involved in the hit.

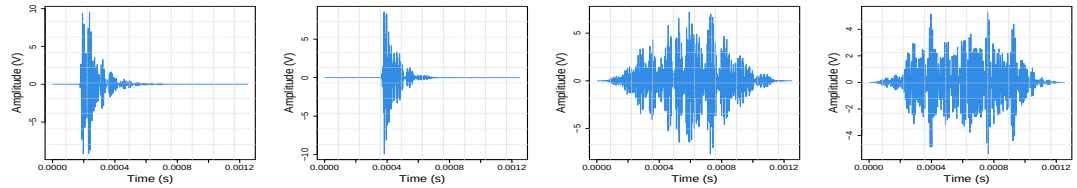
As the source mechanism corresponding to each hit is supposed to be perfectly known, there is no need to manage most of the registration problems reflected in Section 2.1. However, a horizontal shift has been randomly introduced. The model (24) has been used to create four clusters of 100 hits size. The firsts two being associated with discrete hit signals (clad failures for example) and the last two ones with continuous hit signals (gas ejections for example).

It is obvious that using only the envelopes does not allow to separate the clusters 1 and 2, respectively the clusters 3 and 4 (Figure 9b). Moreover, it is also obvious that using only the spectra does not lead to the separation of the clusters 1 and 4, respectively the clusters 2 and 3 (Figure 9c).

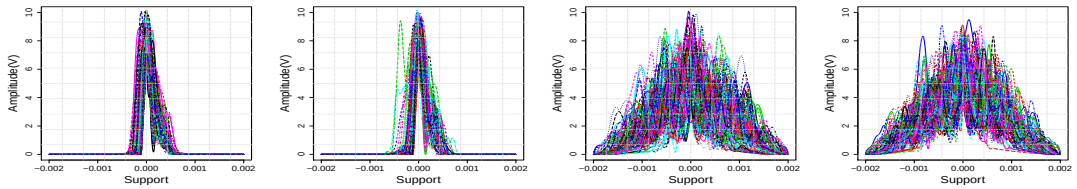
Then, this example requires a combination of the envelopes and the spectra. Table 1 shows the results of the comparison of the three strategies according to the ratio of wrong classification. It allows to conclude that for the simulated data of interest here, the three strategies lead to very good clustering results. Indeed, the ratio of wrong classification is less than 14% whatever the method and the cluster.

If one has to make a hierarchy between the three methods in this case, we can observe that the mix strategy leads to the best results. However, one has to keep in mind that the performances of the various methods depends on several parameters among which the raw data and the preprocessing strategy.

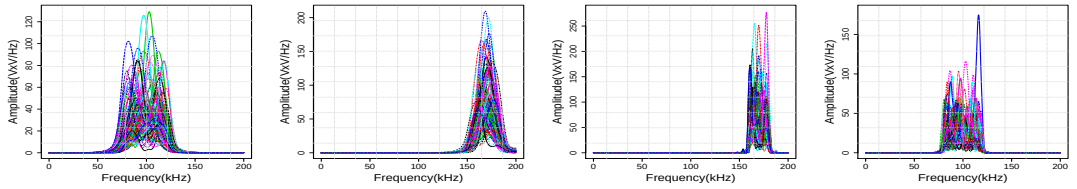
For the sake of brevity, we limit this simulation section to this unique case. The reader can rely on the proposed strategy to simulate AE data corresponding to his context and then choose the appropriate clustering strategy.



(a) Illustrative exemple of hit according to the four simulated clusters (from left to right: clusters 1 to 4)



(b) Envelopes according to the cluster (from left to right: clusters 1 to 4)



(c) Spectra according to the cluster (from left to right: clusters 1 to 4)

Fig. 9: Four clusters of simulated hits (from left to right: clusters 1 to 4).

Table 1: Ratio (in %) of wrong classification according to initial clusters

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Dimension reduction	14	2	0	0
Nonparametric Functional	0	6	2	0
Mix strategy	0	3	0	0

5.2 Case of nuclear safety experiment data

5.2.1 Comparison of the three strategies

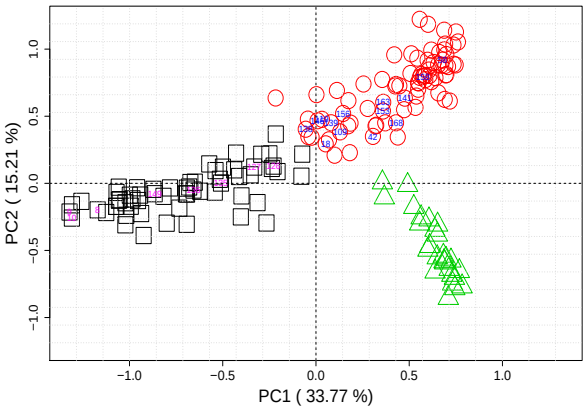
Table 2 summarizes the results of the repartition of the 168 hits into 6 clusters for the three strategies. According to the comparison criteria (mean silhouette or *a priori* information), we observe that the two best strategies for these data are the nonparametric functional and the mix one. Taking into account that the mean silhouettes are computed from different data matrices (or curves), here we prefer the physical criteria of the *a priori* information and choose the mix strategy.

Table 2: Comparative table of the results of the various approaches of unsupervised classification by hierarchical clustering in the functional case. Note that DCR means drop of control rods.

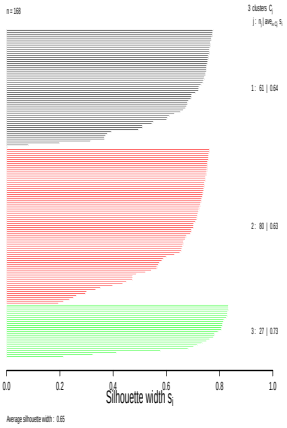
	loss of clad failures	**loss of DCR**	**mean silhouette**
Dimension reduction	4	6	0.5
Nonparametric Functional	3	6	0.56
Mix strategy	0	5	0.11

5.2.2 Detailed results of the mix strategy

Step 1: Hierarchical clustering on the data matrix obtained from the CA on the spectra. Figure 10 allows to conclude that three main profiles of spectra with high mean silhouette values (> 0.6) can be drawn. In particular, we can observe that we have no loss of clad failure or drop on the control rod. Furthermore, we know from the experimentalists that there exists at least six clusters, this highlights the fact that different source mechanisms can have same spectrum shape. This allows to conclude that relying only on the hits spectra is not enough for clustering the data considered here and then confirms the necessity to also use the envelopes.



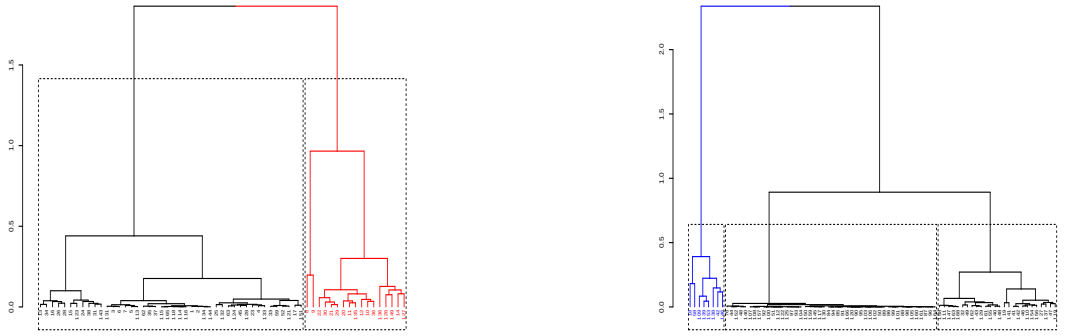
(a) Individuals in the plane ($PC1$,$PC2$) issued from the CA process. The numbers of the hits associated with a clad failure are red (numbers in the squares) and those associated with drop of the control rods are blue (numbers in the circles).



(b) Silhouettes of the hits by group.

Fig. 10: Hits map from PCA and silhouettes after clustering

Step 2: Refining the clusters by relying on the envelopes. We identify the cluster of hits of green color (third group) on Figure 10 as a very homogeneous cluster (mean silhouette > 0.70). In addition, the representation of the hits in the factor plane (1,2) allows to observe that this cluster (represented by the triangles) is well separated from the two others. We then keep it unchanged.
The analysis of the clusters 1 (represented by the squares) and 2 (represented by the circles) leads to refine them into 2 and 3 sub-clusters, respectively (Figure 11). The final clusters associated with the clad failures and the drops of the control rods contain 17 and 8, respectively. These numbers being different from those estimated by the experimentalists (12 and 13, respectively), we conclude that our statistical method reaches its limitation. In particular, the hits associated with the drops of the control rods seem difficult to isolate (Table 2). Then it is necessary to refine the clusters manually in order to reach the requirements of the experimentalists.



(a) Refining the cluster 1 into 2 sub-clusters

(b) Refining the cluster 2 into 3 sub-clusters

Fig. 11: Dendrograms associated with the refining of clusters 1 and 2 of the first step

5.2.3 Back to physics: Identification of the physical mechanisms associated with the clusters

We have just seen that in order to reach the requirements of the experimentalists, it is necessary to make a manual refining of the clusters. This being out of the scope of this work, here we restrict ourselves to the analysis of the six clusters obtained by statistical ways. The objective is to identify the physical mechanisms associated with some clusters and to give perspectives for the manual refining.

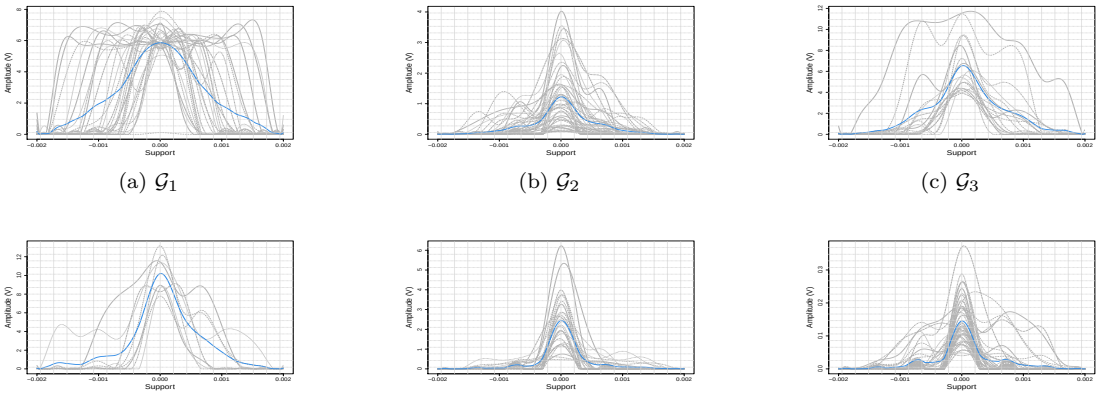
In addition to the clusters already identified, which are the clad failures (cluster $\mathcal{G}_3$) and the drops of the control rods (cluster $\mathcal{G}_4$), relatively large cluster of 46 hits ($\mathcal{G}_6$) does not correspond to physical mechanisms of interest. Indeed, the broadband spectra and the low energy of the hits belonging to this cluster allow to conclude that they are associated with noise.

The cluster $\mathcal{G}_1$ is composed of very energetic hits with a waveform (Figure 12) presenting a plate. This observation being unusual for AE signals, we conclude that they are corrupted by the limitations of the acquisition system.

Concerning the unspecified clusters $\mathcal{G}_2$ and $\mathcal{G}_5$, it is clear that they are those on which one has to focus for a manual refining. In particular, the remaining 13 hits associated with the drop of the control rods belong to the cluster $\mathcal{G}_5$.

Table 3: Obtained clusters according to the third clustering strategy

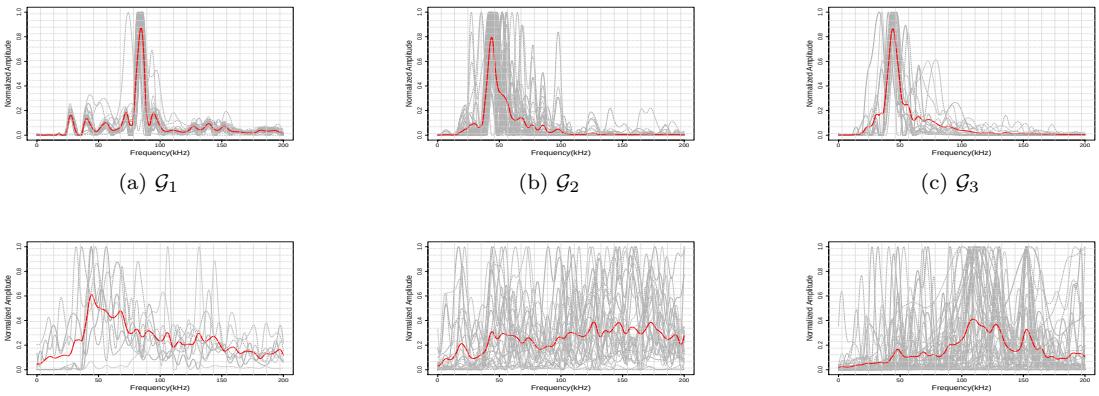
Classe	Identification of the cluster	size
$\mathcal{G}_1$	Microphone saturation	27
$\mathcal{G}_2$	Unspecified cluster 1	44
$\mathcal{G}_3$	Clad failures	17
$\mathcal{G}_4$	Drop of the control rods	8
$\mathcal{G}_5$	Unspecified cluster 2	26
$\mathcal{G}_6$	Noise	46



(a) $\mathcal{G}_1$ (b) $\mathcal{G}_2$ (c) $\mathcal{G}_3$

(d) $\mathcal{G}_4$ (e) $\mathcal{G}_5$ (f) $\mathcal{G}_6$

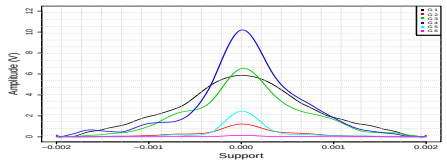
Fig. 12: Superposition of the envelopes according to the cluster



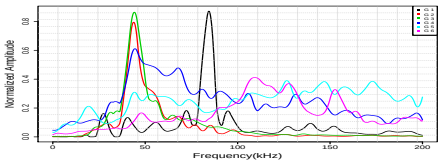
(a) $\mathcal{G}_1$ (b) $\mathcal{G}_2$ (c) $\mathcal{G}_3$

(d) $\mathcal{G}_4$ (e) $\mathcal{G}_5$ (f) $\mathcal{G}_6$

Fig. 13: Superposition of the normalized spectra according to the cluster.



(a) Superposition of the mean envelopes for the six clusters.



(b) Superposition of the normalized mean spectra for the six clusters.

Fig. 14: Superposition of the mean curves according to the cluster

6 Conclusion

The aim of this article was to explore the abilities of clustering methods based on functional statistics in order to cope with the complexity of acoustic emission signals recorded in nuclear context. To this end, we have first proposed a preprocessing strategy in order to manage classical problems (horizontal shifts, differences in curves supports...) and more specific problems (hit detection biases). Then, three strategies of clustering using the spectra and the envelopes of the hits have been proposed.

An analysis of the results based on the mean silhouette index and *a priori* information allows to conclude that a two step strategy combining the dimension reduction and the use of semi-metrics is particularly well adapted in our context. We obtain six homogeneous clusters and conclude that both spectra and envelopes are relevant to use in this study.

From a physical point of view, in addition to the already identified physical mechanisms which are the clad failures and the drops of the control rods, our work have led to isolate two more clusters. The microphone saturations, reflecting the limits of the signal acquisition system and noises corresponding to the hits having no physical interest. However, two clusters ($\mathcal{G}_2$ and $\mathcal{G}_5$) associated with unknown physical mechanisms still remain to identify. If it is difficult to unambiguously assign a physical mechanism to the cluster $\mathcal{G}_2$, the analysis of the spectra of its hits and the moment at which they occur during the experiments allow to make some assumptions. One can observe that the shape of the spectra is approximately the same as those of the cluster $\mathcal{G}_3$ corresponding to the clad failures. Moreover, most of the hits of this cluster occurred during experiments in which there has effectively been a clad failure. Then, these hits correspond to physical mechanisms which are precursor or residual of a clad failures. Unfortunately this reasoning is not decisive in the case of the cluster $\mathcal{G}_5$. A perspective is to perform measurements of acoustic wave propagation in realistic mocks and then identify which physical mechanism could be associated with this cluster.

If the results prove that functional methods are effective for this king of complex curves, several questions remain open. We have seen that the proposed clustering method reaches its limitations, in particular to isolate in the same clusters the mechanism associated with the drop of the control rods. This paves the way to the experimentation of more semi-metrics or dimension reduction methods beyond the default ones used in this work. Furthermore, the physical analysis of the obtained clusters have conducted to the need of refining manually the obtained clusters. The result of this refining would allow to get a definitive response variable and to optimize the settings of the different tools (preprocessing, smoothing, dimension reduction and semi-metric construction). This also leads to the perspective of comparative study with the methods classically used in acoustic emission.

References

Abraham et al., 2003. Abraham, C., Cornillon, P. A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using b-splines. *Scandinavian journal of statistics*, 30(3):581–595.

Aguilera et al., 2006. Aguilera, A. M., Escabias, M., and Valderrama, M. J. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50(8):1905–1924.

Ai et al., 2010. Ai, Q., Liu, C., Chen, X., He, P., and Wang, Y. (2010). Acoustic emission of fatigue crack in pressure pipe under cyclic pressure. *Nuclear Engineering and Design*, 240(10):3616–3620.

Anastassopoulos and Philippidis, 1995. Anastassopoulos, A. A. and Philippidis, T. P. (1995). Clustering methodology for the evaluation of acoustic emission from composites. *Journal of Acoustic Emission*, 13(1-2):11–22.

Boudou and Viguier-Pla, 2017. Boudou, A. and Viguier-Pla, S. (2017). Commutator of projectors and of unitary operators. In *Functional Statistics and Related Fields*, pages 67–75. Springer.

Cardot et al., 1999. Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45(1):11–22.

Chiou et al., 2004. Chiou, J. M., Müller, H. G., and Wang, J. L. (2004). Functional response models. *Statistica Sinica*, pages 675–693.

Cuesta-Albertos and Febrero-Bande, 2010. Cuesta-Albertos, J. A. and Febrero-Bande, M. (2010). A simple multiway anova for functional data. *Test*, 19(3):537–557.

Cuevas, 2014. Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23.

Desgraupes, 2013. Desgraupes, B. (2013). Clustering indices. *University of Paris Ouest-Lab Modal'X*, 1:34.

Favretto-Cristini et al., 2016. Favretto-Cristini, N., Hégron, L., and Sornay, P. (2016). Identification of the fragmentation of brittle particles during compaction process by the acoustic emission technique. *Ultrasonics*, 67:178–189.

Febrero-Bande et al., 2017. Febrero-Bande, M., González-Manteiga, W., and de la Fuente, M. O. (2017). Variable selection in functional additive regression models. In *Functional Statistics and Related Fields*, pages 113–122. Springer.

Ferraty and Vieu, 2003. Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1):161–173.

Ferraty and Vieu, 2006. Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Friedman et al., 2001. Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer Series in Statistics, Berlin.

Gautschi, 2002. Gautschi, G. (2002). *Piezoelectric sensorics: force, strain, pressure, acceleration and acoustic emission sensors, materials and amplifiers*. Springer Science & Business Media.

Goia and Vieu, 2016. Goia, A. and Vieu, P. (2016). An introduction to recent advances in high/infinite dimensional statistics.

Huckemann and Eltzner, 2017. Huckemann, S. F. and Eltzner, B. (2017). Essentials of backward nested descriptors inference. In *Functional Statistics and Related Fields*, pages 137–144. Springer.

Ieva et al., 2013. Ieva, F., Paganoni, A. M., Pigoli, D., and Vitelli, V. (2013). Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):401–418.

Jacques and Preda, 2014. Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255.

Jernkvist and Massih, 2010. Jernkvist, L. O. and Massih, A. R. (2010). Nuclear fuel behavior under reactivity-initiated accident (ria) conditions. Technical report, Nuclear Energy Agency.

Jiang et al., 2017. Jiang, Q., Meintanis, S. G., and Zhu, L. (2017). Two-sample tests for multivariate functional data. In *Functional Statistics and Related Fields*, pages 145–154. Springer.

Keyvan and Nagaraj, 1996. Keyvan, S. and Nagaraj, J. (1996). Pattern recognition of acoustic signatures using art2: A neural network. *Journal of Acoustic Emission*, 14(2):97–102.

Lila et al., 2017. Lila, E., Aston, J. A., and Sangalli, L. M. (2017). Functional data analysis of neuroimaging signals associated with cerebral activity in the brain cortex. In *Functional Statistics and Related Fields*, pages 169–172. Springer.

Murtagh and Contreras, 2017. Murtagh, F. and Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, ii. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(6).

Pantera and Traore, 2015. Pantera, L. and Traore, O. I. (2015). Reproducible data processing research for the CABRI RIA experiments acoustic emission signal analysis. In *4th International Conference on Advancements in Nuclear Instrumentation Measurement Methods and their Applications (ANIMMA)*, pages 1–8. IEEE.

Preda et al., 2007. Preda, C., Saporta, G., and Lévéder, C. (2007). Pls classification of functional data. *Computational Statistics*, 22(2):223–235.

R Development Core Team, 2008. R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramsay and Silverman, 2005. Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Science & Business Media.

Ramsay and Silverman, 2002. Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. New York, NY: Springer New York.

Roget, 1988. Roget, J. (1988). *Essais non destructifs: L'émission acoustique: Mise en œuvre et applications*. AFNOR; CETIM.

Rossi et al., 2004. Rossi, F., Conan-Guez, B., and El Golli, A. (2004). Clustering functional data with the som algorithm. In *ESANN*, pages 305–312.

Rossi and Villa, 2006. Rossi, F. and Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing*, 69(7):730–742.

Rudling et al., 2016. Rudling, P., Jernkvist, L. O., Garzarolli, F., Adamson, R., Mahmood, T., Strasser, A., and Patterson, C. (2016). Nuclear fuel behaviour under ria conditions. *Advanced Nuclear Technology International*.

Scrucca, 2004. Scrucca, L. (2004). qcc: an r package for quality control charting and statistical process control. *R News*, 4/1:11–17.

signal developers, 2014. signal developers (2014). *signal: Signal processing*.

Sueur et al., 2008. Sueur, J., Aubin, T., and Simonis, C. (2008). Seewave: a free modular tool for sound analysis and synthesis. *Bioacoustics*, 18:213–226.

Traore et al., 2017a. Traore, O. I., Cristini, P., Favretto-Cristini, N., Pantera, L., Vieu, P., and Viguier-Pla, S. (2017a). Contribution of functional approach to the classification and the identification of acoustic emission source mechanisms. In *Functional Statistics and Related Fields*, pages 251–259. Springer.

Traore et al., 2017b. Traore, O. I., Cristini, P., Favretto-Cristini, N., Pantera, L., and Viguier-Pla, S. (2017b). Impact of the test device on the behavior of the acoustic emission signals: Contribution of the numerical modeling to signal processing. In *5th International Conference on Advancements in Nuclear Instrumentation Measurement Methods and their Applications (ANIMMA)*. IEEE.

Traore et al., 2017c. Traore, O. I., Favretto-Cristini, N., Pantera, L., Cristini, P., Viguier-Pla, S., and Vieu, P. (2017c). Which methods and strategies to cope with noise complexity for an effective interpretation of acoustic emission signals in noisy nuclear environment ? *Acta-Acustica united with Acustica*.

Traore et al., 2017d. Traore, O. I., Pantera, L., Favretto-Cristini, N., Cristini, P., Viguier-Pla, S., and Vieu, P. (2017d). Structure analysis and denoising using singular spectrum analysis: application to acoustic emission signals from nuclear safety experiments. *Measurement*, 104:78 – 88.

Ullah and Finch, 2013. Ullah, S. and Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 13(1):43.

Walders and Liebl, 2017. Walders, F. and Liebl, D. (2017). Parameter regimes in partially functional linear regression for panel data. In *Functional Statistics and Related Fields*, pages 261–270. Springer.

Wotzka, 2014. Wotzka, D. (2014). Mathematical model and regression analysis of acoustic emission signals generated by partial discharges. *Applied and Computational Mathematics*, 3(5):225–230.