Engineering Structures 56 (2013) 273–285

Contents lists available at SciVerse ScienceDirect

Engineering Structures

journal homepage: www.elsevier.com/locate/engstruct

Multivariate statistical analysis for early damage detection

João Pedro Santos [a,*], Christian Crémona [b], André D. Orcesi [c], Paulo Silveira [a]

[a] LNEC, Structure's Department, Av. Brasil 101, 1700-066 Lisbon, Portugal
[b] SÉTRA, Technical Center for Bridge Engineering, B.P. 214, 77487 Provins Cedex, France
[c] IFSTTAR, Bridges and Structures Department, 58 boulevard Lefebvre, F-75732 Paris Cedex 15, France

ARTICLE INFO

Article history:
Received 3 September 2012
Revised 8 April 2013
Accepted 13 May 2013

Keywords:
Structural Health Monitoring
Early-damage detection
Principal component analysis
Symbolic data
Symbolic dissimilarity measures
Cluster analysis
Numerical model
Damage simulations

ABSTRACT

A large amount of researches and studies have been recently performed by applying statistical methods for vibration-based damage detection. However, the global character inherent to the limited number of modal properties issued from operational modal analysis may be not appropriate for early damage, which has generally a local character.

The present paper aims at detecting this type of damage by using static SHM data and by assuming that early damage produces dead load redistribution. To achieve this objective a data driven strategy is proposed, consisting in the combination of advanced multivariate statistical methods and quantities, such as principal components, symbolic data and cluster analysis.

From this analysis it was observed that, under the noise levels measured on site, the proposed strategy is able to automatically detect stiffness reduction in stay cables reaching at least 1%.

1. Introduction

Structural Health Monitoring (SHM) can be defined as the development and application of strategies to identify abnormal behaviors (such as damage) in structural systems [1]. In civil engineering structures, damage may lead to expensive maintenance actions and, if it occurs with significant magnitude, may result in dramatic social and human consequences. An efficient SHM should aim at identifying damage in an early stage, which is generally related to local phenomena with small magnitude.

Damage identification has been extensively studied in the framework of mechanical, airspace and civil engineering structural systems by using model based or data driven approaches [1,2]. The first type typically aims at identifying damage by fitting a numerical model to real data, a procedure which is usually combined with optimization techniques. Conversely, data driven approaches are usually based on data processing obtained from monitoring without relying on *a priori* models.

Damage detection has been described in the literature as a four-level scale [3]: (i) damage detection, (ii) localization, (iii) type and severity assessment and (iv) lifetime prediction update. While the first and second levels can be carried out by data driven methods alone, the fourth (and partly the third stage) level requires the use of numerical models. The last two levels may also require local non-destructive testing, visual inspection, human expertise and additional theoretical concepts such as fracture mechanics or fatigue analysis to enhance the damage detection analysis [4].

This paper is mainly focused on the first level of the previous scale (damage detection) by means of data driven techniques: early damage detection is targeted. To fulfill this objective, three main operations are required after data acquisition [5]: (i) feature extraction, (ii) data normalization and (iii) statistical classification.

Damage sensitive feature extraction is mandatory for early damage detection approaches since the acquired data, alone, may not be informative about the presence of damage. Modal or modal-based quantities are by far the most reported features in the literature [6–8]. Autoregressive models [9–11] and wavelet components [9,12,13] have also been reported as damage sensitive feature extractors for both static and dynamic monitoring. Principal component analysis (PCA) is usually applied after feature extraction for dimensionality reduction. However, for long-term static monitoring, Posenato et al. [9], showed that damage sensitive features can also be obtained from this multivariate statistical method. Symbolic data have shown to be useful in compressing and representing data without loss of information [14] and proved to be a sensitive feature extractor, by Cury et al. [15,16], in SHM works applied to bridges.

[*] Corresponding author. Address: LNEC, National Laboratory for Civil Engineering, Monitoring of Structures Division, Structure's Department, Av. Brasil 101, 1700-066 Lisbon, Portugal. Tel.: +351 964570563; fax: +351 218443025.

E-mail addresses: josantos@lnec.pt (J.P. Santos), christian.cremona@developpement-durable.gouv.fr (C. Crémona), andre.orcesi@ifsttar.fr (A.D. Orcesi), paulo.silveira@lnec.pt (P. Silveira).

0141-0296/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engstruct.2013.05.022

Data Normalization can be defined as the process of separating data changes caused by environmental and operational effects from those caused by damage occurrences [17]. This process is crucial for early damage detection and false alarm prevention since environmental conditions, such as temperature, may impose larger variations than those due to damage [18,19]. It was observed that modal quantities can change by 17% [20] due to seasonal temperature and by 5% on a single daily cycle [21], a fact which is mainly due to changes in structural stiffness and boundary conditions [17]. For static based monitoring, temperature can impose complex and important structural changes, which are generated by the induced strain on each structural element. Assuming that dynamic effects like traffic and wind are properly filtered *in situ* by on-line monitoring systems [22–24], this action is usually the only requiring normalization.

Several normalization strategies are available based in regression methods such as Multivariate Linear Regression, Multi-Layer Perceptron Neural Networks or Support Vector Regression [16,18,19,25,26]. However, these are greatly dependent on an appropriate characterization of temperature effects throughout the target structure. When monitoring applications do not include broad temperature measurement, latent variable methods may be more efficient in normalizing temperature effects. These methods are able to characterize and suppress independent actions and effects using only structural measurements. Among these, the principal component analysis (PCA) [6,27] has been found very efficient, even if restricted to linear effects.

Statistical classification aims at distinguishing damaged from undamaged related data. It can be divided into supervised methods, which require knowledge (sensitive features) from both damaged and undamaged states, and unsupervised ones, which are frequently used to detect deviations from an undamaged baseline structural state (a procedure known as novelty detection). Since data obtained from damaged structures is scarce or inexistent, unsupervised techniques have been used more often for damage detection purposes [5]. Most of these are based in outlier detection, carried out after a training procedure over a period in which structures are assumed undamaged. This procedure is known as Statistical Process Control and can be performed on single variables [6,28] or multivariate data sets [29,30]. Cluster analysis can be seen as an alternative to this approach since it enables to distinguish different groups in data without any prior knowledge or known reference baseline. Even though this type of unsupervised analysis has been reported has an efficient damage detection approach [31,32], the limited structural representativeness of measured data (static or dynamic) and its computational complexity have discouraged its use in SHM of large civil structures. To circumvent these disadvantages, symbolic dissimilarity measures have been used as input on cluster analysis, providing greater structural representativeness and smaller computational complexity [7,16,33].

The large majority of data-driven damage detection methods are vibration-based approaches [4,5,7,8,28]. Based on the assumption that damage produces changes in structural stiffness, feature extraction is made with frequencies, mode shapes or damping ratios. The global character of these features makes damage detection algorithms less sensitive to early damage which has, in general, a local character [1,5,34]. Furthermore, dynamic SHM systems require fast, solid-state relay switching data loggers, with dedicated high resolution analog-to-digital converters (ADC) and precision accelerometers, which are not only sensitive to ambient noise but also make SHM systems extremely expensive.

Static based SHM systems can be deployed with less cost. But, surprisingly, not so many works report damage detection using static SHM data in civil structures. Recently, numerical detection and localization was performed under the principle that damage produces changes in measured effects generated by dead loads [1,35]. However, its application was carried out based in optimization procedures and on simple theoretical structures. A similar optimization procedure, based on the same principle, was applied to a cable-stayed bridge numerical model [34] considering, as monitored quantities, the forces in all the stay cables of a bridge and without taking into account environmental influences. This method did not gain attention in the SHM active fields of mechanical or aerospace systems since its application requires that dead load effects must prevail significantly above the remaining ones [1], a fact which, in general, is only observed in civil engineering structures.

This paper attempts to provide answers to the above questions:

- development of an automatic unsupervised data driven strategy for early damage detection by continuously controlling dead load redistribution effects,
- feature extraction by selecting appropriate principal components,
- data normalization by eliminating spurious principal components,
- statistical classification by transforming features into symbolic objects and by enhancing detection with cluster analysis.

To address all these topics and to show the efficiency of the developed procedure, a cable stayed bridge located at the South of the Iberian Peninsula was used as case study. A numerical model was performed and calibrated according to experimental modal data [36]. Damage occurrence was simulated by performing finite element time-history analysis using, as input, time series of real temperature and noise effects measured on site. Section 2 of the paper describes the case study, the numerical model and the damage simulation procedure while Section 3 details the feature extraction/data normalization/damage classification procedure. In Section 4, conclusions are drawn by discussing the obtained results.

2. Case study – International Bridge over River Guadiana

2.1. Description of the structure and the SHM system

The International Bridge over River Guadiana (Fig. 1) is a cable-stayed bridge located in the South West of the Iberian Peninsula, and was built to connect the regions of Algarve (Portugal) and Andalucía (Spain). The bridge has a central span of 324 m and two lateral and transition spans of 135 m and 36 m, respectively. The deck is a pre-stressed concrete box girder with 18 m wide and 2.5 m high, which is suspended by one hundred and twenty-eight stay cables that are composed by individually sheathed mono strands, varying from 22 to 55 (Fig. 1a). Their length varies from 48 m to 167 m, and the stay cables are equally spaced every 9.0 m on the deck and every 1.8 m on the pylons. Shorter cables are clamped at mid length while longer at third length. The A-shaped pylons are 95 m and 96 m high and consist in concrete hollow sections which, besides anchoring the cables, support the deck at a height of 35 m by means of hollow section transverse beam.

The bridge was open to traffic in 1991 and was the target of extensive studies prior, during and after construction. In addition, a permanent monitoring system consisting of acoustic strain gages and resistance thermometers was installed for periodic manual data acquisition. In December 2010 an autonomous on-line SHM system [22–24], was installed on the bridge with the aim of carrying out early damage detection thus contributing to an increase in safety and to a reduction of maintenance costs. Sensors' location (Fig. 1b) was based on the principle that any damage in a cable stayed bridge may be revealed by load redistribution in cables' ten-

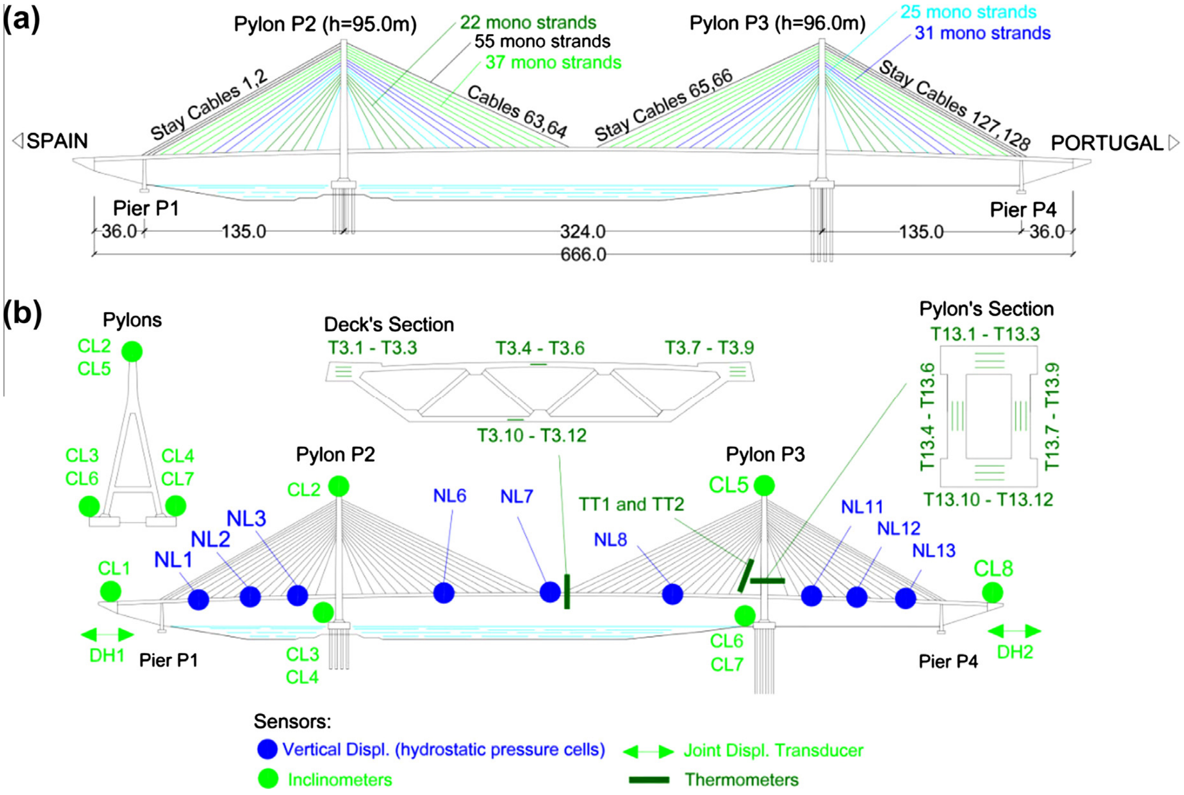

Fig. 1. International Bridge over River Guadiana: (a) side schematic view and (b) SHM sensors.

sion, with displacement and rotation changes close to the anchorages. Load cells, which would directly measure the cables' forces, were not installed and are usually avoided due to operational, economical and applicability constraints. Instead, hydrostatic pressure cells (named as NL throughout the present paper) and magnetostrictive transducers (named herein as DH) were respectively used for measuring deck and joint displacements. Bi-axial inclinometers (named herein as CL) were installed on the top of the pylons. Infrastructure differential displacements and rotations are also controlled using the same type of inclinometers (CL), installed in each foundation and abutment (Fig. 1b).

Data acquisition is carried out synchronously and hourly by a locally deployed industrial computer. Each hour, scale and covariance robust estimators are computed and compared to thresholds established by outlier and goodness-of-fit statistical tests. This strategy successfully removes values related to sensor malfunction and dynamical effects such as wind and traffic [22–24]. Data is sent daily to an FTP server which is hourly queried by a routine that automatically stores data in a MySQL database server [22].

2.2. Numerical simulation

Structural behavior is simulated, in the present work, by running finite element time-history analysis using only experimental data as input. The tri-dimensional numerical model is geometrically and physically linear for the sake of computational simplicity and is composed of (Fig. 2) 404 beam elements and 543 nodes, reproducing the geometry of the original design. Stay cables were defined as beam elements free of bending moments and compression. Piles' shafts were continuously restrained by linear elastic Winkler springs with stiffness values varying from 24 MN/m to 105 MN/m, according to the design studies. The Young Modulus and unitary weights were defined as E_c = 42 GPa and γ_c = 25 kN/m^3 for concrete and E_{sp} = 195 GPa and γ_{sp} = 78.5 kN/m^3 for stay cable steel [36]. Coefficients of linear thermal expansion are α_{sp} = 1.2 °C^{-1} and α_c = 1.0 °C^{-1} for stay cable steel and concrete, respectively.

To guarantee that the numerical model accurately simulates the structural behavior of the bridge, its natural frequencies were compared with those identified in the last experimental modal analysis performed and reported in [36]. Fig. 3 shows the good agreement between the natural frequencies obtained using the developed numerical model and the ones obtained experimentally. In Table 1, both quantities are presented as well as the associated fitting error, *FE*, defined in percentage as,

$$FE_i = |f_{i,\exp} - f_{i,num}|/f_{i,\exp} \times 100 \tag{1}$$

where $f_{i,num}$ and $f_{i,\exp}$ are the numerical and experimental frequencies obtained for mode *i*. The average value of this error, across the identified mode shapes, is 1.76%.The time-history numerical simulation aims at reproducing, as truthfully as possible, the structural behavior using as input measured temperature and noise data (Fig. 4). Measured temperature data (average temperatures in deck, pylons and cables, and differential temperature in the latter two) are used as input in the numerical time-history simulation (Fig. 5a–c). This analysis generates simulated displacements and rotations (Fig. 5e). To obtain the most similar and trustworthy reproduction of the real SHM data, the uniformly distributed noise (Fig. 5d) measured on site by the sensors is added to the numerical output.Uniform noise distributions with 3″ and 0.1 mm spans were observed for rotations and displacements, respectively. Each time series used in the numerical simulation is constituted by over 6500 data points, spanning a 9 months period (11th of January 2011 to the 1st of February 2012). The gaps shown in the time series of Figs. 5a, e and 6 are related to maintenance actions carried out by the structure's owner on the bridge's power supply system.

Damage is simulated by applying controlled temperature time series (Fig. 5b) to selected stay cables. The applied temperature

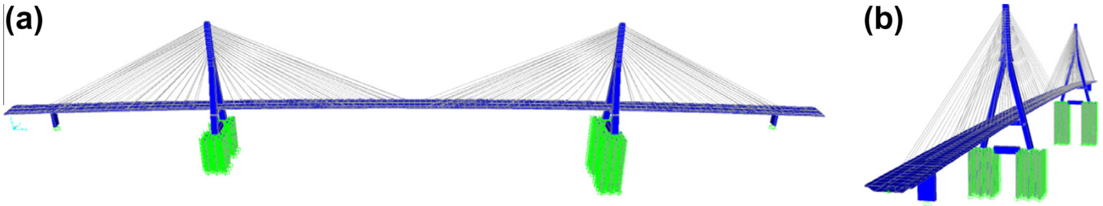

Fig. 2. Numerical model: (a) lateral view and (b) perspective view.

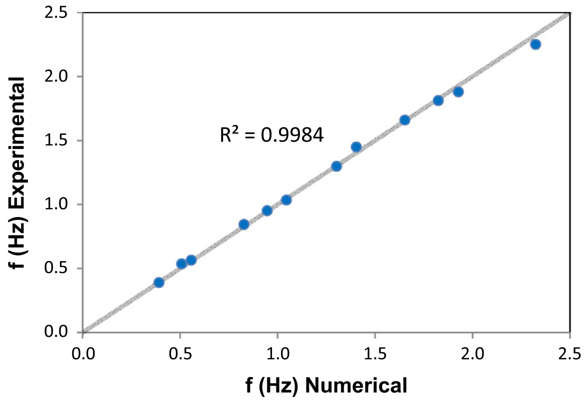

Fig. 3. Experimental and numerical natural frequencies.

Table 1
Experimental and numerical natural frequencies, mode shapes and fitting error.

Numerical model		Experimental		*FE* (%)	Mode type
Mode	*f* (Hz)	Mode	*f* (Hz)		
1	0.391	1	0.391	0.00	1st VS
2	0.508	2	0.537	5.49	1st LS
3	0.557	3	0.566	1.66	1st VAS
4	0.827	4	0.845	2.11	2nd VS
5	0.946	5	0.952	0.60	2nd VAS
6	1.045	6	1.035	0.95	3rd VS
7	1.302	7	1.299	0.25	3rd VAS
8	1.403	8	1.450	3.21	1st LAS
11	1.652	12	1.660	0.46	4th VS
13	1.824	14	1.812	0.65	4th VAS
15	1.927	15	1.880	2.48	5th VS
19	2.323	20	2.251	3.21	5th VAS

VS – vertical symmetric; LS – lateral symmetric; VAS – vertical anti-symmetric; LAS – lateral anti-symmetric.

values were calculated to reproduce equivalent stiffness losses under dead load. This numerical simulation procedure is used to accurately reproduce damage scenarios and to test the novel data driven proposed strategy.

To conduct the analyses carried out herein and explained in the following section, time-series of 15 structural measurements were obtained for several numerically simulated scenarios. Fig. 6 presents these series for an undamaged scenario. The locations of the 15 sensors, installed in the real structure, can be found in Fig. 1, where CL2 and CL5 are bi-axial inclinometers capable of measuring rotations along the longitudinal and transversal horizontal axes of the structure (suffixes "L" and "T" in Fig. 6, respectively). By comparing Figs. 5a and 6, a high correlation between the 15 simulated structural quantities and the temperatures measured on site can be observed.

3. Damage detection strategy

3.1. Principal component analysis

Principal component analysis (PCA), or Karhunen–Loève transform, is a well-known multivariate statistical method which allows obtaining, from a group of correlated variables, a set of linearly uncorrelated vectors called principal components or scores [37]. In static SHM, where measurements are highly correlated, this method can be useful to distinguish, without significant computational complexity, the uncorrelated ("independent") effects generated by different loads acting on a structure.

Let us consider a centered data matrix X^{nxp}, with n measurements performed in p sensors, The PCA consists in a linear mapping between the original variables, X^{nxp}, and the principal components, Y^{nxp}, as follows,

$$Y^{nxp} = U^{pxp}(X^{nxp})^T \qquad (2)$$

where the orthonormal linear transformation matrix, U^{pxp}, is given by the solution of an eigenproblem formulated on the covariance or correlation matrices of X^{nxp}. When using the covariance matrix of X^{nxp}, the eigensolution takes into account the scales of different variables, a fact that can bias the multivariate analysis of data acquired from different types of sensors (with distinct measurement

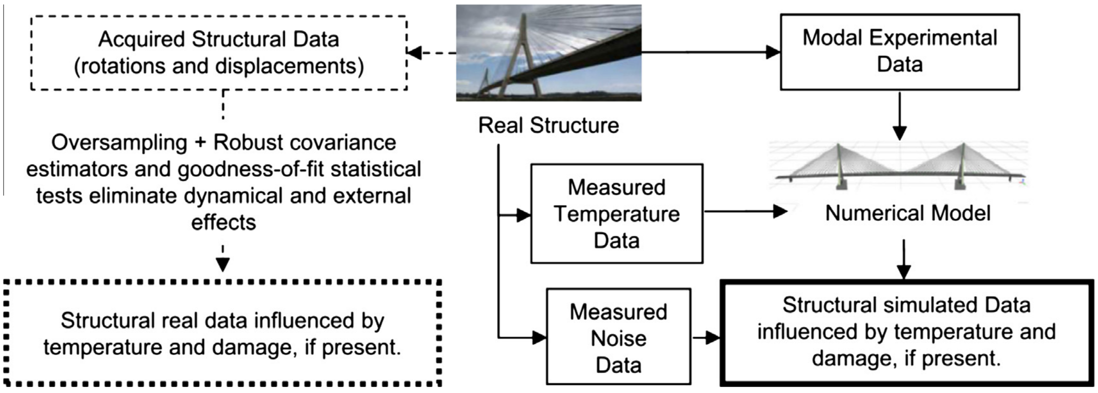
Acquired Structural Data
(rotations and displacements)
Oversampling + Robust covariance estimators and goodness-of-fit statistical tests eliminate dynamical and external effects
Structural real data influenced by temperature and damage, if present.
Real Structure
Modal Experimental Data
Measured Temperature Data
Numerical Model
Measured Noise Data
Structural simulated Data influenced by temperature and damage, if present.

Fig. 4. Acquisition of real and numerical structural data.

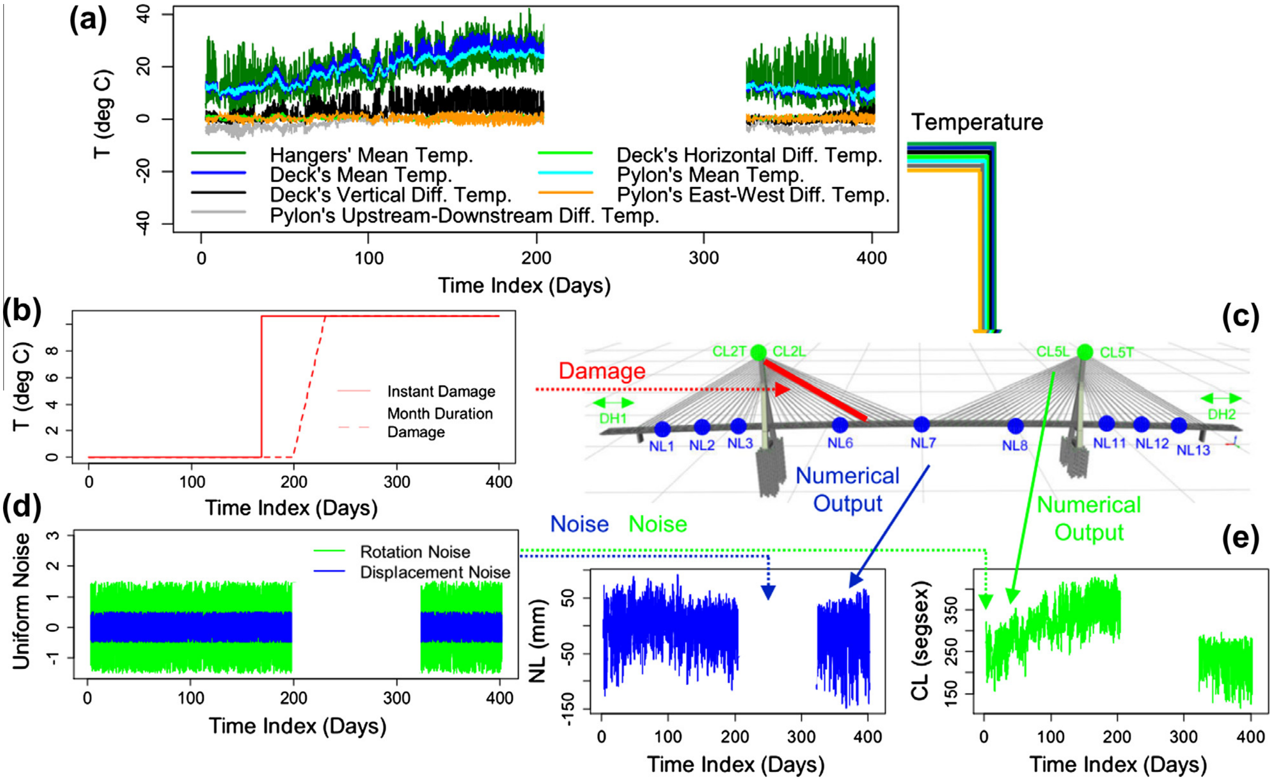

Fig. 5. Numerical simulation procedure: (a) temperature input, (b) damage time series, (c) numerical model, (d) noise time series, and (e) numerical output time series. Time index origin (day 0): 11th January 2011.

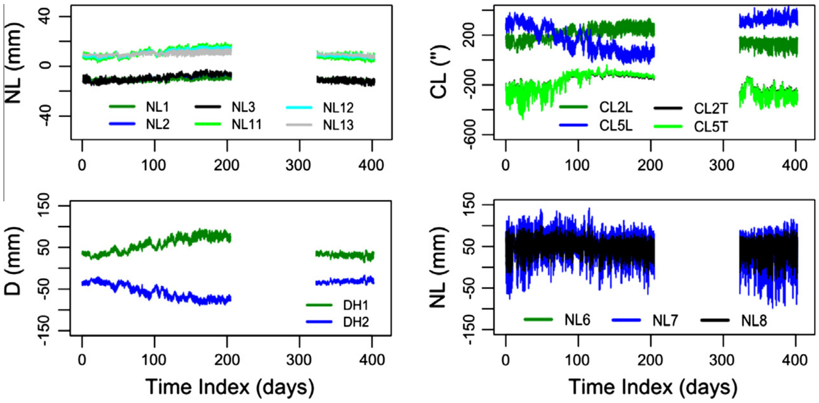

Fig. 6. Numerically generated measurements for the undamaged scenario. Time index origin (day 0): 11th January 2011.

magnitudes). When the correlation matrix C^{pxp} is used, a standardized form of PCA [37] is then considered. Such a form is equivalent to standardizing the original variables prior to the computation of the covariance matrix. Hence, since the case study addressed herein is based on data acquired from several types of sensors (as most of monitoring applications), PCA is applied using the correlation matrix, C^{pxp}, as shown in

$$C^{pxp}U^{pxp} = \Lambda^{pxp}U^{pxp} \tag{3}$$

The Lambda matrix is a diagonal matrix with positive or null values that are the eigenvalues of the correlation matrix. The transformation shown in Eq. (2) is defined such that each column of Y^{nxp} (called principal component – PC) corresponds to the higher value of Lambda under the constraint of orthogonality to the preceding ones. Hence, the elements of the diagonal matrix Λ^{pxp} are usually placed in descending order. These values express the relative importance of each principal component in the entire data set variation [6] and are usually named as "active energies".

In static SHM, early damage can be outlined by PCA based on the principle that, since it is generated by local dead load redistribution, it produces distinct variations from environmental loading acting globally on the structure [38]. From the damage simulation it was observed that, regardless the fact that each PC is a linear combination of all original variables (Eq. (2)), local dead load redistribution produces clear shifts in one or two successive principal component(s). Moreover, these simulations also allowed observing that larger damage magnitudes produce more sensitive shifts in higher "active energy" scores. These remarks were obtained by observing the sets of all principal components obtained for each damage scenario simulated. Examples of shifts are given in Fig. 7 for damage simulations on stay cable 78: this cable is located in the central span at approximately mid-fan and is sustaining a dead load force of 2676 kN. Simulated stiffness reductions consist in 1%, 2%, 5% and 10%, occurring instantly on the 1st of July, 2011. For this stay cable, the stiffness reduction percentages correspond to the rupture of approximately 2, 4, 11 and 22 wires respectively, out of the 217 wires gathered in the 31 mono strands.

To assess which principal components seem to be affected by local dead load redistribution, Kolmogorov–Smirnov (K–S) goodness-of-fit tests [39] are performed on each of the 15 principal components. Each K–S test is performed on pairs of principal components of the same order. In each pair, one of the components is obtained from the undamaged reference state and the other from one of the damage scenarios. The results of this analysis are summarized in Fig. 8, where a p-value close to 1.0 suggests that the principal component is independent from the corresponding damage scenario. For the undamaged scenario, it can be observed that all principal components are reported as independent from damage, with p-values equal to 1.0. From this figure it can also be observed that, for the four damage magnitudes tested, the first five principal components seem not to be influenced by damage and are therefore assumed to be related to global variations caused by temperature.

As observed in Fig. 8, PCA is able to retain meaningful information, related to global effects such as temperature, in the first axes whereas variations related to measurement inaccuracy, noise or other small magnitude effects such as early damage, may be summarized in latter axes. However, the issue of determining whether or not a given axis summarizes meaningful variation remains unclear in many cases. When the correct number of principal components is not retained for subsequent analysis, either relevant information is lost (underestimation) or surplus effects are included (overestimation), causing sensitivity of damage detection algorithms to decrease. Determining the number of meaningful principal components remains one of the greatest challenges in providing a truthful interpretation of multivariate data. This has been a long-standing issue in both biological and statistical literature, and a variety of stopping rules have been proposed for its estimation without resorting to external comparison or baseline information [37,40]. These rules include the establishment of eigenvalues' distributions which are directly compared to the ones extracted from data. Principal components extracted from data exhibiting greater values than the ones provided by the rules are considered meaningful [41]. Since the aim of the present work is to detect early damage, which has generally a local character, the normalization procedure consists, thus, in removing the meaningful principal components from the data set and retaining the remaining ones for subsequent statistical analysis.

Among the different stopping rules found in the literature, the Kaiser–Guttmann parameter [41] is very popular: it consists in considering as meaningful only principal components with eigenvalues, λ_k, larger than 1 (constant eigenvalue distribution across all components). The Kaiser–Guttmann parameter was tested on the five simulated scenarios and returned three as the amount of meaningful principal components related to global effects, as can be observed in Fig. 9. According to this rule, data normalization would consist in removing these components, a result which is not in agreement with the baseline analysis performed using the K–S statistical test.

Another stopping rule is based in the Broken-Stick method. The idea of this rule is that if a stick is randomly broken into p pieces, b_1 would be the expected value of the largest piece in each set of broken sticks; b_2 the expected value of the second largest piece, and so on. In the case of correlation matrices (i.e., standardized variables), p equals both the number of components and the sum of the eigenvalues, λ_k, and the proportion of total variation associated with the eigenvalue of the kth component, according to the Broken-Stick model, is obtained from

$$b_k = b(p,k) = \frac{1}{p}\sum_{i=k}^{p}\frac{1}{i} \tag{4}$$

If the kth component has an eigenvalue larger than b_k, then the component is considered as related to global effects and removed for data normalization. The factor $1/p$ is included in Eq. (4) when the correlation matrix of X^{nxp} is used to perform PCA. When the covariance matrix is used, this factor is suppressed.

One the main advantages of both rules described herein is their non-dependence of the acquired data, making the choice of the number of principal components independent from the existence of damage. While the Kaiser–Guttmann is based on a simple threshold ($\lambda_k \geqslant 1$) regardless the value of p, the Broken-Stick method takes only this parameter into account for the definition of the Broken-Stick eigenvalue distribution. Considering that a

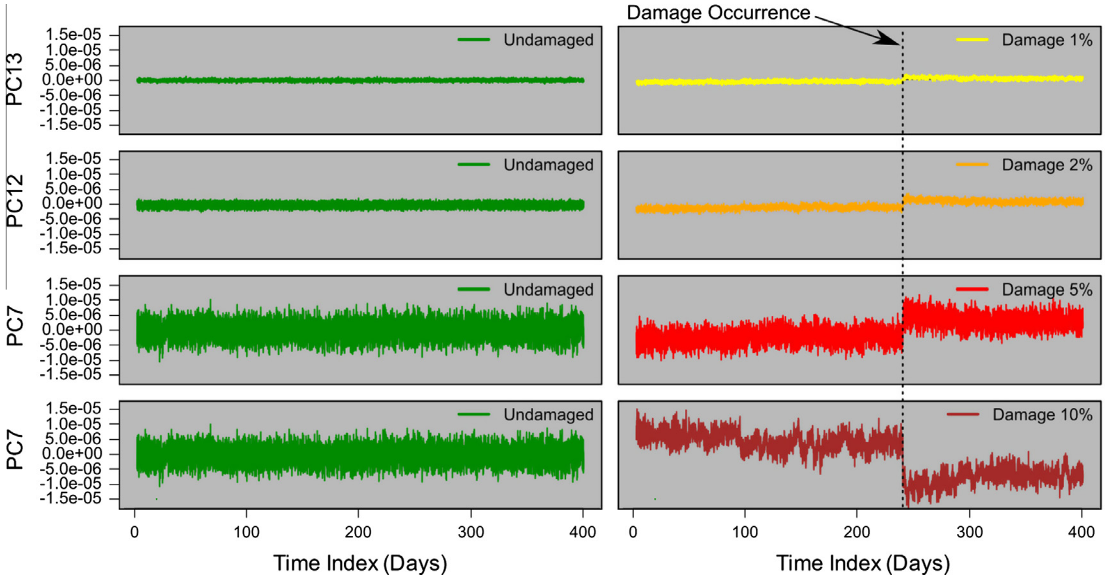

Fig. 7. Principal components obtained for the simulations of instantaneous stiffness reductions in stay cable 78, on the 1st of July 2011. Time index origin (day 0): 11th January 2011.

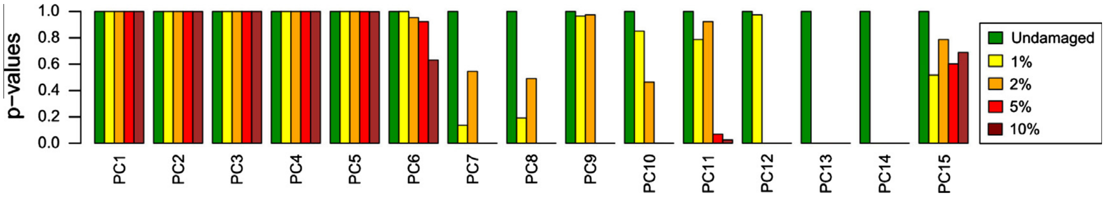

Fig. 8. Kolmogorov–Smirnov goodness-of-fit for the simulations of instantaneous stiffness reductions in stay cable 78, on the 1st of July 2011.

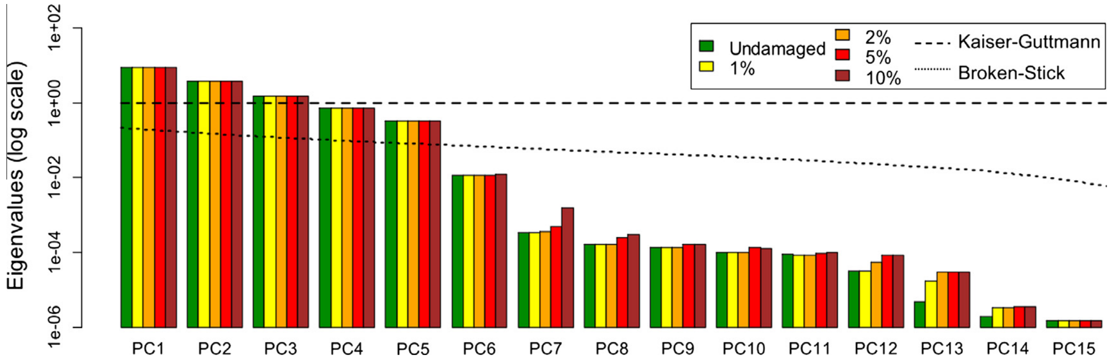

Fig. 9. PCA eigenvalues for the five simulation scenarios in stay cable 78. Bars represent the data's eigenvalues ("active energies") and lines the Kaiser–Guttmann and Broken-Stick rules' eigenvalues distributions.

unit-length stick is broken into two pieces (two principal components), the expected length for the larger piece, $b_1 = b(1, 1)$, can be obtained according to Eq. (4),

$$b_1 = b(2,1) = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{i} = \frac{1}{2}\left(\frac{1}{1}+\frac{1}{2}\right) = \frac{3}{4} \tag{5}$$

while the expected length of the smaller piece is,

$$b_2 = b(2,2) = \frac{1}{2}\sum_{i=2}^{2}\frac{1}{i} = \frac{1}{2}\left(\frac{1}{2}\right) = \frac{1}{4} \tag{6}$$

Bearing in mind that the larger piece cannot measure less than ½ and that it is equally likely to be anywhere between ½ and 1, its expected value is naturally 3/4. Therefore, the shorter piece must exhibit a length between 0 and ½ and its expected value is naturally ¼. For a set of 15 measured variables (and principle components), the Broken-Stick eigenvalue distribution, for correlation-based PCA, is presented in Fig. 9. This distribution is, by definition, monotonous; however, the logarithmic representation used in the y-axis of this figure provides a non-monotonous appearance to its line plot.

Unlike the Kaiser–Guttmann, this method identified the first five principal components as related to global effects, a result obtained by observing which eigenvalues surpass the Broken-Stick distribution in Fig. 9. This result is in agreement with the baseline comparison performed with the Kolmogorov–Smirnov tests (Fig. 8) and suggests that this method is efficient in distinguishing principal components related to temperature action from the ones related to early damage, with local character.

3.2. Symbolic data objects and dissimilarity measures

Symbolic data can be defined as a richer, less voluminous and less specific type of information, when compared to classical data [7,15,16]. While classical data mining focuses in detecting groups or patterns in individual measurements, symbolic data deals with concepts, which must be properly defined to statistically describe the analyzed data. For instance, a week of measurements for the data set considered in this work can be described by 2520 individual measurements (168 for each of the 15 sensors) or it can consist of a single symbolic object named "week of acquired data". This object must be described by statistical quantities such as histograms or interquartile intervals providing data compression without significant loss of generality or information [14]. In the case of interquartile intervals, a week of data is reduced to only 30 values (15 intervals). This type of statistical quantity has proved to be sensitive enough in detecting structural changes [7,15,16] and is consequently used in the present paper.

The effectiveness of symbolic data analysis in SHM heavily relies on the definition of symbolic dissimilarities and distances between data objects [14,42]. These measures can supply numerical values which reflect the distance between a pair of data objects. In a common sense, the lower these values are the more similar the objects may be according to their intrinsic features. Conversely, the objects with the highest distances are the ones which evidence greater discrepancies between them. A distance measure can therefore be used to quantify similarities as well as dissimilarities in data. Dissimilarity and distance measures can take a variety of forms and some applications might require specific ones. For the present work, three distinct symbolic dissimilarity measures were considered (Appendix A): the Normalized Euclidean Ichino–Yaguchi distance [43], the Gowda–Diday dissimilarity measure [44] and the Normalized Euclidean Hausdorff distance [42]. The choice of a dissimilarity measure is an important step for any clustering method and may strongly influence the shape of the clusters. This point will be highlighted later in this section.

In the present work, symbolic data objects (Fig. 10) and distances (Fig. 11) are obtained from multivariate sets of principal components, to combine them into univariate types of information. As observed in Fig. 7, these multivariate sets contain damage-related information, composed by the shifts in the time-series of principal components. However, depending of the magnitude of the damage occurrences, these shifts may be observed in principal components with different variation magnitudes (Fig. 7). When these components are combined, using the theoretical background

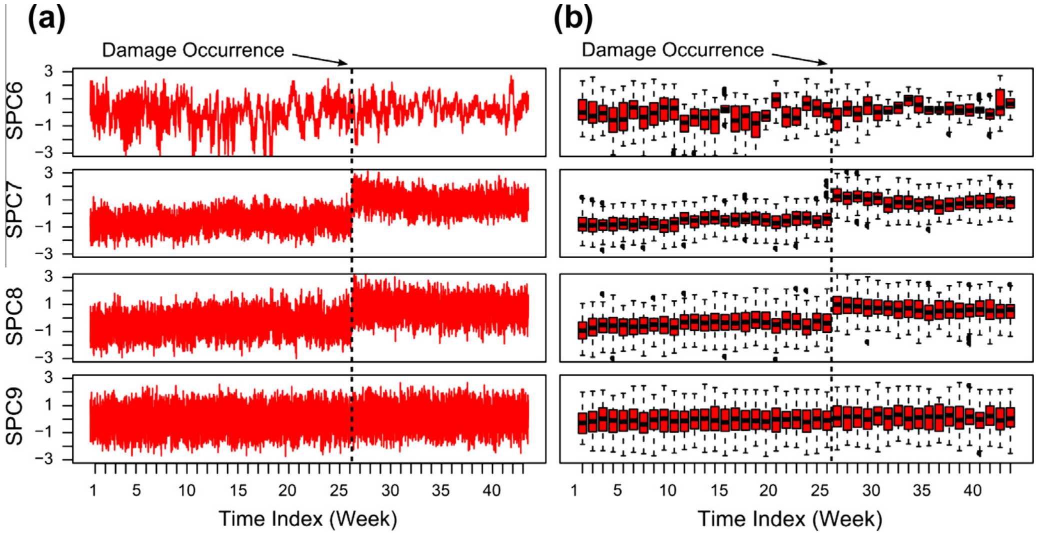

Fig. 10. Symbolic objects of standardized principal components for 5% of instantaneous stiffness reduction in cable 78. Values presented in the *y* axis have no units and stand for standard deviations: (a) classical data time series and (b) symbolic data series represented by the colored regions of box-and-whiskers plots.

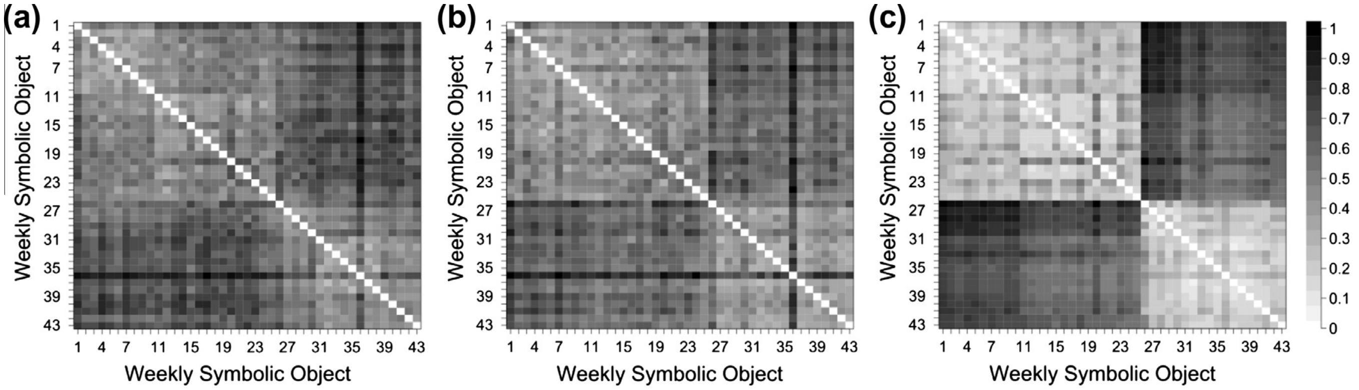

Fig. 11. Cluster dissimilarities (unit-scale) for instantaneous 5% of stiffness reduction in cable 78: (a) Gowda–Diday dissimilarity measure, (b) Normalized Euclidean Hausdorff distance, and (c) Normalized Euclidean Ichino–Yaguchi distance.

presented in Appendix A, the shifts produced by damage with smaller magnitude (top-right plots of Fig. 7), become less noticed than the ones produced by more significant damaged (bottom-right plots of Fig. 7). Hence, a simple and efficient procedure that highly increases the sensitivity to early damage consists in standardizing the principal component data set prior to the definition of symbolic objects and distances. This standardization grants similar importance to all principal components, regardless of the magnitude of damage that they highlight, and is performed as defined in Eq. (7), where $\mu(PC_i)$, $\sigma(PC_i)$ are the average and standard deviation of PC_i.

$$SPC_i = \frac{PC_i - \mu(PC_i)}{\sigma(PC_i)} \tag{7}$$

Considering the simulation of a 5% instantaneous stiffness reduction in stay cable 78 (Fig. 7), symbolic objects defined as “weeks of standardized principal components” were obtained and described by ten interquartile intervals (corresponding to the smaller ten principal components, obtained according to the Broken-Stick rule). Four series of these statistical quantities are presented by the colored regions of the box-and-whiskers plots shown in Fig. 10b. By comparison with Fig. 10a, it can easily be observed that, even though significant data compression took place, damage-related shifts are present in both types of data and appear to be outlined by the symbolic data. This fact seems to be related to the stability and generalization capacity of interquartile values in representing the data’s structure.

From the 43 symbolic objects, described by ten interquartile objects each, symbolic dissimilarity matrices have been obtained and are presented in Fig. 11. These matrices contain the pair-wise dissimilarities between all symbolic objects and constitute the input for cluster analysis. As it can be observed in Fig. 11, variations related to small magnitude damage are clearly highlighted in this type of information, where two distinct groups of data can be identified, regardless of the calculated dissimilarity. However, the Ichino–Yaguchi dissimilarity (Fig. 11c) produces a more sensitive outline of the effect of dead load redistribution than the other two (Fig. 11a and b), leading to a better sensitivity, of the proposed strategy, to early damage.

3.3. Cluster analysis

Clustering methods consist in unsupervised multivariate statistical algorithms which aim at classifying objects as members of different subsets (or clusters). Unlike supervised algorithms such as decision trees or neural networks, clustering methods do not require previous information about the objects' memberships, which are obtained according to the data's intrinsic characteristics, or dissimilarities.

The aim of a clustering method can be defined as the division of a data set into groups, which must be as compact and separated as possible. To fulfill this objective, allocation rules must be defined so that pair-wise dissimilarities between objects assigned to the same cluster tend to be smaller than those allocated in different clusters [7]. Let us consider a given partition containing K clusters, $P_k = \{C_1, \ldots, C_k\}$. The within-cluster dissimilarity $W(P_k)$ can be defined as [45]:

$$W(P_k) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(j)=k} d_{ij} \tag{8}$$

where $C(i)$ is an allocation rule which assigns element i to cluster k, based on the dissimilarity measure d_{ij}. The total variation of data can be defined as in Eq. (9), where N is the total number of objects considered in the cluster analysis. Finally, $B(P_k)$ can be simply obtained by subtracting the other two defined distances, $B(P_k) = T - W(P_k)$.

$$T = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij} \tag{9}$$

Each cluster can be described by a prototype, which generally consists of an object of the same type as the ones being clustered. The location of each prototype is obtained, in the present work, by computing the centroid of the clusters' members (Appendix B).

Several families of cluster methods can be found in the literature [45,46], however the most used are the combinatorial and hierarchical methods. While the first is iterative in nature and requires the input of an initial set of clusters' prototypes (and their centroids), the hierarchical methods provide a merging (or separation) hierarchy so that all partitions are defined, regardless of their number of elements [45]. Hierarchical methods can be classified as divisive (top-down) or agglomerative (bottom-up). Agglomerative strategies start by considering single-object clusters and, at each level, merge a selected pair of clusters into a new, single, cluster. This merging produces a new level in the hierarchy (which contains one less cluster). Divisive methods start by considering a single cluster containing all objects and, at each level, split one of the existing clusters into two new clusters. Both strategies generate hierarchies with $N - 1$ levels, where N is the number of data objects. This hierarchy can be displayed in a dendrogram plot, which is generically schematized in Fig. 12, and considered one of the main advantages of hierarchical methods [45] since it allows for a clear visualization of the structure of high dimensional data, in a single and unambiguous plot.

In the present work, agglomerative clustering is used since it was reported as computationally simplest and efficient in detecting structural changes, in previous SHM works [7,34]. Under this approach, the definition of which clusters should be merged, at each level, is based on merging rules [45,46]. The two simplest are the "single link" and "complete link". While the first states that the two clusters containing the closest objects should be merged, the latter chooses the ones with the farthest objects [46]. As a consequence, the first tends to find elongated clusters while the latter is more appropriate for finding more compact clusters [46]. The most widely used merging rules are the "average link" and the "ward", or "minimum variance" rule. The "average link" states that the pair of clusters to be merged, at each level, is the one exhibiting smaller average (element-wise) distance, thus leading to round-shaped clusters [46]. The "ward" rule defines that, at each level, the pair of merged clusters must generate a new partition with the smallest variance possible [46]. This rule was chosen for application in the present work since it does not favor any particular cluster-shape and due to the fact that it has already been reported as efficient in detecting structural changes [7,34].

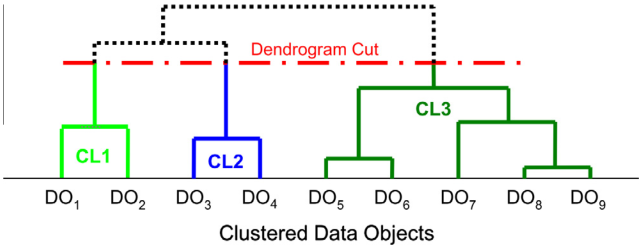

Fig. 12. Schematic representation of dendrogram plot.

From the defined agglomerative hierarchy, cluster partitions containing any number of clusters (from 1 to N) can be obtained by cutting the dendrogram plots (horizontally) between two hierarchy levels. In this type of plot, clusters are represented by vertical lines and hierarchy levels by horizontal lines (see Fig. 12). The number of clusters resulting from a dendrogram cut is equal to the number of the vertical lines intercepting the horizontal cutting line (Fig. 12). To assess which data objects belong to each of the defined clusters, one needs to observe the sub-dendrograms, located below the cutting line. From the generic example presented in Fig. 12, it can be readily observed that the cutting line intercepts three vertical lines of the dendrogram, thus generating the three clusters represented by solid lines. From the three corresponding sub-dendrograms, it can be observed that: data objects 1 and 2 belong to cluster one, data objects 3 and 4 belong to cluster two and data objects 5–9 are assigned to cluster three.

In the present section, the application of cluster analysis for early damage detection is highlighted using data from: (i) an undamaged simulated scenario and, (ii) 5% instantaneous stiffness reduction in stay cable 78. The Ichino–Yaguchi dissimilarity measure, which has exhibited greater sensitivity to data changes generated by early damage (Fig. 11), was used as input to hierarchical agglomerative clustering. The dendrogram plots are presented for both simulated scenarios, in Fig. 13a and b, along with six dendrogram cuts, which generate partitions comprising two, three and five clusters. These partitions are presented in Fig. 13c, d, Fig. 13e, f and Fig. 13g, h (for two, three and five clusters, respectively), using interquartile interval time-series of two standardized principal components (SPC6 and SCP7) and the clusters are defined by the different colors.

From the interval time-series representing the undamaged-related data (Fig. 13c, e, and g), it can be observed that objects belonging to different clusters appear to be randomly located in time. This fact suggests that no structural changes occurred during the period analyzed. Conversely, the two clusters belonging to the partition presented in Fig. 13d are compact in time and divided by the date in which damage was simulated. This result suggests that a change in the intrinsic structure of the analyzed data was observed and suggests the instant of this occurrence: the time instant that serves as boundary between the two clusters. The same conclusions can also be made for the partitions comprising three and five clusters (Fig. 13f and h, respectively), where no clusters comprise objects acquired in both monitoring periods (before and after the damage occurrence).

It can be, therefore, concluded that the constitution of cluster partitions is indicative of damage; however, its analysis may not be conclusive when no knowledge about the structural condition exists. Hence, a statistical sensitive feature of easier analysis and able to highlight damage occurrences, is required. A general comparison of cluster partitions obtained for the damaged (Fig. 13d, f, and h) and undamaged scenarios (Fig. 13c, e, and g),

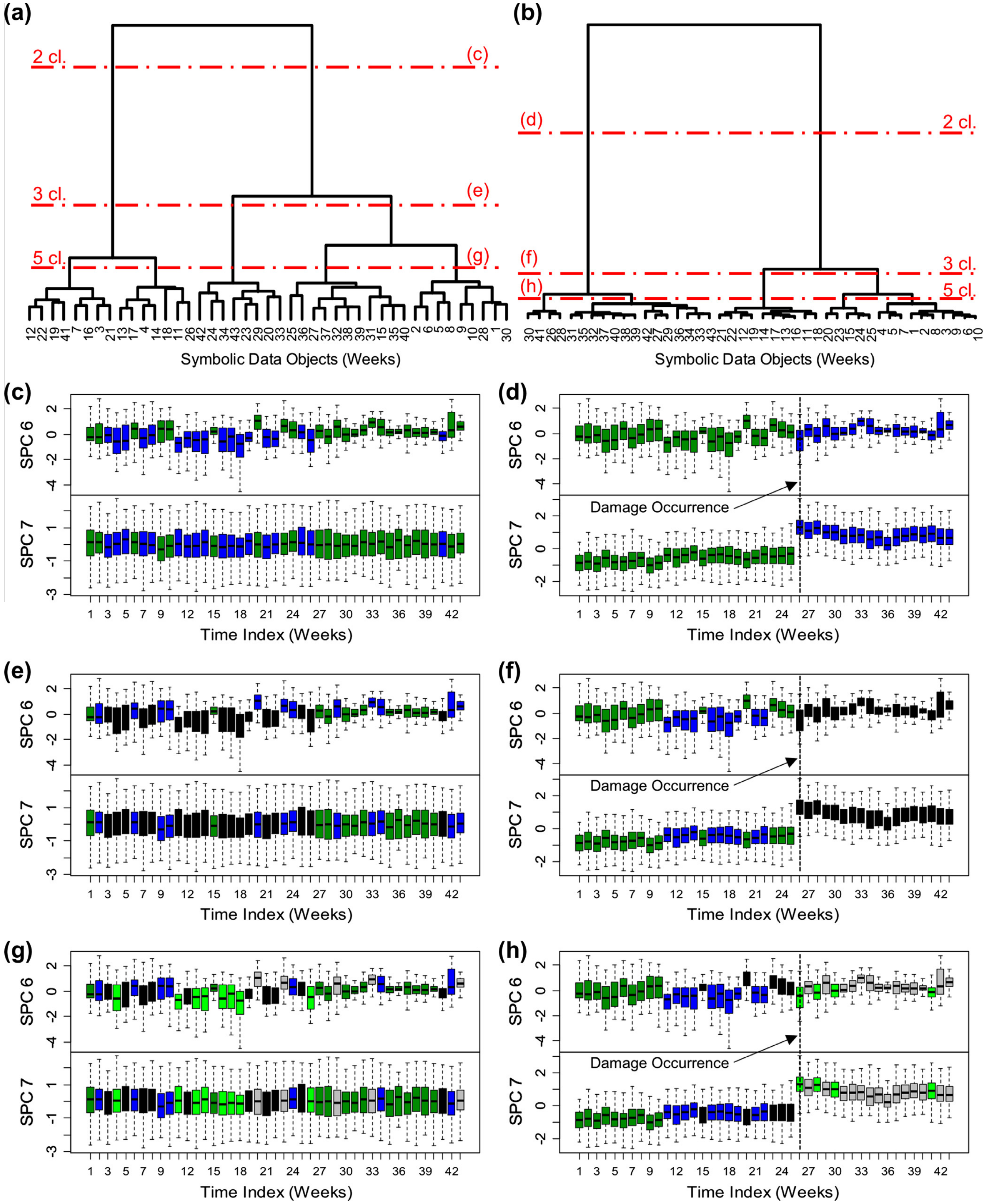

Fig. 13. Cluster analyses results: (a) dendrogram (undamaged), (b) dendrogram (damaged – 5% instantaneous stiffness reduction in stay cable 78), (c) interval-series (undamaged – 2 cl.), (d) interval-series (damaged – 2 cl.), (e) interval-series (undamaged – 3 cl.), (f) interval-series (damaged – 3 cl.), (g) interval-series (undamaged – 5 cl.), and (h) interval-series (damaged – 5 cl.).

puts in evidence that the damage occurrence results in important changes to the average distance between clusters, $B(P_k)$. This change is generated by the principle component's drifts, presented in Fig. 7 and also observed in Fig. 13 (mainly for SPC7). The sensitivity of this feature can also be observed in the dendrogram plots (Fig. 13a and b), where the average distance between clusters,

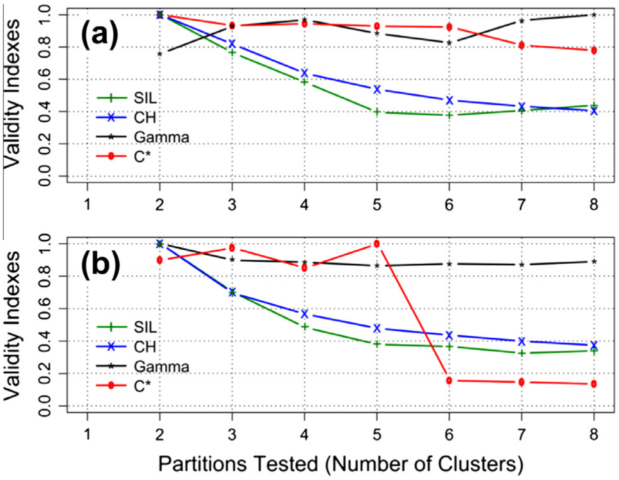

Fig. 14. Cluster validity indexes: (a) undamaged scenario and (b) simulation of 5% instantaneous stiffness reduction in stay cable 78.

$B(P_k)$, is inversely related to the total length of its vertical lines. These remarks suggest that this distance is a single-valued statistical feature capable of representing the structure of multivariate data and of pointing out changes generated by early damage, with small magnitude.

Hierarchical clustering algorithms are able to define a hierarchy of $N - 1$ levels and N partitions. However, some of these partitions generate high values of within-cluster distance (Eq. (8)) and, thus, solutions which are far from being truthful [46]. To assess which partition, among the ones obtained by cutting the hierarchy, is the most suitable for representing the data's structures, a quantitative evaluation known as cluster validity is usually performed. It consists in computing validity indexes for k pre-chosen dendrogram cuts, which result in cluster partitions with different number of k clusters [46]. Structure and dimensionality of data can influence the outcome of this task and, thus, the choice of the most appropriate validity index, which can identify the most truthful partition by its maximum or minimum value. In the present work, four validity indexes were tested (Appendix C), three which were reported as exhibiting better performance in the reference study [47]: the Calinski and Harabaz (CH) index, the C^* index and the Gamma index; and a more recent named Global Silhouette (SIL) index [46]. The optimal partitions are provided, for the CH, Gamma and SIL indexes, by their maximum values. Conversely, the C^* index identifies the optimal partition with its minimum value (Appendix C).

For the two time-history simulation scenarios used to plot Fig. 13, the four studied validity indexes were calculated. These are presented, in Fig. 14, considering cluster partitions with two to eight clusters. Each of the index values, presented in this figure, are divided by their maximum so that all plotted values do not surpass 1.0, hence easing comparison between indexes.

By observing Fig. 14a and b, it can be observed that the CH (index with best performance in [47]) and the SIL indexes exhibit great correlation and evidence a more stable behavior, regardless of the presence of damage. These two indexes reveal two as the optimal cluster partition, for both data sets considered. Conversely, the C^* and Gamma indexes vary significantly. While the first exhibits great changes but suggests the same optimal cluster partition for both simulated scenarios, the latter changes its optimal partition from eight in an undamaged state, to two under a 5% stiffness reduction. Hence, it can be concluded that the SIL and CH indexes are more efficient in understanding the data's structure, and that the real number of clusters, present in the analyzed data set, is two.

3.4. Real time simulation

To study the effectiveness of the proposed statistical strategy for real time SHM applications, the five stiffness reductions (0%, 1%, 2%, 5% and 10%) occurring instantaneously in stay cable 78, on the 1st of July, were considered.

For each of these simulated scenarios, 43 analyses corresponding to the 43 weeks of measured data were performed to simulate the functioning of an on-line SHM system, which collects data weekly from the *in situ* deployed hardware. At each data collection, the time-series used as input comprise data from instant 0 (11th January, 2011) to the last instant of collected data, resulting in an increasing size of the analyzed data set.

The output obtained at each week consists in the average distance between clusters, $B(P_k)$, which was observed to be a single-valued early damage sensitive feature, in the previous subsection. Fig. 15 presents its values for each of the five simulated scenarios, during the period of SHM on the Guadiana Bridge (11th January 2011 to the 1st of February 2012). From this figure, the great stability of the chosen sensitive feature can be observed in the series corresponding to the undamaged scenario values, where no increasing trend and small variability are exhibited. For the damage scenarios analyzed in real time, very significant increases of the features' values are observed in the series presented in Fig. 15, even for a stiffness reduction as small as 1%, thus allowing for clear early damage detections under the amount of noise measured *in situ*.

4. Conclusions

The present paper describes a novel data-driven strategy to detect early damage under environmental effects, based in static monitoring and in multivariate statistical methods. The developed strategy consists in fusing sets of measurements developed in such a way that it is of computational simplicity, allows a real time implementation, and consists in the combination of: (i) PCA, (ii)

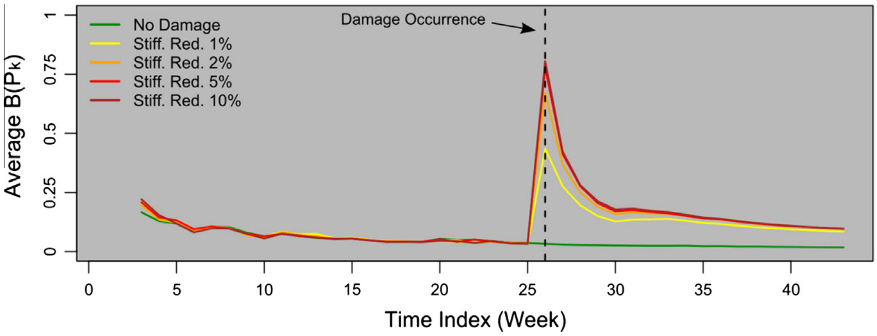

Fig. 15. Average distance between clusters, $B(P_k)$, for simulations of instantaneous damage occurrences on the 1st of July 2011.

the Broken-Stick, (iii) symbolic data, (iv) interquartile based dissimilarity measures and (v) cluster analysis.

The proposed strategy led to detections of 1% structural stiffness reductions in a stay cable. These results, obtained using a static SHM system composed of few inexpensive sensors, clearly evidence the great efficiency of the proposed strategy.

From the analyses performed, with respect to environmental normalization, it was observed that combining PCA with the Broken-Stick method provides an efficient distinction of temperature and damage-related effects. Great sensitivity to early damage was achieved: (i) by computing the Ichino–Yaguchi dissimilarity measure from symbolic data objects described by interquartile intervals, and (ii) by standardizing the principal components prior to the definition of the symbolic objects, leading to a normalization of the components' eigenvalues or "active energies". Strategies including the Gowda–Diday and Hausdorff dissimilarity measures and the use of non-standardized principal components lead to a less efficient damage detection strategy.

Four cluster validity indexes were tested to automatically and objectively obtain the data's optimal cluster partitions. All of them revealed sensitivity to early damage, however two of them, the Calinski and Harabaz and the Global Silhouette indexes, revealed greater stability and were, therefore, considered on the proposed strategy.

The efficiency of the proposed strategy was checked, in the present paper, by performing a real time procedure simulation. From the results, it was concluded that the average distance between clusters is an effective early damage sensitive feature which, in spite of being single-valued, reflects changes from all sensors installed throughout an entire structure. This simulation showed that all early damage scenarios simulated (1%, 2%, 5% and 10%) can be clearly and unambiguously detected and distinguished from the undamaged scenario.

Appendix A. Interquartile based dissimilarity measures

Let us consider two symbolic objects T_i and T_j, obtained from a data set of $s = 1, \ldots, N$ symbolic objects, which are described by their r interquartile intervals, respectively $\left(T^{(r)}_{i,inf};\ T^{(r)}_{i,sup}\right)$ and $\left(T^{(r)}_{j,inf};\ T^{(r)}_{j,sup}\right)$. The index $r = 1, \ldots, p$ stands for each of the p variables used to define the dissimilarity measure.

The Gowda–Diday dissimilarity measure, $d_{ij} = d(T_i, T_j)$, defined between the pair of objects T_i and T_j, is given by:

$$\begin{aligned}
d_{ij} &= \sum_{r=1}^{p} \varphi_r(T_i, T_j) \\
\varphi_r(T_i, T_j) &= \frac{\left|\left|T^{(r)}_{i,sup} - T^{(r)}_{i,inf}\right| - \left|T^{(r)}_{j,sup} - T^{(r)}_{j,inf}\right|\right|}{k_r} \\
&+ \frac{\left(\left|T^{(r)}_{i,sup} - T^{(r)}_{i,inf}\right| - \left|T^{(r)}_{j,sup} - T^{(r)}_{j,inf}\right| - 2I_r\right)}{k_r} + \frac{\left|T^{(r)}_{i,inf} - T^{(r)}_{j,inf}\right|}{|Y_r|} \\
k_r &= \left|\max\left(T^{(r)}_{i,sup}, T^{(r)}_{j,sup}\right) - \min\left(T^{(r)}_{i,inf}, T^{(r)}_{j,inf}\right)\right| \\
I_r &= \left|\max\left(T^{(r)}_{i,inf}, T^{(r)}_{j,inf}\right) - \min\left(T^{(r)}_{i,sup}, T^{(r)}_{j,sup}\right)\right| \\
|Y_r| &= \left|\max_s\left(T^{(r)}_{sup}\right) - \min_s\left(T^{(r)}_{inf}\right)\right|
\end{aligned} \tag{A.1}$$

The Normalized Euclidean Ichino–Yaguchi distance measure $d_{ij} = d(T_i, T_j)$ can be calculated as follows:

$$\begin{aligned}
d_{ij} &= \left(\frac{1}{p}\sum_{r=1}^{p}\frac{1}{|Y_r|}\left[\varphi_r(T_i, T_j)\right]^2\right)^{1/2} \\
\varphi_r(T_i, T_j) &= \left|T^{(r)}_i \oplus T^{(r)}_j\right| - \left|T^{(r)}_i \otimes T^{(r)}_j\right| + \gamma\left(2\left|T^{(r)}_i \otimes T^{(r)}_j\right| - \left|T^{(r)}_i\right| - \left|T^{(r)}_j\right|\right)
\end{aligned} \tag{A.2}$$

where γ is a pre specified constant ranging from 0 to 0.5. The operators $\oplus$ and $\otimes$, as well as the norm $|[\ldots]|$, are defined by,

$$\begin{aligned}
T^{(r)}_i \oplus T^{(r)}_j &= \left[\min\left(T^{(r)}_{i,inf}, T^{(r)}_{j,inf}\right) \max\left(T^{(r)}_{i,sup}, T^{(r)}_{j,sup}\right)\right] \\
T^{(r)}_i \otimes T^{(r)}_j &= \left[\max\left(T^{(r)}_{i,inf}, T^{(r)}_{j,inf}\right) \min\left(T^{(r)}_{i,sup}, T^{(r)}_{j,sup}\right)\right] \\
|A| &= |[A_{inf}, A_{sup}]| = A_{sup} - A_{inf}
\end{aligned} \tag{A.3}$$

and the quantity $|Y_r|$ is defined in (A.1).

The Normalized Euclidean Hausdorff distance $d_{ij} = d(T_i, T_j)$ is given by:

$$\begin{aligned}
d_{ij} &= \left(\sum_{r=1}^{p}\left[\frac{\varphi_r(T_i, T_j)}{H_r}\right]^2\right)^{1/2} \\
H_r^2 &= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\varphi_r(T_i, T_j)\right]^2 \\
\varphi_r(T_i, T_j) &= \max\left(\left|T^{(r)}_{i,inf} - T^{(r)}_{j,inf}\right|, \left|T^{(r)}_{i,sup} - T^{(r)}_{j,sup}\right|\right)
\end{aligned} \tag{A.4}$$

Appendix B. Cluster centroid

Let us consider a cluster C containing the objects $(T_i)_{1 \leqslant i \leqslant N}$, each object described by interquartile values $\left(T^{(r)}_{i,inf},\ T^{(r)}_{i,sup}\right)$ with $r = 1, \ldots, p$. The centroid of cluster C is defined by:

$$\begin{aligned}
\overline{C} &= \left(\left[\overline{T}^{(r)}_{i,inf}, \overline{T}^{(r)}_{i,sup}\right]\right)_{1 \leqslant r \leqslant p} \\
\overline{T}^{(r)}_{inf} &= \frac{1}{N}\sum_i T^{(r)}_{i,inf}; \quad \overline{T}^{(r)}_{sup} = \frac{1}{N}\sum_i T^{(r)}_{i,sup}
\end{aligned} \tag{B.1}$$

Appendix C. Cluster validity indexes

Let us consider a symbolic data set of N objects and K clustering partitions, chosen for validity purposes. Considering a partition containing t distinct clusters $P_t = (C_1, \ldots, C_t)$, let $C_k = \left(T^{(k)}_1, \ldots, T^{(k)}_{M_k}\right)$ be the kth cluster, constituted by M_k objects and $1 \leqslant M_k \leqslant N$.

The Calinski and Harabaz (CH) index is given by:

$$\mathrm{CH}(P_t) = \frac{B(P_t)}{W(P_t)} \times \frac{N - t}{t - 1}, \quad t = 2 \ldots K \tag{C.1}$$

where $B(P_t)$ is the between-cluster variation, $W(P_t)$ is the total within-cluster variation of partition P_t. The partition corresponding to the maximal CH absolute value is identified as the optimal clustering partition (i.e. the optimal number of clusters).The C^* index can be calculated as:

$$C^*(P_t) = \frac{1}{N}\sum_{k=1}^{t} M_k \frac{S^{(k)} - S^{(k)}_{\min}}{S^{(k)}_{\max} - S^{(k)}_{\min}}, \quad t = 2 \ldots K \quad C^* \in [0, 1] \tag{C.2}$$

where $S^{(k)}$ represents the sum of distances among the k objects within a cluster C_k, $S^{(k)}_{\min}$ is the sum of the k smallest distances among all objects and, conversely, $S^{(k)}_{\max}$ is the sum of the k largest distances among all objects. The optimal partition is given by the minimal absolute value of C^* index.The Gamma (Γ) index is obtained by:

$$\Gamma(P_t) = \frac{\Gamma_+(P_t) - \Gamma_-(P_t)}{\Gamma_+(P_t) + \Gamma_-(P_t)}, \quad t = 2, \ldots, K \tag{C.3}$$

where $\Gamma_+(P_t)$ represents the number of within-cluster distances which are smaller than between-cluster distances, and $\Gamma_-(P_t)$ is the number of within-cluster distances larger than between-cluster distances. The optimal partition is given for Γ maximal absolute value.

The silhouette width of the ith object in the cluster C_k is defined in the following way:

$$s_i^{(k)} = \frac{b_i^{(k)} - a_i^{(k)}}{\max\left(a_i^{(k)}, b_i^{(k)}\right)} \in [-1, 1] \tag{C.4}$$

The average distance $a_i^{(k)}$ between the ith object in the cluster C_k and the remaining j objects assigned to the same cluster is given by:

$$a_i^{(k)} = \frac{1}{M_k - 1} \sum_{\substack{j=1 \\ i \neq j}}^{M_k - 1} d_{ij}, \quad 1 \leqslant i \leqslant M_k \tag{C.5}$$

The minimum average distance $b_i^{(k)}$ between the same object i and all the objects clustered in one of the remaining clusters is given by:

$$b_i^{(k)} = \min_{\substack{r = 1, \ldots, K \\ r \neq k}} \left(\frac{1}{M_r} \sum_{j=1}^{M_r} d_{ij} \right), \quad 1 \leqslant i \leqslant M_k \tag{C.6}$$

where r is any cluster of partition P_t with a number of elements equal to M_r.The silhouette index of cluster C_k, s_k, and the global silhouette index of partition t, $SIL(P_t)$, are respectively given by:

$$\begin{aligned} s_k &= \frac{1}{M_k} \sum_{i=1}^{M_k} s_i^k \\ SIL(P_t) &= \frac{1}{K} \sum_{k=1}^{K} s_k, \quad t = 2, \ldots, K \end{aligned} \tag{C.7}$$

The higher the Silhouette, the more compact and separate are the clusters. Hence, its maximal value indicates the optimal partition.

References

[1] Hu X, Shenton HW. Damage identification based on dead load redistribution methodology. J Struct Eng 2006;132(8):1254–63.

[2] Teughels A, De Roeck G. Damage detection and parameter identification by finite element model updating. Rev Eur Génie Civ 2005;9(1):109–58.

[3] Rytter A. Vibration based inspection of civil engineering structures. Aalborg University; 1993.

[4] Doebling SW, Farrar CR, Prime MB, Shevitz DW. Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: a literature review. Distribution. Los Alamos; 1996. p. 134.

[5] Sohn H, Farrar C, Hemez FM, Shunk DD, Stinemates DW, Nadler BR. A review of structural health monitoring literature: 1996–2001. Struct Health Monit 2004:311.

[6] Yan A, Kerschen G, De Boe P, Golinval J-C. Structural damage diagnosis under varying environmental conditions—part I: a linear analysis. Mech Syst Signal Process 2005;19(4):865–80.

[7] Cury A, Crémona C, Diday E. Application of symbolic data analysis for structural modification assessment. Eng Struct 2010;32(3):762–75.

[8] Alvandi A, Crémona C. Assessment of vibration-based damage identification techniques. J Sound Vib 2006;292(1–2):179–202.

[9] Posenato D, Kripakaran P, Inaudi D, Smith IFC. Methodologies for model-free data interpretation of civil engineering structures. Comput Struct 2010;88(7–8):467–82.

[10] Figueiredo E, Figueiras J, Park G, Farrar C. Influence of the autoregressive model order on damage detection. Comput-Aid Infrastruct Eng 2011;26:225–38.

[11] Nair KK, Kiremidjian AS, Law KH. Time series-based damage detection and localization algorithm with application to the ASCE benchmark structure. J Sound Vib 2006;291:349–68.

[12] Moyo P, Brownjohn JMW. Detection of anomalous structural behaviour using wavelet analysis. Mech Syst Signal Process 2002;16:429–45.

[13] Ni YQ, Xia HW, Wong KY, Ko JM. In-service condition assessment of bridge deck using long-term monitoring data of strain response. J Bridge Eng 2012;17:876–85.

[14] Diday E, Noirhomme-Fraiture. Symbolic data analysis and the SODAS Software. Chicester: John Wiley and Sons; 2008. p. 445.

[15] Cury A, Crémona C. Assignment of structural behaviours in long-term monitoring: Application to a strengthened railway bridge. Struct Health Monit 2012;11(4):422–41.

[16] Cury A. Téchniques D'Anormalité Appliquées a la Surveillance de Santé Structurale. Université Paris-Est; 2010. p. 369.

[17] Kook C, Sohn H, Oh CK. Damage diagnosis under environmental and operational variations using unsupervised support vector machine. J Sound Vib 2009;325(1–2):224–39.

[18] Hua XG, Ni YQ, Ko JM, Asce F, Wong KY. Modeling of temperature–frequency correlation using combined principal component analysis and support vector regression technique. J Comput Civ Eng 2007;21(2):122–35.

[19] Zhou HF, Ni YQ, Ko JM. Constructing input to neural networks for modelling temperature-caused modal variability: mean temperatures, effective temperatures, and principal components of temperatures. Eng Struct 2010;32(6):1747–59.

[20] Peeters B, de Roeck G. One-year monitoring of the Z24 Bridge environmental effects versus damage events. Earthq Eng Struct Dynam 2001;30:149–71.

[21] Sohn H. Effects of environmental and operational variability on structural health monitoring. Philos Trans A Math Phys Eng Sci 2007;365(1851):539–60.

[22] Santos J, Silveira P. A SHM framework comprising real time data validation. In: Strauss A, Bergmeister K, Frangopol DM, editors. IALCCE 2012 – third international symposium on life cycle civil engineering. Vienna: IALCCE – International Association for Life-Cycle Civil Engineering; 2012.

[23] Santos J, Silveira P, Santos LO, Calado L. Monitoring of road structures – real time acquisition and control of data. In: Pinelo A, Rahja J, Roffé J-C, editors. 16th IRF world road meeting Lisbon, Lisbon; 2010.

[24] Santos J, Orcesi AD, Silveira P, Pina C. Damage detection under environmental and operational loads on large span bridges. V Congresso brasileiro de Pontes e Estruturas – Soluções Inovadores para Projeto, Execuçao e Manutençao. Rio de Janeiro: ABPE – Associação Brasileira de Pontes e Estruturas; 2012.

[25] Posenato D. Model-free data interpretation for continuous monitoring of complex structures. Commun École Polytech Féd Lausanne 2009:155.

[26] Ni YQ, Hua XG, Fan KQ, Ko JM. Correlating modal properties with temperature using long-term monitoring data and support vector machine technique. Eng Struct 2005;27(12):1762–73.

[27] Bellino A, Fasana A, Garibaldi L, Marchesiello SÃ. PCA-based detection of damage in time-varying systems. Mech Syst Signal Process 2010;24(7):2250–60.

[28] Zhou HF, Ni YQ, Ko JM. Structural damage alarming using auto-associative neural network technique: exploration of environment-tolerant capacity and setup of alarming threshold. Mech Syst Signal Process 2011;25(5):1508–26.

[29] Lee J-M, Yoo C, Choi SW, Vanrolleghem Pa, Lee I-B. Nonlinear process monitoring using kernel principal component analysis. Chem Eng Sci 2004;59(1):223–34.

[30] Mujica L, Rodellar J, Fernandez A, Guemes A. Q-statistic and T2-statistic PCA-based measures for damage assessment in structures. Struct Health Monit 2010;10(5):539–53.

[31] Silva S, Dias Júnior M, Lopes Junior V, Brennan MJ. Structural damage detection by fuzzy clustering. Mech Syst Signal Process 2008;22(7):1636–49.

[32] Sohn H, Kim SD. Reference-free damage classification based on cluster analysis. Comput-Aid Civ Infrastruct Eng 2008;23:324–38.

[33] Santos J, Orcesi AD, Silveira P, Guo W. Real time assessment of rehabilitation works under operational loads. In: Salta M, Moura R, Basheer M, Gonçalves A, editors. ICDS12-international conference durable structures: from construction to rehabilitation. Lisboa: LNEC – National Laboratory for Civil Engineering; 2012.

[34] Hua XG, Ni YQ, Chen ZQ, Ko JM. Structural damage detection of cable-stayed bridges using changes in cable forces and model updating. J Struct Eng 2009;135(9):1093–106.

[35] Hu X, Shenton HW. Damage identification based on dead load redistribution effect of measurement error. J Struct Eng 2006;132(8):1264–73.

[36] Caetano E, Cunha Á, Gattulli V, Lepidi M. Cable–deck dynamic interactions at the International Guadiana Bridge on-site measurements and finite element modelling. Struct Control Health Monit 2008;15:237–64.

[37] Jolliffe IT. Principal component analysis. 2nd ed. Aberdeen: Springer; 2002. p. 518.

[38] Hsu T-Y, Loh C-H. Damage detection accommodating nonlinear environmental effects by nonlinear principal component analysis. Struct Control Health Monit 2010;17:338–54.

[39] Massey FJ. The Kolmogorov–Smirnov test for goodness of fit. J Am Statist Assoc 1951;46(253):68–78.

[40] Jackson JE. A user's guide to principal components. Wiley-Interscience; 1991.

[41] Jackson D. Stopping rules for principal component analysis. Ecology 1993;74(8):2204–14.

[42] Billard L, Diday E. Symbolic data analysis. Merrill-Palmer quarterly. Chichester: John Wiley and Sons; 2006. p. 321.

[43] Ichino M, Yaguchi H. Generalized Minkowski metrics for mixed feature-type data analysis. IEEE Trans Syst Man Cybernet 1994;24(4):698–708.

[44] Gowda KC, Diday E. Symbolic clustering using a new dissimilarity measure. IEEE Trans Syst Man Cybernet 1991;24(6):567–78.

[45] Hastie T. In: Hastie T, Tibshirani R, Friedman J, editors. The elements of statistical learning, data mining, inference and prediction. Stanford: Springer; 2011. p. 763.

[46] Theodoridis S, Koutroumbas K. Pattern recognition. 4th ed. London: Elsevier; 2009. p. 961.

[47] Milligan G, Cooper M. An examination of procedures for determining the number of clusters in a data set. Psychometrika 1985;50(2):159–79.