Symbolic data analysis and supervised/ non supervised learning algorithms for bridge health monitoring

Christian Crémona [a], Alexandre Cury [b], André Orcesi [b], Luc Dieleman [c]

[a] Commissariat Général au Développement Durable, Direction de la Recherche et de l'Innovation (CGDD/DRI), La Défense Cedex, France
[b] Université Paris-Est, IFSTTAR, Paris 75015, France
[c] Société Nationale des Chemins de fer Français (SNCF), La Plaine St-Denis, France

ABSTRACT

In the past few years, numerous methods for damage assessment in connection with structural health monitoring were proposed in the literature. Several problems are raised for making these approaches practical for the engineer. The first concern is to determine whether a structure presents an abnormal behavior or not. Statistical inference is concerned with the implementation of algorithms that analyze the distribution of extracted features in an effort to make decisions on damage diagnosis.

Learning algorithms have extensively been applied to classification and pattern recognition problems in the past years and deserve to be used for structural health monitoring. Two approaches are nevertheless available depending on the ability to perform supervised or unsupervised learning. The first group of methods forms the family of classification methods whereas the second group is referred to clustering techniques. In addition, data acquisition campaigns of civil engineering structures can last from several minutes to years. Dealing with large amounts of data is not an easy task and suitable tools are required to correctly extract important features from them. To deal with this issue, symbolic data analysis (SDA) is introduced for managing complex, aggregated, relational, and higher-level data. SDA is then coupled with supervised and non supervised learning algorithms to form a new family of hybrid techniques. From the non supervised learning side, dynamic clouds and hierarchy-divisive method have been used. From the supervised learning side, neural networks and support vector machines have been introduced. All these techniques have been developed within the concept of symbolic data analysis in order to compress data without losing its inherent variability.

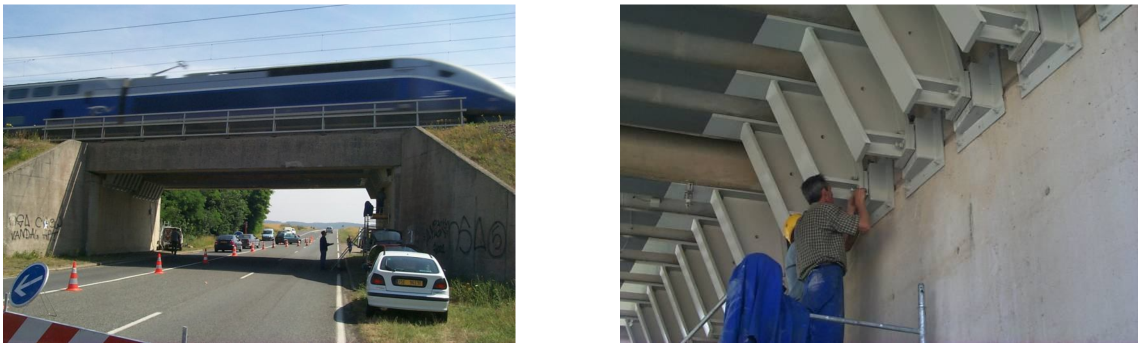
To highlight the different features of these techniques for structural health monitoring, this paper focuses attention on the monitoring of a railway bridge belonging to the high speed track between Paris and Lyon. During the month of June 2003, a strengthening procedure was carried out in this bridge. In so doing, vibration measurements were recorded under three different structural conditions: before, during and strengthening. In the following years (2004, 2005 and 2006), new tests were performed to observe how the dynamic behavior of the bridge evolved, especially for the case of frequency changes. The objective was to verify whether the strengthening procedure was still effective or not, in order terms if the new data could be still assigned to the condition "after strengthening". This paper reports the major results obtained and shows how the techniques can be applied to cluster structural behaviors and classify new data.

INTRODUCTION

Structural monitoring consists in observing, measuring and recording information. The development of high performance sensors, precision signal conditioning, Analog-to-Digital converters, optical or wireless networks, global positioning systems, etc have drastically changed the vision of structural monitoring giving to engineers a large amount of data, and consequently performance indicators. In connection with advanced software for structural analysis, significant developments can be expected regarding the detection of deterioration mechanisms. These developments open the way for a wide range of applications dedicated to efficient operation and maintenance of civil structures.

Vibration-based monitoring is such an approach since the acquisition of the structural dynamic response and its analysis are intended to give knowledge about the actual mechanical behaviour of a structure. Dynamic investigations have been widely developed over the years [1]. Various reasons justify them, but among them, novelty detection (or diagnosis of abnormal behaviour) and structural characterization are of paramount importance for operating and maintaining structures. Detecting faults structural changes in a timely manner or understanding the mechanical behaviour are critical to ensure that the resulting disruption and the economical management issues are optimized. This explains why a lot of expectations have been placed in vibration-based monitoring.

The objective of this paper is to present a practical application of these two problems. The studied bridge is an embedded steel bridge (fig.1a) located on the South-East high speed track in France at the kilometric point 075+317; it crosses the secondary road D939 between the towns of Sens and Soucy in the Yonne county. This bridge was built in the early eighties; the increase of the operating speed of high speed trains (TGV) has moved the excitation frequency of the trains close to the first natural frequency of the bridge. This risk of resonance was furthermore emphasized by the uncertainties in the mass of the ballast disposed on the bridge. To avoid this problem, the French railways SNCF set up a system of rods near the bearings tightened by torque wrench (fig.1b); this strengthening brings stiffness and increase natural frequencies. In 2003 a strengthening intervention was scheduled and led to a change in natural frequencies and mode shapes as given in Tab.1 and Fig.2 [2].

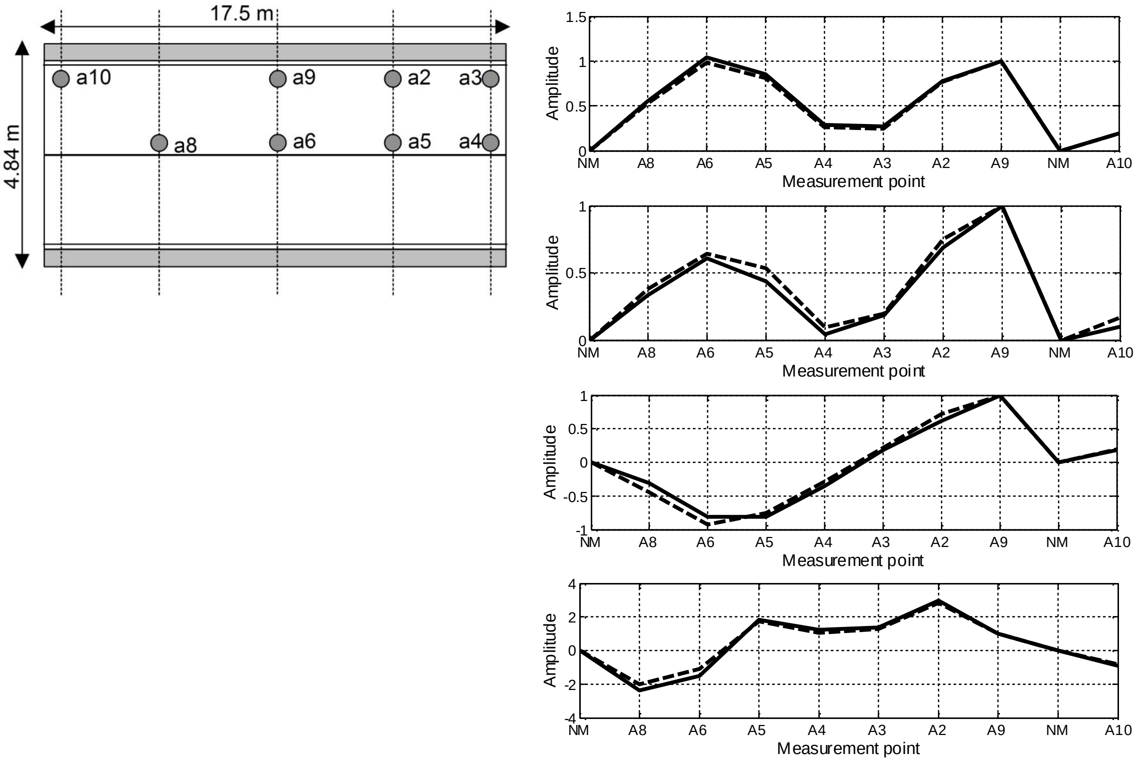


(a) (b)

Figure 1 – PK 075+317 bridge on the Paris-Lyon high speed track and strengthening system

Frequency	Before strengthening		After strengthening	
	Mean [Hz]	Standard deviation	Mean [Hz]	Standard deviation
1	5.848	0.242	6.461	0.267
2	8.507	0.322	8.592	0.415
3	13.017	0.305	13.078	0.296
4	16.850	0.502	17.142	0.507

Table 1– Natural frequencies of the PK 075+317 bridge before and after strengthening (2003)



(a) (b)

Figure 2 – Instrumentation setup (a) and evolution of the mode shapes (b)

A long term monitoring was later on decided to appraise the efficiency of this strengthening over a non continuous 2 years period (December 2004 to March 2006). The arising problem is therefore to know if the bridge dynamic behaviour remained close to the behaviour after strengthening or if it had moved to the initial structural condition, or to a completely different behaviour. This problem is an assignment problem: assuming different clusters of data, is it possible to affect any new information to one of these clusters, or eventually to identify a new type of information that cannot be related to any of these clusters?

In previous paper [3-4], the first authors introduced supervised and non supervised learning algorithms in order to build clusters and then to assign new data to them. Due to the variability in the data and the large amount of data to manipulate, Symbolic Data Analysis (SDA) was also introduced. Supervised learning refers to the case where available data from different structural behaviours are available and then new data is assigned to one of this structural condition. Unsupervised learning refers to the case where the number of structural conditions is not know and must be first determined. Once the clusters formed, new data can be assigned. The purpose of this paper is to apply the coupled SDA/learning algorithms approach to evaluate how the data from 2004-2006 differs from the "after strengthening condition" highlighting an eventual deterioration of the strengthening.

SYMBOLIC DATA ANALYSIS OVERVIEW

In general, data acquisition campaigns in civil engineering structures gather thousands of accelerations values measured by several sensors. As a consequence, analyzing all of these data (classical data) directly may usually be time-consuming or even prohibitive. In this sense, transforming this massive quantity of data into a compact but also rich descriptive type of data (symbolic data) becomes an attractive approach [5]. Let us consider, for instance, a signal X (which is part of a dynamic test) containing *n* acceleration values measured by one single sensor. There are several ways to transform classical data into symbolic data. This signal can be represented by:

- a *k*-category histogram: $X=\{a_1(n_1), a_2(n_2), a_3(n_3), \ldots, a_k(n_k)\}$;
- an interquartile interval: $X = [a_{25\%}; a_{75\%}]$;
- a min/max interval: $X = [a_{\min}; a_{\max}]$;

The same representation can be applied to modal parameters, i.e. natural frequencies and mode shapes. In other words, both of these quantities can be represented by intervals or histograms. Transforming classical data to symbolic data is carried out almost instantly which does not prohibit or make difficult the use of this methodology for a large ensemble of dynamic tests.

PRINCIPLES OF SYMBOLIC CLUSTERING METHODS

Data clustering is a common technique for statistical data analysis, which is used in many fields [6]. A clustering procedure is a non supervised learning process and can be defined as a way of classifying a number of objects into different groups. More precisely, it can be described as the partitioning of a data set into subsets (clusters), so that the data in each subset share some common properties. For an appropriate clustering, it is necessary to minimize the within-cluster variation to obtain the most homogeneous clusters as possible and, as a natural consequence, to maximize the between-cluster variation to obtain the most dissimilar clusters among each other.

To define these clusters and determine the proximity (or similarity) among the tests, it is necessary to define suitable dissimilarity measures. In a common sense, the lower these values are the more similar the objects are and thus, they are gathered in the same cluster. Conversely, the objects allocated into different clusters are the ones which have greater distances between them. Dissimilarity measures can take a variety of forms and some applications might require specific ones. In the present study, for the interval-valued data, the Hausdorff distance has been used, since it is faster to evaluate and it shows a better performance when compared to the other symbolic data distances presented in the literature. For the histogram-valued data, a standard categorical distance measure is used. More details can be found in [7].

The clustering simulations in this paper were carried out by using the software SODAS (Symbolic Official Data Analysis System), developed under the project ASSO (Analysis System of Symbolic Official Data) [8]. This paper used two well-known clustering methods: the *Hierarchy-Divisive clustering* (HD) which is based on successive top-down divisions, and the *Dynamic Clouds* (DC) method which consists in gathering the nearest tests according to a specific iterative algorithm. If a number of clusters *r* is fixed, divisive clustering is a top-down clustering process that starts with the entire dataset as one cluster and then proceeds downward through as many levels as necessary to produce the *r* optimal clusters. Dynamic clouds method consists in minimizing a general optimized criterion that measures the adequacy between the partition and the representation of the clusters, denoted *prototype*.

PRINCIPLES OF SYMBOLIC CLASSIFICATION METHODS

Any classification method uses a set of features or parameters to characterize each object. Supervised classification means that an expert both has already determined into what classes an object may be categorized and also has provided a set of sample objects with known classes. This set of known objects is called the training set because it is used by the classification programs to learn how to classify objects. There are two phases to constructing a classifier. In the training phase, the training set is used to decide how the parameters ought to be weighted and combined in order to separate the various classes of objects. In the application phase, the weights determined in the training set are applied to a set of objects that do not have known classes in order to determine what their classes are likely to be. Two supervised algorithms have been introduced in this paper: the *neural networks* and the *support vector machines*.

Supervised learning is a machine learning technique for deducing mapping functions from a training dataset consisted of input-output pairs. The goal is to predict output values of the mapping function for any valid input after having seen a number of training examples (i.e. pairs of inputs and target outputs). To achieve this, the network has to generalize from the training data to unseen situations in a "reasonable" way.

For neural networks (NN), mapping functions are obtained through an optimization scheme based on the evaluation of the mean-squared error. This scheme tries to minimize the average squared error between the network's output and the target value over all the training dataset pairs. In this paper, a feed-forward multilayer perceptron neural network is used for classifying dynamic tests. Multilayer networks use a variety of learning techniques, the most popular being back-propagation. In this case, the output values are compared with the correct answer to compute the value of some predefined error-function and the error is then fed back through the network. Using this information, the algorithm adjusts the weights of each connection to reduce the value of the error function by some small amount. After repeating this process for a sufficiently large number of training cycles, the network will usually converge to some state where the error of the calculations is negligible. At this point, it is said that the network is “trained” [4].

Support Vector Machines (SVM) are also useful techniques for data classification problems. As previously, the objective is to separate two different classes by a function which is induced from available examples (training dataset). This technique came up from statistical learning theory and is based on the structural risk minimization principle. Commonly, SVMs are used for two-groups classification problems but can easily be extended to multi-groups classification problems [4].

CLASSIFICATION OF NEW TESTS

For the classification methods (neural networks and support vector machines), once the mapping functions are determined then new data is placed as input and the mapping function returns an assignment value to one of the classes.

For the clustering methods (dynamic clouds and hierarchy-divisive method), the data is assigned to one the cluster by evaluating the dissimilarity measure of the new test to each cluster prototype in the case of dynamic clouds, the new data being assigned to the cluster presenting the smallest dissimilarity measure. For the hierarchy-divisive methods, the feature (accelerations, frequencies or mode shapes transformed into symbolic data) is compared to a cut-off value automatically generated from the algorithm.

Nevertheless, the authors [9] have shown that it is possible to improve these clustering methods by adding threshold values. Basically, for the two techniques, each cluster prototypes (this is automatically generated for the dynamic clouds but has to be built for the hierarchy-divisive method) are taken and the dissimilarity measures of each data in each cluster are calculated and a probability density function is fitted on the normalized histograms. Choosing any fractile value (5% in general), a threshold value can be calculated. For the cluster to which the new data has been assigned, the dissimilarity measure to the prototype is larger than this threshold value, then it is stated that the new data has to be rejected from the identified cluster. It therefore no longer belongs to any cluster and can be considered as a new structural behaviour.

It shows how superior to the classification techniques are the clustering methods: while the classifications methods are forced to assign to one of the initial cluster any new data, the clustering methods are able to identify new structural condition.

APPLICATION THE PK 075+317 BRIDGE MONITORING

The instrumentation of the bridge comprises 8 vertical accelerometers under the bridge deck (fig.3). The sampling frequency was fixed at 4096Hz during the strengthening phase, then to 500 Hz during the long term monitoring phase. The signal analysis due to high speed train crossings highlights a good repeatability [2], mainly because the operating conditions are mostly constant over time. In the following analyses, only acceleration measurements are used for classification. Modal parameters are extracted from the response measurements by the random decrement technique in connection with the Ibrahim Time Domain for modal parameters extraction [11]. This method requires fixing the model size (i.e. the number of modes present in the response). This disadvantage can be turned into an advantage by noting that automatically varying the model size can lead to get an evaluation of the stability of the modal parameters and therefore to assess the good quality of the modal identification by generating histograms.

In 2003, three sets of dynamic tests were performed: 15 before strengthening (represented by the letter "*A*"), 13 during strengthening (represented by the letter "*D*") and 13 after strengthening (represented by the letter "*B*"). The idea is to apply the SDA in association with supervised/non supervised learning methods outlined in the previous section to discriminate these three different stages *A, D* and *B*. In other words, the goal is to breakdown the whole set of 41 tests into three specific groups (before, during and after). Measured data (signals), natural frequencies and mode shapes of the bridge were employed.

Since the clustering methods are non supervised learning approaches, the 41 dynamic tests are classified into three groups that are not necessary perfect; each group can be composed of tests *A*, *D*, and *B*. Similarly the non supervised techniques require using a training data set (28 tests randomly chosen from 41 dynamic tests) and a validation data set (13 tests). The testing data set in the present case is composed of the 575 dynamic tests measured between 2004 and 2006. As for the non supervised learning algorithms, the classification may include some errors since all the data of the validation set could be not split in the three perfect groups *A*, *D*, and *B*. For each technique, the cluster that possesses the maximum of tests *A* is noted C^A. Similarly the clusters that have respectively the maximum of tests *B* or tests *D* are noted C^B and C^D. As the probabilities of correct classification do not reach 100%, it implies that a new test assigned to cluster C^A can only be stated close to a state "*A*" with a confidence probability that is the probability of correct classification. Tab.2 and Tab.3 provide the average probabilities of correct classification for the supervised and non supervised learning methods. These tables highlight that the clustering methods (DC, HD) are very effective in terms of classification since the average probabilities of detection are very high. For the classification techniques, the NN technique is very efficient but the SVM provides more degraded results for modal parameters (frequencies and mode shapes).

Signals		Frequencies		Mode shapes	
DC	HD	DC	HD	DC	HD
83%	73%	84%	88%	78%	82%

Table 2 – Average probabilities of correct classification for the dynamics clouds (DC) and hierarchy-divisive (HD) methods (data transformed into histograms)

Signals		Frequencies		Mode shapes	
NN	SVM	NN	SVM	NN	SVM
89%	80%	93%	56%	40%	50%

Table 3 – Average probabilities of correct classification for the neural networks (NN) and support vector machines (SVM) methods (data transformed into histograms)

In December 2004, further dynamic investigations were scheduled to analyse the dynamic properties of the bridge. In this measurement campaign 21 tests were recorded. During the year of 2005, three extra campaigns were also performed. The first one, during May-June, 107 tests were recorded. For the second one, during July-October, 254 tests were made. The third campaign, during November-December, comprised 52 tests. Finally, the last monitoring took place between the months of January-Mars 2006, with a total of 141 tests. With the non supervised learning methods, new data can be assigned to the three initial clusters or to a totally new one based on the threshold value estimated from the normalized histograms of dissimilarity measures. For the supervised learning methods, this extra-assignment is not possible and assignments can only be made to one of the three groups. Tabs.4-7 give the probabilities of assignment for each method over the different dynamic investigation periods. It come from the supervised learning methods that the structural behaviour of the bridge after two years can be mostly assigned to the state "during" highlighting a loss of the strengthening effect. The results from the non supervised learning methods are more erratic since the possibility to assign tests to a new cluster is available. Nevertheless it seems that the structural behaviour can be detected as close to state "before" or most often to a "new" state. In any case, it appears clearly that the state "after" is no longer selected; this tends to prove that the efficiency of the strengthening is no longer valid. Of course to be fully correct, these probabilities of assignment must be corrected by the average probabilities of correct detection [9].

Date	Signals				Frequencies				Mode shapes			
	C^A	C^D	C^B	C'	C^A	C^D	C^B	C'	C^A	C^D	C^B	C'
Dec/04	0	0	0	100	67	14	0	19	0	0	0	100
May-Jun/05	0	0	0	100	56	24	0	20	0	0	0	100
Jul-Oct/05	0	0	0	100	51	19	0	30	0	0	0	100
Nov-Dec/05	0	0	0	100	27	10	0	63	0	0	0	100
Jan-Mar/06	0	0	0	100	31	19	0	50	0	0	0	100

Table 4 – Probabilities of assignments obtained using the HD method

Date	Signals				Frequencies				Mode shapes			
	C^A	C^D	C^B	C'	C^A	C^D	C^B	C'	C^A	C^D	C^B	C'
Dec/04	0	0	0	100	52	10	38	0	100	0	0	0
May-Jun/05	0	0	0	100	71	6	3	21	100	0	0	0
Jul-Oct/05	0	0	0	100	56	5	31	8	100	0	0	0
Nov-Dec/05	0	0	0	100	33	33	8	27	100	0	0	0
Jan-Mar/06	0	0	0	100	38	26	7	30	100	0	0	0

Table 5 – Probabilities of assignments obtained using the DC method

Date	Signals				Frequencies				Mode shapes			
	C^A	C^D	C^B	C'	C^A	C^D	C^B	C'	C^A	C^D	C^B	C'
Dec/04	18	80	2	▨	0	100	0	▨	0	100	0	▨
May-Jun/05	5	79	16	▨	0	100	0	▨	0	100	0	▨
Jul-Oct/05	12	70	18	▨	0	100	0	▨	0	100	0	▨
Nov-Dec/05	29	56	15	▨	0	100	0	▨	0	100	0	▨
Jan-Mar/06	24	61	15	▨	0	100	0	▨	0	100	0	▨

Table 6 – Probabilities of assignments obtained using the NN method

Date	Signals				Frequencies				Mode shapes			
	C^A	C^D	C^B	C'	C^A	C^D	C^B	C'	C^A	C^D	C^B	C'
Dec/04	0	100	0	▨	0	100	0	▨	8	45	47	▨
May-Jun/05	0	100	0	▨	0	100	0	▨	28	67	15	▨
Jul-Oct/05	0	100	0	▨	0	100	0	▨	22	75	3	▨
Nov-Dec/05	0	100	0	▨	0	100	0	▨	15	81	4	▨
Jan-Mar/06	0	100	0	▨	0	100	0	▨	12	88	0	▨

Table 7 – Probabilities of assignments obtained using the SVM method

CONCLUSION

In this paper a novelty technique based on Symbolic Data Analysis and supervised/non supervised learning methods (neural networks, support vector machines, dynamic clouds and hierarchy-divisive) has been introduced to classify different structural behaviours. For this purpose, raw information (acceleration measurements) and processed information (modal data) are used for feature extraction. The approach has been applied to the monitoring of a railway bridge located in France. In June 2003, this bridge has been strengthened in order to avoid any resonance effects from high speed train crossing. Dynamic measurements before and after strengthening were performed in order to compare the structural changes induced by the repair process. The first objective of this paper was to use the clustering methods and the symbolic data analysis to discriminate these two different conditions. The results obtained showed that the methods are efficient to classify and to discriminate structural modifications either considering the vibration data or the modal parameters.

During the years of 2004 to 2006, 575 new measurements were recorded to attest the efficiency of the reinforcement works over time. Preliminary analyses considering natural frequencies or mode shapes demonstrated to be insufficient to classify new tests [9]. As second objective, symbolic data analysis and pattern recognition methods were introduced to assign these tests to the previously identified clusters, or eventually to a totally different structural behaviour (novelty detection). Results obtained showed that new tests were often classified into the cluster representing the state “before” strengthening or into a new cluster rather than into the state “after” strengthening. Overall, natural frequencies seem to be more sensitive to structural changes compared with signals and mode shapes.

REFERENCES

[1] LCPC (2009). Dynamic investigations and assessments on bridges, Technical Guide, LCPC Press.

[2] Cremona, C. (2004). Dynamic monitoring applied to the detection of structural modifications. A high speed railway bridge study, Progress in Structural Engineering and Materials, 3, 147-161.

[3] Cury, A., Cremona, C., Diday, E. (2010). Symbolic Data Analysis applied to structural damage assessment, Engineering Structures 32, 762-775.

[4] Cury, A., Cremona, C. (2010). Pattern recognition of structural behaviors based on learning algorithms and symbolic data concepts, Structural Control and Health Monitoring, DOI: 10.1002/stc.412

[5] Billard, L.; Diday, E. (2006). Symbolic Data Analysis, Wiley.

[6] Bock, H. H.; Diday, E (2001). Analysis of Symbolic Data. Springer Verlag.

[7] Malerba, D.; Esposito, F.; Gioviale, V.; Tamma V. (2001). Comparing dissimilarity measures for symbolic data analysis. ETK-NTTS 2001, Hersonissos, Crete, 1, 473-481.

[8] Diday, E.; Noirhomme-Fraiture, M. (2008). Symbolic Data Analysis and the SODAS Software, Wiley.

[9] Cury, A., Cremona, C. (2011). Assignment of structural behaviours in long term monitoring: application to a strengthened railway bridge. Structural Health Monitoring, submitted.

[11] Alvandi, A.; Cremona, C. (2006). Assessment of vibration-based damage identification techniques, Journal of Sound and Vibration, 292, 179–202.