Early damage assessment in large-scale structures by innovative statistical pattern recognition methods based on time series modeling and novelty detection

Alireza Entezami[1,2†], Hashem Shariatmadar[2], and Stefano Mariani[1]

[1]*Department of Civil and Environmental Engineering, Politecnico di Milano, Piazza L. da Vinci 32, 20133 Milano, Italy*

[2]*Department of Civil Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Azadi Square, Mashhad, Iran*

Abstract

Time series analysis and novelty detection are effective and promising methods for data-driven structural health monitoring (SHM) based on the statistical pattern recognition paradigm. However, processing substantially large volumes of vibration measurements may represent a serious limitation, especially for long-term SHM programs of large-scale civil structures. Moreover, shortcomings like the choice of an appropriate time series model in an automatic manner, the determination of optimal orders of the identified model and the classification of random high-dimensional features for damage detection, can strongly affect the performance of these approaches. This study is intended to propose statistical pattern recognition methods regarding time series modeling for feature extraction and novelty detection in feature classification in the presence of big data. These methods include an automatic model identification algorithm, an improved order determination approach and a hybrid distance-based novelty detection through a combination of Partition-based Kullback-Leibler divergence and Mahalanobis-squared distance. Experimental datasets relevant to a

† **Corresponding Author:** Alireza Entezami, Department of Civil and Environmental Engineering, Politecnico di Milano, Piazza L. da Vinci 32, 20133 Milano, Italy, Email: alireza.entezami@polimi.it,

cable-stayed bridge are considered to validate the effectiveness of the proposed methods. Results demonstrate that: the AutoRegressive-AutoRegressive with eXogenous input (AR-ARX) model turns out to be the most suitable representation for feature extraction; the orders of this model are efficiently and automatically determined; the proposed novelty detection approach is highly successful in detecting damage, even in case of large volumes of random high-dimensional features.

Keywords: Structural health monitoring; early damage detection; statistical pattern recognition; time series modeling; novelty detection; big data.

---

1. Introduction

Damage is an adverse change in a structure, which may threaten its integrity and safety. It may consist of cracks in concrete structures, loose bolts, broken welds and corrosion in steel connections, all of which may lead to stresses and displacements exceeding thresholds, inappropriate vibrations, local failure and even collapse. In order to avoid life and economic losses, structural health monitoring (SHM) is strongly needed for civil structures as it can increase the structural safety and performance, reduce the maintenance costs and prevent irreparable damage [1, 2]. Based on the assessment of the structural health or detection of any possible damage, SHM strategies can be classified into four levels: early damage detection (level 1), damage localization (level 2), damage quantification (level 3), and damage evaluation (level 4) [1]. The first level attempts to discern whether a damage exists in the structure, whereas the second and third levels are respectively intended to identify/estimate the location and severity of an already existing damage. Finally, the fourth level aims to predict the remaining service life and to provide a risk assessment for the structure [1]. As the level increases, the knowledge about damage and the complexity of the adopted methods increase as well. It thus looks essential to implement an early damage

detection method to avoid as much as possible damage occurrence, particularly in the case of large-scale structures and infrastructures [3]. In the context of SHM, this process can be carried out in short-term or long-term monitoring scenarios [4].

Damage diagnosis can be carried out by either model-driven or data-driven methods [1]. The main premise behind the first approach is the use of an elaborate and accurate analytical or numerical model of the structure. Due to discrepancies between the real structure and its digital twin, model updating turns out to be a mandatory task [5-11]. Although the model-driven approaches are often successful in damage diagnosis [12-20], some limitations and difficulties such as the requirement of a highly detailed model, the mandatory implementation of model updating, the transformation of raw vibration measurements into frequency or modal domains, make the data-driven techniques a more feasible option. Most of the data-driven SHM methods have adopted statistical pattern recognition paradigms within a framework combining feature extraction and statistical decision-making [2, 21-28]. Since such methods depend directly on measured data, some issues regarding sensor types, sensor placements and networks should be accounted for properly [10, 29]. Feature extraction is defined as a process of modeling measurements of the structural response (e.g. in terms of accelerations, displacements or strains) in the time-domain and extracting meaningful information, known as damage-sensitive features, by means of advanced signal processing techniques [30]. Time series analysis is a statistical tool for modeling raw time series data, with a high capability to analyze long sequences of data [31]. Feature extraction by time series analysis consists of fitting an appropriate time series model to the vibration data, and then extracting model coefficients and residuals as damage-sensitive features [32]. Considering stationary and linear time series, different representations can be adopted to extract the aforementioned features: Autoregressive (AR) [21, 26, 33], Autoregressive with eXogenous input (ARX) [34], Autoregressive Moving Average (ARMA) [24, 31],

Autoregressive Moving Average with eXogenous input (ARMAX) [35], and Autoregressive-Autoregressive with eXogenous input (AR-ARX) [23] ones.

Despite the reliability of such time series models for feature extraction, some challenging issues still look unsolved. A major problem is related to the analysis of substantially large volumes of time-domain responses. There is no doubt that data-driven SHM has entered the era of Big Data [31, 36, 37]; the great challenge is that the procedure of feature extraction via time-invariant linear models in such a case is time-consuming, cumbersome, and computationaly inefficient. A second important challenge is concerned with the model accuracy, which directly depends on the correct selection of the model orders. This is important because an improper choice of the order may lead to insensitivity to damage [26, 38, 39]; on the other hand, overfitting must be always prevented during order determination [40]. Another issue is the identification of a suitable model for high-dimensional vibration responses acquired under unknown ambient excitations. As a wide range of time series representations is available, the selection of the most appropriate model can be challenging when the excitation source is unknown or unmeasurable. Finally, it may not be effective and efficient to use customarily adopted graphical tools for model identification, particularly in the case of big data [41].

Statistical decision-making for feature classification is associated with the implementation of machine learning algorithms, which are generally categorized into supervised or unsupervised classes. The supervised learning class needs to handle features relevant to both undamaged and damaged conditions to train a statistical model; the unsupervised learning class uses instead only features of the undamaged state to learn the proper model. Despite the applicability of supervised learning to SHM, particularly to locate and quantify damage [42], the obvious benefit of unsupervised learning is that one does not need information related to the structure in any damaged condition. In this regard, one of the

most efficient unsupervised learning methods is novelty detection [43]. Although novelty detection methods may provide some limitations in detecting damage type and damage prognosis [44], they are widely used in SHM applications. In general, a novelty detection strategy for SHM is subdivided into a baseline (training) phase and an inspection (monitoring) phase. During the baseline stage, the damage-sensitive features extracted from the responses of the structure under known conditions are used as training datasets, to learn the baseline model. In the subsequent inspection stage, the same features (testing datasets) are extracted from the responses of the structure in the current unknown conditions, to compare them with the baseline ones for decision-making. Within this approach, any deviation of the features from those related to the baseline state is indicative of damage occurrence [21, 43].

Distance-based novelty detection methods are powerful statistical similarity measures that utilize well-defined distance metrics for feature classification [21, 23, 26, 33, 34, 43]. These methods rely upon statistical distances that measure the (dis)similarity between two sets of features (either univariate or multivariate), respectively relevant to the undamaged and damaged conditions. However, such methods do not prove effective and efficient when large volumes of random high-dimensional features are handled. In fact, the use of high-dimensional features, such as the residuals of time series models, may also have an adverse impact on the performance of novelty detection and machine learning algorithms [31, 45], and lead to serious limitations at the stage of decision-making due to complex implementation, high computational costs, and huge data storage requirements [37]. Therefore, feature classification must be dealt with by a robust and reliable novelty detection method.

To deal with all the above-mentioned issues, the focus of this article is on innovative methods based on the statistical pattern recognition paradigm, for feature extraction via time series modeling and feature classification by a distance-based novelty detection approach. An automatic, non-graphical algorithm termed Autoregressive Moving Average selection

(ARMAsel) is first presented, to choose the most suitable time-invariant model compatible with the measured vibration responses. This algorithm is able to provide an automatic model identification strategy and addresses the limitations of graphical tools, needing specific user expertise and being computationally expensive for big datasets. An improved, efficient order determination method is proposed to determine the optimal orders of the time series model to be identified. This method, which improves one authors' recently proposed technique [26], guarantees the accuracy of time series modeling, avoiding overfitting within a cost-efficient approach. A distance-based hybrid novelty detection method is finally proposed to detect damage by handling large volumes of random high-dimensional feature samples. This methodology is based on the combination of a univariate statistical measure called partition-based Kullback-Leibler divergence (PKLD) and the Mahalanobis-squared distance (MSD), which provides a multivariate statistical distance measure. This method proves efficient to deal with the high-dimensional features characteristic of the periodic monitoring of structures and early damage detection.

A series of experimental datasets related to a cable-stayed bridge, taken as a benchmark available in the literature [46], is exploited to verify the effectiveness and performance of the proposed methods. Results show that the proposed automatic model identification and improved order determination techniques are effective and efficient for response modeling and feature extraction. Additionally, the proposed PKLD-MSD method is shown to be able to detect damage from the random high-dimensional feature samples provided by the two aforementioned techniques.

The remainder of this paper is organized as follows. Section 2 briefly discusses the vibration response modeling by time series analysis. Section 3 introduces the proposed order determination approach. A description of the new novelty detection methodology is then given in Section 4. The results, in terms of feature extraction and early damage detection for

the cable-stayed bridge, are gathered in Section 5. Finally, Section 6 draws the main conclusions of this work.

2. Time series analysis

Time series analysis is a statistical method that can analyze data sequences for the purposes of model identification, parameter estimation, model validation and prediction [40]. Statistical time series methods for feature extraction in vibration-based applications can be classified as either non-parametric or parametric. The former class is based on non-parameterized representations of the structural responses, such as auto power spectral density and auto-covariance function; the advantages of this approach are its simplicity and computational efficiency. The latter class relies instead upon parameterized representations of the same responses, such as polynomial model classes, which offers a superior performance in terms of damage detection [32].

The process of feature extraction by the polynomial model classes is usually obtained with coefficient-based or residual-based algorithms [26]. The first approach relies on the estimation of the model coefficients in the baseline and inspection phases, adopting series orders that are generally set in the initial undamaged state. The second approach is instead based on the extraction of model residuals in the two SHM phases, on the basis of model orders and coefficients both set in the initial state. The main idea behind the residual-based feature extraction approach is that the model used in the healthy state is no longer able to accurately predict the response in the damaged state, due to the changes in the structural properties [26, 33]. The major benefit of the residual-based feature extraction approach is that order determination and parameter estimation do not have to be carried out again in the inspection phase [32].

In the following, the polynomial model classes adopted in this study are first discussed, and then the offered automatic model identification procedure via statistical criteria is discussed.

2.1. Polynomial model classes

Polynomial model classes provide parametric time series representations, such as the AR, ARX, ARMA, ARMAX and AR-ARX ones [47]. These models can provide input-output or output-only representations: in the first case, both the excitations (input) and the responses (output) of the structure are considered available; in the second case, only the structural responses need to be modelled, since the excitations may be either unknown or unmeasurable [32]. Generally, parametric time series representations consist of input, output and error terms, which are respectively represented by eXogenous (X), Autoregressive (AR), and Moving Average (MA) polynomials. A combination of all of them leads to the ARMAX model, which reads:

$$y(t)=\sum_{\mathrm{i}=1}^{p}\theta_{\mathrm{i}}y(t-\mathrm{i})+\sum_{\mathrm{j}=1}^{r}\varphi_{\mathrm{j}}x(t-\mathrm{j})+\sum_{\mathrm{k}=1}^{q}\psi_{\mathrm{k}}e(t-\mathrm{k})+e(t) \tag{1}$$

where $x(t)$ and $y(t)$ denote the input and output data at time t; p, r, and q respectively represent the orders of the output, input and error terms; the vectors $\mathbf{\Theta}=[\theta_1\ldots\theta_p]$, $\mathbf{\Phi}=[\varphi_1\ldots\ \varphi_r]$ and $\mathbf{\Psi}=[\psi_1\ldots\ \psi_q]$ are the model coefficients; $e(t)$ is the residual at time t, namely the difference between the measured and the predicted outputs. Eq. (1) refers to a single-input single-output system, and provides a relationship between the input loading and each sensed output of the structure; by combining similar representations for all the system outputs, a multi-output formulation is obtained. All the polynomial classes mentioned above can be obtained by neglecting some terms of the ARMAX model in Eq. (1): the ARX model is given by q=0; the ARMA model is given by r=0; finally, the AR model is given by r=q=0.

Besides these representations, one can adopt a two-stage model class named AR-ARX, which is a combination of the AR and ARX ones. Although the formulation of the AR-ARX model is similar to the ARMA one, it entirely conforms to an AR process [47]. To build the AR-ARX model, an AR representation is first fitted to the vibration response of the structure; next, the residuals are used as input data for a further ARX model. For the structural response $y(t)$, the AR model is formulated as follows:

$$y(t) = \sum_{\mathrm{i}=1}^{p} \theta_{\mathrm{i}} y(t-\mathrm{i}) + e(t) \tag{2}$$

The ARX model is then built upon $e(t)$ as input, according to:

$$y(t) = \sum_{\mathrm{i}=1}^{\bar{p}} \bar{\theta}_{\mathrm{i}} y(t-\mathrm{i}) + \sum_{\mathrm{j}=1}^{\bar{r}} \bar{\varphi}_{\mathrm{j}} e(t-\mathrm{j}) + \varepsilon(t) \tag{3}$$

where: $\bar{p}$ and $\bar{r}$ denote the orders of the output and input terms of the ARX model; $\bar{\boldsymbol{\Theta}}=[\bar{\theta}_1 \dots \bar{\theta}_{\bar{p}}]$ and $\bar{\boldsymbol{\Phi}}=[\bar{\varphi}_1 \dots \bar{\varphi}_{\bar{r}}]$ are the vectors of the relevant model coefficients; $\varepsilon(t)$ represents the model residual at time t. According to Ljung's suggestion [47], the sum of ARX orders has to be smaller than the AR order, that is $\bar{p} + \bar{r} \leq p$. Since the AR-ARX model is an enhancement of the AR one to also allow for the error term (necessary in case of ambient vibrations [23, 24]), the aforementioned constraint arises to avoid overfitting in the second stage of the AR-ARX modeling.

Even if ARMA and AR-ARX models look similar, they differ in how they model time series data: with ARMA models, two types of polynomials are adopted for the AR (output) and MA (error) terms, while one only is needed to model the output term. Accordingly, it is clear that the AR-ARX model represents an extension of the AR representation.

2.2. Automatic model identification

Even though there exists a broad range of time series representations, all of which may be applicable for the modeling based on collected data, some are not suitable for feature extraction from vibration-based measurements due to their complexity, possible overfitting, and lack of efficiency.

The selection of the most appropriate time series model can be carried out on the basis of engineering aspects, related to the acquisition and availability of input data, kind of structural system, excitation source and test devices. When both the input (excitation) and output (structural response) data are available, the problem at hand is defined as input-output; otherwise, in case of unavailability of the input data (for ambient excitation sources), the problem is classified as output-only. The ARX and ARMAX models are well suited for the input-output problem; in contrast, AR, AR-ARX and ARMA models are preferred in case of the output-only problem. It should be noted that the unmeasurable ambient excitations affect the error term in the model in such a way that a change of their amplitude leads to a change of the coefficients of the error interpolation [23, 24]. Under such circumstances, it is therefore necessary to allow for a representation such as the ARMA and AR-ARX ones, resting on a purposely defined equation for the error term.

From a statistical aspect, a common graphical tool for model identification is represented by the Box-Jenkins methodology, which is suitable to choose a polynomial model among AR, MA, and ARMA via the autocorrelation function (ACF) and the partial autocorrelation function (PACF) [40]. If the ACF tails off with an exponential decay or a damped sine wave and the PACF becomes zero after a lag, the time series conforms to an AR process. Therefore, the time series representations compatible with this process are the AR, ARX and AR-ARX ones. On the contrary, if the PACF tails off with an exponential decay or a damped sine wave and the ACF cuts off after a lag, an MA process is at hand. In this case,

the error term modeling only is required. The MA representation alone is therefore not useful for response modeling and feature extraction because, unlike the AR term, it does not depend on structural parameters. Finally, if both the ACF and the PACF tail off with an exponential decay or a damped sine wave, time series data conforms to an ARMA process [40]. In this case, the modeling of time series data needs to allocate polynomial representations for both the output and the error terms.

Despite the described simplicity of the Box-Jenkins methodology, it may become time-consuming for the selection of a proper model in case of high-dimensional time sequences and large datasets. Since the approach relies strongly on the full inspection of time series data, the decision about the kind of representation to select depends on user inference and expertise [41]. Alternatively, one can exploit numerical methods that are mainly intended to automatically choose the most appropriate time series representation via statistical criteria. The method presented in this work and belonging to this latter category, is equivalent to the Box-Jenkins methodology: the selection of one out of the AR, MA, and ARMA representations is carried out through finite sample criteria, based on the ARMAsel algorithm [48]. With this algorithm, the automatic model identification is based on the identification of AR, MA and ARMA models, and then on the selection of the overall best according to the procedure detailed in what follows.

Given a n-dimensional time series dataset (that may be constituted by acceleration measurements), the ARMAsel algorithm estimates the AR(p) models, p=1,2,…,n/2 being the order of the AR representation, and selects the best one via the Combined Information Criterion (CIC), which reads:

$$\mathrm{CIC}=\ln\left(\sigma_e^2\right)+\max\left\{\prod_{\mathrm{i}=0}^{p}\frac{1+\frac{1}{n+1-\mathrm{i}}}{1-\frac{1}{n+1-\mathrm{i}}}-1,\ 3\sum_{\mathrm{i}=0}^{p}\frac{1}{n+1-\mathrm{i}}\right\} \tag{4}$$

where σ_e^2 is the variance of the model residuals. The best AR model is the one featuring the smallest CIC value; CIC represents a trade-off between the finite sample estimator for the Kullback-Leibler information and the optimal asymptotic penalty function [48]. In Eq. (4), the penalty factor used for the second term and equal to 3, is adopted to reduce the probability of underfitting and overfitting [49]. For the estimation of MA(q) models, with order q=1,2,…,n/5, the Generalized Information Criterion (GIC) is used in the following form:

$$\mathrm{GIC}=\ln\left(\sigma_e^2\right)+\frac{3q}{n} \tag{5}$$

where the same penalty factor adopted for CIC is adopted. Finally, the estimation of ARMA(p^*,p^*-1) models, of order p^*=2,3,…,n/10, is obtained via the GIC in the following form:

$$\mathrm{GIC}=\ln\left(\sigma_e^2\right)+\frac{3\left(2p^*-1\right)}{n} \tag{6}$$

For both the MA and ARMA representations, the best model is the one that minimizes the GIC value.

The prediction error (PE) for each model of the three classes is next computed, and the model class leading to the smallest PE value is finally chosen. For the AR model, PE is given by:

$$PE\left(p\right)=\sigma_e^2\left(\prod_{\mathrm{i}=0}^{p}\frac{1+\frac{1}{n+1-\mathrm{i}}}{1-\frac{1}{n+1-\mathrm{i}}}\right) \tag{7}$$

while for the MA and ARMA models, PE reads:

$$PE(a) = \sigma_e^2 \left(\frac{1 + \frac{a}{n}}{1 - \frac{a}{n}} \right) \tag{8}$$

where a is the number of coefficients to be estimated. It is to note that, if a is smaller than $n/10$, there is no difference between the values of PE provided by Eqs. (7) and (8). To reduce the computational costs of this identification and model selection task, according to [48] it looks appropriate to upper bound the order of each model by 1000, rather than using the limits $n/2$, $n/5$ and $n/10$ mentioned before.

3. An efficient order determination method

Besides the importance of the initial data analysis and model identification, an imperative issue is the determination of sufficient and optimal orders of the polynomial models. From a statistical viewpoint, such orders should enable to provide uncorrelated (independent) residuals [40], thought of as a primary criterion to assess the accuracy of the identified model. In this section, an improvement of the iterative-only order determination approach of [26], hereafter defined as the original technique, is proposed aiming to develop an efficient model order determination. The main strength of the new method is represented by its computational efficiency in the case of so-called Big Data. Although both the original and improved techniques are based upon the residual analysis through the Ljung-Box Q-test (LBQ) [40], the newly proposed method will be shown to be faster and more effective. In other words, both methods attain the same results, but the improved approach needs shorter time intervals than the original technique.

The residual analysis via the LBQ test can be carried out by evaluating the test statistics, to be compared with a critical value (c-value) under an assigned significance level α. The statistical hypothesis test assesses the correlation among model residuals through:

$$Q_{LB} = n(n+2)\sum_{z=1}^{K}\frac{\rho_z^2}{n-z} \tag{9}$$

where n denotes, as before, the number of samples; ρ_z is the sample ACF at lag z; K represents the number of lags. The test decision can be either based on the null or an alternative hypothesis linked to a specific level α; under the null hypothesis, if Q_{LB} turns out to be smaller than the c-value the model residuals are considered uncorrelated.

The main limitation of the iterative order determination method proposed in [26] is its computational inefficiency to attain $Q_{LB}<c$-value for high-dimensional time series data. Overfitting, which occurs when the attained model contains redundant coefficients, must be avoided otherwise undesirable forecasts can be obtained, see [40]. Overfitting is thus also coped with in the proposed improved method.

Taking the above-mentioned limitations and shortcomings into consideration, the improved iterative method for the model order determination consists of non-iterative and iterative phases, as shown in Fig. 1. The non-iterative first stage of the method is intended to set the initial model orders (namely, p_0, r_0, and q_0) by one of the well-known statistical criteria, such as the Akaike Information Criterion (AIC), the corrected AIC (AICc), or the Bayesian Information Criterion (BIC). While the AIC often tends to give rise to overfitting, the AICc and BIC enhance the procedure by adding rigorous penalty terms. The AICc is also known to properly behave for small samples, whereas the BIC is better suited for large data sequences [40]. Due to the superiority of BIC over the other mentioned criteria, it is adopted in this non-iterative stage. Given an n-dimensional time series data and a model made by a

coefficients (namely $a=p+r+q$ for ARMAX, $a=p+q$ for ARMA, $a=p+r$ for ARX, and $a=p$ for AR), the BIC is given by [40]:

$$\mathrm{BIC} = -2\ln\left(L_{\max}\right) + a\ln\left(n\right) \tag{10}$$

where $L_{\max}$ represents the maximum value of the likelihood estimate of the model. The model order determination via the BIC (and similarly with other information criterion techniques) begins with computing (10) with a trial for the sample orders (e.g. 100); the initial orders are those that provide the smallest BIC values. If the model residuals obtained according with these orders (p_0, r_0, and q_0) satisfy the LBQ test (Q_{LB}<c-value), they are selected as the optimal model orders; otherwise, the iterative stage of the algorithm is started to progressively increase them by setting $p_i=p_0+i$, $r_j=r_0+j$, and $q_k=q_0+k$, where i,j,k=1,2,…, until when Q_{LB} becomes smaller than the c-value. Meanwhile, overfitting is evaluated through the R-squared (R^2) and the adjusted R-squared ($Adj\text{-}R^2$) statistics [50]. If $\mathbf{y}=[y(1)\ y(2)\ \ldots\ y(n)]$ is the response vector and $\mathbf{e}=[e(1)\ e(2)\ \ldots\ e(n)]$ is the residual vector of the time series model with the mentioned a coefficients to tune, the R-squared is given by:

$$R^2 = 1 - \frac{\sum_{t=1}^{n} e\left(t\right)^2}{\sum_{t=1}^{n}\left(y\left(t\right) - \hat{y}\right)^2} \tag{11}$$

where $\hat{y}$ denotes the mean value of y. R^2 always varies from zero to one, so that values close to one suggest a good fit to the series data. However, it does not necessarily imply that the time series model is appropriate, since an increasing model order never leads to a reduction of R^2. As under such circumstances the occurrence of overfitting is rather common, the adjusted R-squared is also considered according to:

$$Adj\text{-}R^2 = 1 - \frac{\left(\frac{1}{n-a}\right)\sum_{t=1}^{n} e(t)^2}{\left(\frac{1}{n-1}\right)\sum_{t=1}^{n}\left(y(t) - \hat{y}\right)^2} \tag{12}$$

We recall that variables a and n in the previous equations respectively denote the total number of coefficients of the selected model and the number of time series samples.

Similarly to R^2, the adjusted R-squared usually varies in the range of zero to one, but it may also provide negative values. In general, $Adj\text{-}R^2$ does not always increase by increasing the model order: if unnecessary terms are added to the model, the value of the adjusted R-squared statistics often decrease. As a result, an adjusted R-squared value close to 1 (i.e. $Adj\text{-}R^2 \approx 1$) is indicative of a good fit. Overfitting is avoided if the value of $Adj\text{-}R^2$ is positive and slightly smaller than R^2, see [50]; if instead R^2 and $Adj\text{-}R^2$ are far different or $Adj\text{-}R^2$ becomes negative, it is suggested to modify the model class and select a more parsimonious time series representation [50].

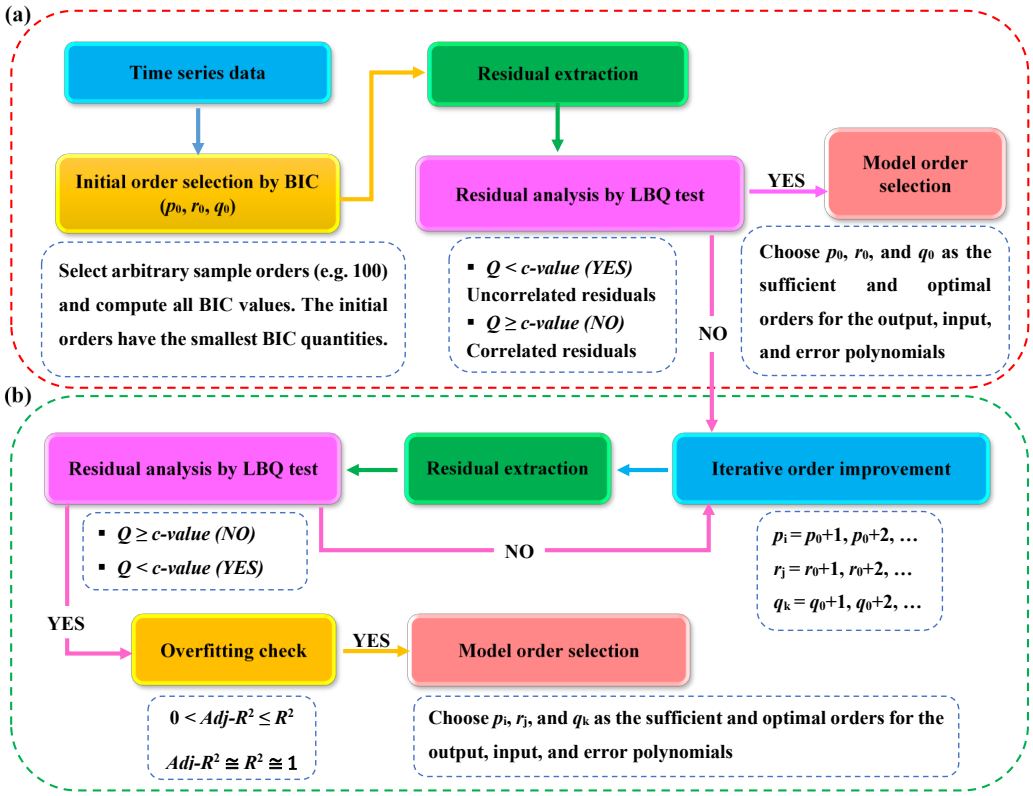


Fig. 1. Schematic representation of the improved order determination method: (a) the non-iterative algorithm, (b) the iterative algorithm

4. An innovative distance-based novelty detection methodology

The proposed distance-based novelty detection PKLD-MSD is a hybrid methodology for feature classification. As stated earlier, it is based on the combination of the univariate statistical PKLD measure and of the multivariate MSD. First, the PKLD works on two sets (vectors) of high-dimensional residual samples to provide a scalar distance quantity; therefore, it is adopted as a data dimensionality reduction. By considering all the residual datasets at all sensor locations and for all the measurements of the structural states in the baseline and inspection phases, one can obtain the training and testing sets for feature classification. Finally, these sets are fed to the MSD, in an effort to detect early damage. The whole procedure is detailed in the following.

4.1. Partition-based Kullback-Leibler divergence method

Distance methods are powerful statistical tools to measure the (dis)similarity between two sets of samples. If the samples of interest are time series, it can be difficult to introduce a robust statistical distance method for measuring the (dis)similarity of such samples, since time series are essentially high-dimensional data that lead to a major obstacle to decision-making [51]. To overcome this limitation in the process of novelty detection, the PKLD method is adopted and it gives the following advantages. First, PKLD measures the (dis)similarity of random high-dimensional time series datasets; working on two sets of samples, it provides a scalar distance measure. This value is obtained through a data-partitioning algorithm, which subdivides the random sequences into independent partitions which are next exploited in place of the original random sets to compute their (dis)similarity. Second, PKLD is capable of computing the (dis)similarity of both correlated and uncorrelated random sets. This property makes it suitable for measuring the distance between the residual samples extracted from the time series. In fact, while the extracted residuals in the baseline phase are uncorrelated, in the inspection phase the already fitted model leads to residual samples that can be either correlated or not [39]. Third, there are no limitations to the dimension of samples, which can also have different sizes.

Assume that $\mathbf{E_x}=[\varepsilon_x(1)\ \varepsilon_x(2)\ \dots\ \varepsilon_x(m)]$ and $\mathbf{E_y}=[\varepsilon_y(1)\ \varepsilon_y(2)\ \dots\ \varepsilon_y(n)]$ are m and n-dimensional random vectors (as said, with m and n either equal or different), like e.g. the ARX residuals of the AR-ARX models in two different structural states. Henceforth, $\mathbf{E_x}$ and $\mathbf{E_y}$ are called the reference and current datasets, for which PKLD has to provide the distance or divergence of $\mathbf{E_y}$ from $\mathbf{E_x}$. The partitioning process of these datasets relies on the maximum entropy approach [33]. To this aim, the samples of $\mathbf{E_y}$ are arranged in ascending order in such a way that it begins with $\varepsilon_{y_{\min}}$ and ends with $\varepsilon_{y_{\max}}$. The arranged vector $\mathbf{E_y}$ is then split into c partitions, named $\mathbf{H}_1,\mathbf{H}_2,\dots,\mathbf{H}_c$, in the following form:

$$\begin{aligned}
&\varepsilon_y\big|_{\min} \le \mathbf{H}_1 \le \varepsilon_y\big|_{\sigma} \\
&\varepsilon_y\Big|_{(\mathrm{j}-1)\sigma} < \mathbf{H}_{\mathrm{j}} \le \varepsilon_y\Big|_{(\mathrm{j}\sigma)}, \qquad \mathrm{j}=2,3,\ldots,c-1 \\
&\varepsilon_y\Big|_{\sigma(c-1)} < \mathbf{H}_c \le \varepsilon_y\big|_{\max}
\end{aligned} \tag{13}$$

Here σ denotes the number of samples in each partition of the current set, and is empirically given by $\sigma=\sqrt{n}$ [33]; in such a case, the total number of partitions is $c=n/\sigma$. It is important to mention that σ and c must be positive integers; for positive non-integer values, one next needs to round them off to the nearest integer.

The PKLD equation then reads:

$$D_{\mathrm{PKL}}\left(\mathbf{E}_{\mathbf{x}} \,\|\, \mathbf{E}_{\mathbf{y}}\right) = \sum_{\mathrm{j}=1}^{c\text{-}1} \frac{\nu_{\mathrm{j}}}{m} \cdot \left(\log\left(\frac{\nu_{\mathrm{j}}}{m}\right) - \log\left(\frac{\sigma}{n}\right)\right) + \frac{\nu_c}{m} \cdot \left(\log\left(\frac{\nu_c}{m}\right) - \log\left(2\gamma_c\right)\right) \tag{14}$$

where: v denotes the number of samples of $\mathbf{E_x}$ that fall into the range of values of each partition of $\mathbf{E_y}$; v_c is the number of samples in the last partition of $\mathbf{E_x}$.

4.2. Definition of the training and testing datasets

The main objective of PKLD is to set the multivariate training and testing datasets from large volumes of random high-dimensional features, that are the residuals of the time series model, to next fed the MSD technique. Accordingly, one initially learns a statistical model via the MSD by the training data, and then makes a decision about the current state of the structure for the early detection of damage using the testing data. Training and testing datasets are obtained by computing the distance between the residual sets relevant to the baseline and inspection phases through the PKLD: Fig. 2 shows the flowchart to obtain such training and testing datasets.

Let us assume that the structure in the baseline phase features n_L normal or undamaged states, denoted with $S_{N_1}, S_{N_2}, \ldots, S_{N_L}$ under different potential operational and/or environmental

conditions. For each state, the vibration responses are measured during n_T tests at n_S sensor locations. Furthermore, let $S_{C_1}, S_{C_2}, \dots, S_{C_U}$ denote the n_U current structural conditions in the inspection phase, assuming to handle the same number of test measurements and sensor locations. To obtain the training data, a distance matrix is computed via PKLD between each normal condition and itself and also all the other undamaged states, using the residual sets relevant to all the test measurements and sensor locations. As an example, $\mathbf{D}_{Tr}^{S_{N_1}} \in \mathbb{R}^{n_S \times n_L n_T}$ denotes the matrix for the state S_{N_1}, which is obtained by computing the distances between S_{N_1} and $S_{N_1} \dots S_{N_L}$; the same matrix can be obtained for the other undamaged states. Once the PKLD matrices for all the conditions in the baseline phase have been obtained, the training data $\mathbf{D}_{Tr} \in \mathbb{R}^{n_S \times n_{Tr}}$ are built by combining these matrices, allowing that $n_{Tr}=n_L \times n_L \times n_T$.

In order to build the testing data for each structural state of the inspection phase, the same procedure is repeated by computing the distances between the current state and all the normal conditions in the baseline period. Since each normal condition is characterized by its own model orders and coefficients, which are used to extract the residuals of the current state on the basis of the adopted residual-based feature extraction algorithm, the PKLD is computed for each residual dataset of the current state obtained from each normal condition. For example, $\mathbf{D}_{Te}^{S_{C_1}} \in \mathbb{R}^{n_S \times n_L n_L n_T}$ refers to the matrix for the current state S_{C_1}, which is obtained by computing the distances between the residual datasets of S_{C_1} and the normal conditions $S_{N_1} \dots S_{N_L}$. For all the unknown structural conditions in the inspection phase, one can obtain a distance matrix as $\mathbf{D}_{Te} \in \mathbb{R}^{n_S \times n_{Te}}$ based on the combination of all PKLD matrices of these states, where $n_{Te}=n_U \times n_L \times n_L \times n_T$.

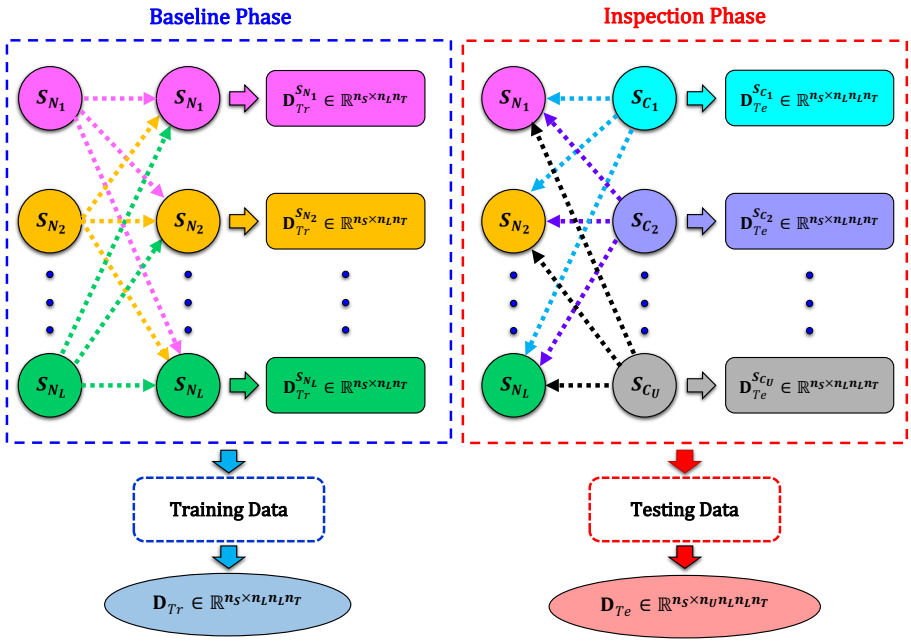


Fig. 2. Schematic representation of data handling to obtain the training and testing datasets

4.3. Mahalanobis-squared distance method

The MSD is a powerful statistical distance metric, that aims to measure the (dis)similarity between two multivariate datasets. The provided measure does not depend on the scale of samples, and performs the similarity computation on the basis of the correlation between variables. Considering the training dataset $\mathbf{D}_{Tr}$, one can learn an unsupervised learning model, which consists of the mean vector $\hat{\mathbf{d}}_{Tr} \in \mathbb{R}^{n_S}$ and covariance matrix $\mathbf{\Sigma}_{Tr} \in \mathbb{R}^{n_S \times n_S}$ of the training data. The MSD computes the distance of each column vector of the testing data $\mathbf{d}_{Te} \in \mathbb{R}^{n_S}$ from the trained model as follows:

$$D_{MS}(l) = \left(\mathbf{d}_{Te}(l) - \hat{\mathbf{d}}_{Tr}\right)^{\mathrm{T}} \mathbf{\Sigma}_{Tr}^{-1} \left(\mathbf{d}_{Te}(l) - \hat{\mathbf{d}}_{Tr}\right), \qquad l = 1,2,...,n_{Te} \tag{15}$$

where n_{Te} denotes the number of testing samples. The fundamental principle behind the MSD method can be stated as follows: if the feature vector is obtained from a damaged structure, the testing dataset $\mathbf{d}_{Te}(l)$ deviates from the mean of the normal conditions; if the feature vector

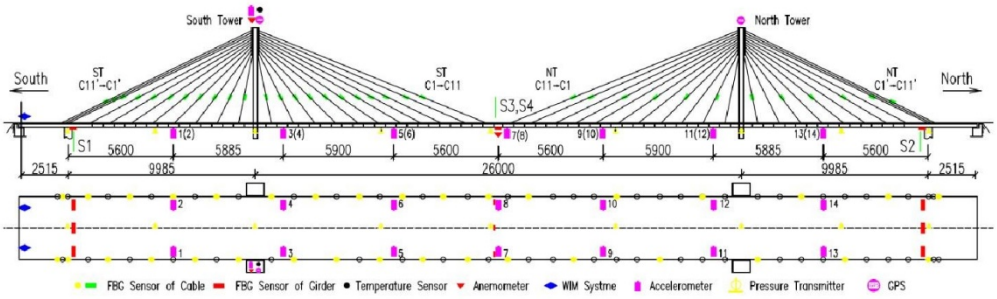
comes from an undamaged structure, it turns out to be close to the mean of the normal conditions [43].

The estimation of a threshold limit for novelty detection is a crucial task, to distinguish a damaged state from a normal condition. A common approach to estimate this limit is by computing a reference value based on the training dataset collected for the normal conditions. The methodology to estimate such threshold value relies on a scalar value from $\mathbf{D}_{Tr}$ obtained via the MSD method. On this basis, each column vector of the training data ($\mathbf{d}_{Tr} \in \mathbb{R}^{n_s}$) is handled in Eq. (14) on place of $\mathbf{d}_{Te}$. By computing all the MSD quantities, the threshold for novelty detection is defined as the 95% of the largest MSD value from the training data.

5. Application to a cable-style bridge

A series of experimental datasets for a full-scale benchmark structure is here employed to validate the accuracy and performance of the proposed methods. The structure is the cable-stayed Tianjin-Yonghe Bridge, shown in Fig. 3. This is one of the earliest cable-stayed bridges constructed in Mainland China; it consists of a main span of 260m and two side spans of 25.15m and 99.85m. The bridge is 510m long and 11m wide, and it includes a 9m wide roadway and two 1m wide sidetracks for pedestrians. The concrete towers, connected by two transverse beams, are 60.5m tall. Additional details on the bridge and on the installed sensor network can be found in [46].

(a)



(b)

Fig. 3. The Tianjin-Yonghe Bridge: (a) general view and (b) main dimensions and sensor locations. Adapted from [46]

The bridge was opened to the traffic on December 1987. After 19 years of operation, in 2005 some serious cracks were detected at the bottom of a girder segment over the mid-span; furthermore, some stay cables, particularly near the anchors, turned out to be severely corroded. With a sophisticated SHM system designed by the Center of SMC at Harbin Institute of Technology, the bridge started to be monitored in 2007 after a major rehabilitation program to replace the damaged girder segment and all the stay cables. In August 2008, new damage patterns were found in the girders during an inspection of the bridge. Based on the description of the second part of this benchmark problem, the damage practically consisted of a separation of the main girder, close to the location S1 (see Fig. 3), from the auxiliary piers and led to the loss of vertical support. This separation reduced the stiffness of the adjacent girder, with a significant effect of the external loading, particularly of the traffic load. Heavy traffic loading might therefore cause severe damage during the short-time period of

monitoring [52]. For this bridge, acceleration time histories were recorded from January to August 2008 by 14 single-axis accelerometers, deployed as shown in Fig. 3(b), under operational (traffic loads) and environmental (temperature and wind) variable conditions. The measured data include 24 datasets of 1 hour, with a sampling frequency of 100Hz; each dataset thus consists of 360,000 samples.

The acceleration datasets from 13 sensors (n_S=13), measured with the accelerometers 1-9 and 11-14 of Fig. 3 on January 1, January 17, February 3, March 19, March 30, April 19, May 5, May 18, and July 31 are here handled to assess the accuracy and reliability of the proposed approach. The vibration datasets recorded on May 31, June 7, and June 16 are not used due to the poor excitation conditions, see [53]. Moreover, the vibration data gathered at sensor 10 are not accounted for, owing to meaningless measurements already highlighted elsewhere. As described in [46, 53], the first 8 sets/days (January 1 - May 18) of measurements are representative of the undamaged conditions, while the last set/day of measurements (July 31) refers to the damaged state of the structure.

Table 1. Undamaged and damaged states for the Tianjin-Yonghe Bridge

Day no.	Date	State	Description
1	January 1	Healthy	Baseline phase
2	January 17	Healthy	
3	February 3	Healthy	
4	March 19	Healthy	
5	March 30	Healthy	
6	April 19	Healthy	
7	May 5	Healthy	
8	May 18	Healthy	Inspection phase
9	July 31	Damaged	

Table 1 lists the undamaged and damaged conditions used in this study. For the procedure of feature classification, the first seven days are considered in the baseline phase; hence, n_L=7. Days eight and nine are instead considered belonging to the inspection phase. In accordance with a common procedure in machine learning, the training data consist of 87.5% of the normal conditions (Days 1-7) while the testing data consist of 12.5% of the remaining undamaged state (Day 8) and of the information on the damaged state (Day 9). The acceleration responses relevant to all test measurements of Day 8 are then exploited as validation data. The total response datasets thus consist of 1,010,880,000 data samples, coming from 3,600,000 acceleration time histories measured with the 13 sensors over 24 hours. Such extremely large numbers are given to highlight that the long-term SHM program for the Tianjin-Yonghe Bridge faces the problem of Big Data.

5.1. Automatic model identification

As the ambient excitations on the bridge were unmeasurable and unknown, the input data (i.e. the excitation loads) are not available. As explained in Section 2.1, the ARMA or AR-ARX representations can then be adopted to model the vibration responses under the ambient excitation sources. From a statistical perspective, the main difference between these two models lies in the fact that AR-ARX is consistent with the AR process, whereas ARMA conforms to both the AR and MA processes [40, 47]. It should be also noted that, according to Eqs. (2) and (3), the eXogenous term of the AR-ARX model is represented by the residuals of the AR representation obtained from the first stage. Hence, the AR-ARX representation does not conflict with the unknown information concerning the real input data.

Using the ARMAsel-based automatic model identification approach, the best AR, MA, and ARMA models are obtained with the statistical criteria detailed in Eqs. (4)-(6). The PE values relevant to these three representations are next computed according to Eqs. (7) and

(8), to choose the most appropriate model. Fig. 4 and Fig. 5 show exemplary results of automatic model identification at some sensor locations, on Days 1 and 9. It can be clearly observed that the PE values for the best AR representation are smaller than those related to the MA and ARMA models. This outcome confirms that the acceleration responses conform to an AR process. As the AR-ARX model is an extension of the AR representation, to account for the AR residuals with a further ARX model, it can be concluded that it is more suitable than an ARMA model. The same conclusion can be attained with all the other measurements obtained with the sensors during the baseline phase.

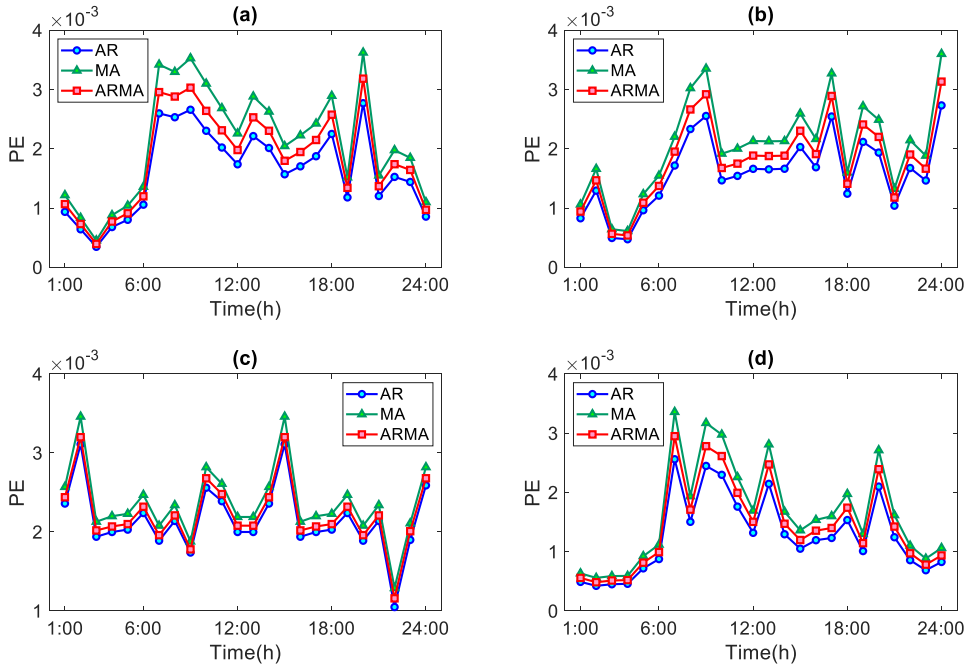


Fig. 4. Automatic model identification by the ARMAsel algorithm on Day 1: (a) Sensor 2, (b) Sensor 6, (c) Sensor 8, (d) Sensor 12

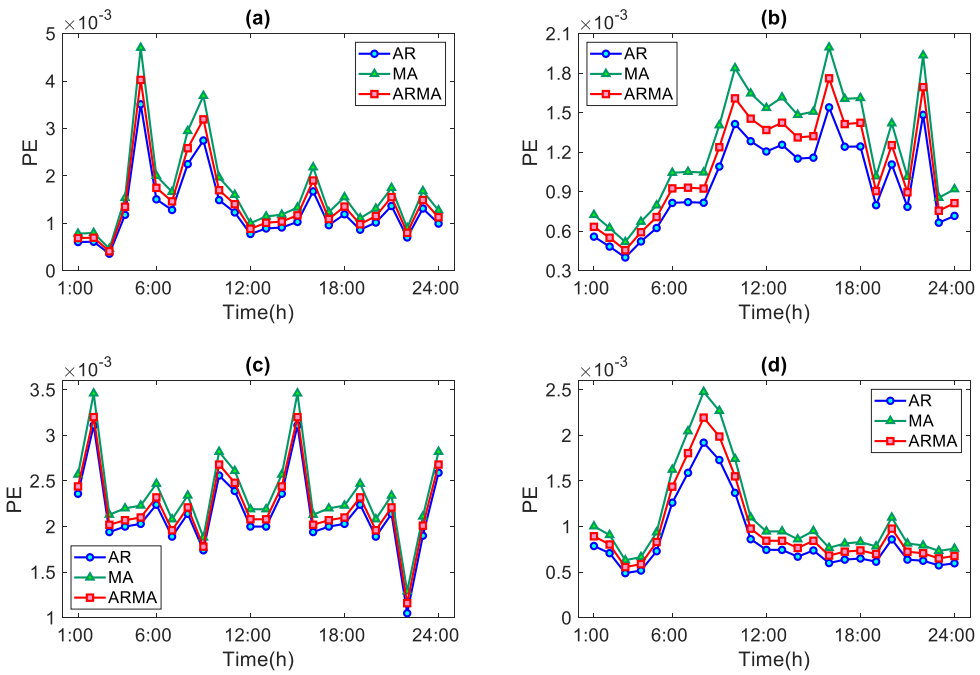


Fig. 5. Automatic model identification by the ARMAsel algorithm on Day 9: (a) Sensor 2, (b) Sensor 6, (c) Sensor 8, (d) Sensor 12

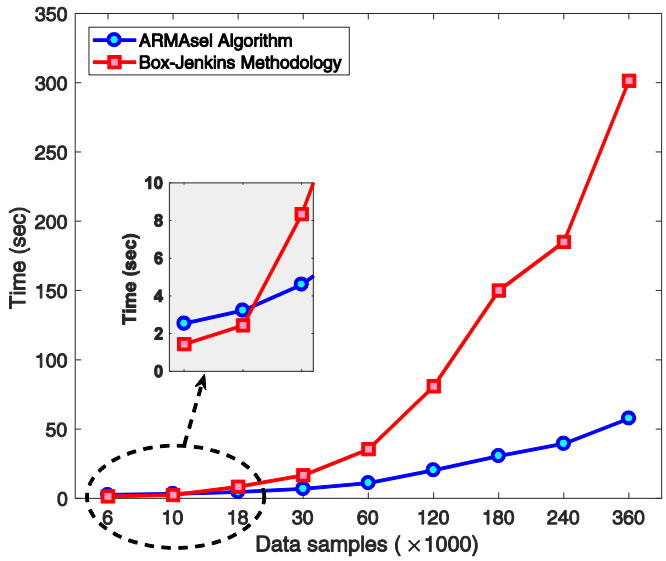


Fig. 6. Comparison between the computational costs of ARMAsel and Box-Jenkins methodologies

Fig. 6 shows a comparison of the computing time required by the automatic model identification method and by the graphical Box-Jenkins methodology, the latter based on plotting both the ACF and PACF. For this comparative analysis, the acceleration response of Sensor 8 at 3:00 PM of Day 1 is used and decomposed into nine datasets consisting of 6000,

10000, 18000, 30000, 60000, 120000, 180000, 240000, and 360000 samples. Results are obtained with a personal computer featuring an Intel™ Core i7-3770 @ 3.90 GHz CPU, and with 16GB of RAM; the MATLAB default function "`tic-toc`" was used to measure the analysis time. Keeping aside the discussed requirements and limitations of the Box-Jenkins methodology, through plots in Fig. 6 it can be observed that this graphical methodology is more efficient for small datasets (the first two ones in the inset), while the proposed automatic approach based on the ARMAsel algorithm becomes highly superior when the dimensionality of the data sequence increases.

5.2. AR-ARX modeling

As detailed in Section 2, the AR-ARX modeling involves three steps related to the determination of the AR and ARX orders $p, \bar{p}$, and $\bar{r}$, to the estimation of the model coefficients $\boldsymbol{\Theta}$, $\bar{\boldsymbol{\Theta}}$, and $\bar{\boldsymbol{\Phi}}$, and to the extraction of the AR and ARX residuals. The first two steps are implemented by handling the normal conditions in the baseline phase, and are based on the 24-h acceleration datasets of Days 1-7. The order of the AR model at each sensor location is set through the proposed order determination method, with an initial order p_0 obtained by the BIC method in a non-iterative form. If such order does not satisfy the LBQ test, it is then increased in the iterative stage. To check the uncorrelatedness of model residuals, Fig. 7 provides the LBQ test statistics relevant to residual analysis of the AR models in all test measurements. In these graphs, the dashed horizontal line represents the *c*-value handled in the LBQ test under a 5% significance limit, which amounts to 31.4. As can be seen, all the test statistics fall below the *c*-value, implying the uncorrelatedness of the AR model residuals. Table 2 gathers the corresponding values of the *R*-squared and adjusted *R*-squared statistics, used to investigate the possible occurrence of overfitting with the identified AR-ARX orders. The values of R^2 and $Adj\text{-}R^2$ in Table 2 are all close to one, with the $Adj\text{-}R^2$

values smaller than the R^2 ones: according to the discussion in Section 3, overfitting has not occurred. Hence, the proposed order determination method turns out to be efficient and capable to determine the series orders, leading to uncorrelated residuals without incurring into the overfitting problem.

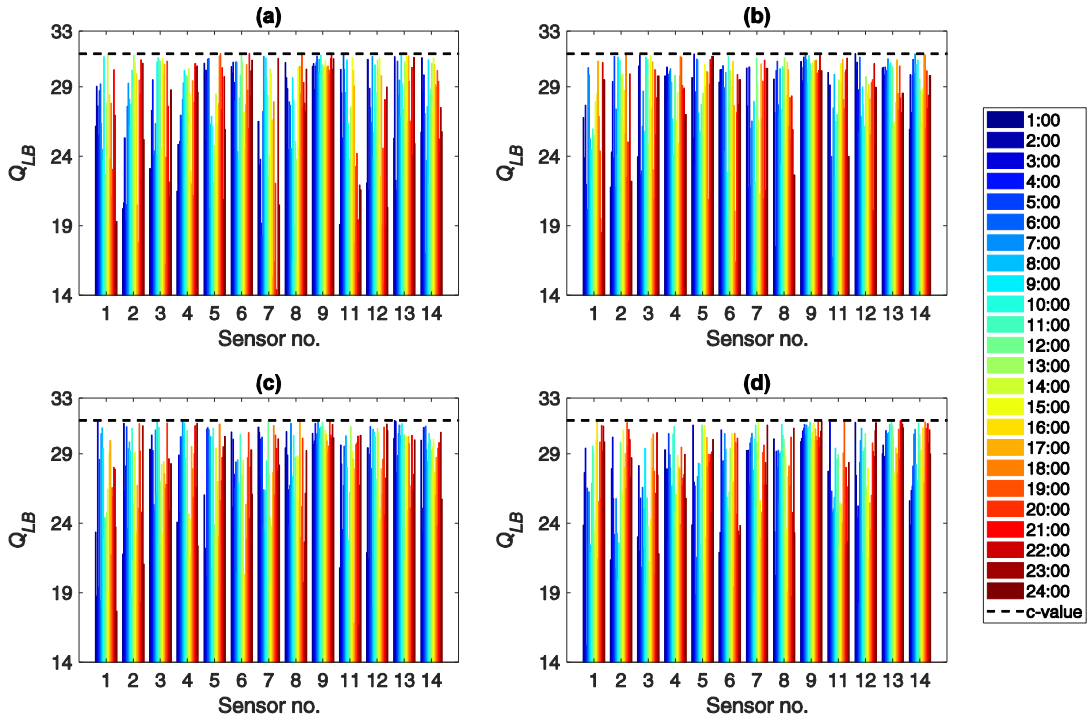


Fig. 7. Residual analysis via the LBQ test statistics under the 24-h vibration measurements on (a) Day 1, (b) Day 3, (c) Day 5, and (d) Day 7

To assess the benefits of the order determination method over the original one, which was suggested in [26], Fig. 8 and Fig. 9 give the evolution of the LBQ test statistics and the computing time needed for setting the minimum and maximum AR orders on Day 1, respectively concerning Sensors 9 and 7 (i.e. $p_{\min}$=59 and $p_{\max}$=174). The results show that the improved and original methods are both able to set the proper order of the AR model and to extract uncorrelated residuals, so that Q_{LB}<c-value. In Fig. 8(a) and Fig. 9(a), the iterative process begins at iteration 1, whereas the improved method first computes the initial orders by the BIC method and checks the correlation of the model residuals. The initial AR orders are

set in this latter case as 20 and 83, and are denoted in Fig. 9(b) and Fig. 9(b) with the vertical dashed lines. Since these values do not satisfy the LBQ test, they are next increased by the iterative algorithm. Although the improved method consists of the two stages (i.e. the non-iterative and iterative algorithms) and assesses the uncorrelatedness of the model residuals and also overfitting, it results to be less computational demanding than the original iterative-only technique. Therefore, it can be concluded that the improved method provides a more cost-efficient order determination algorithm compared to the original technique, especially when attacking the Big Data problem.

Table 2. Evaluation of the overfitting problem for the AR orders at 3:00 PM on Days 1, 3, 5, and 7

Sensor no.	Day no.							
	1		3		5		7	
	R^2	*Adj-*R^2	R^2	*Adj-*R^2	R^2	*Adj-*R^2	R^2	*Adj-*R^2
1	0.9205	0.9069	0.9051	0.8942	0.9175	0.8962	0.9375	0.9111
2	0.9329	0.9195	0.8967	0.8798	0.9019	0.8801	0.9080	0.8962
3	0.9132	0.8944	0.9040	0.8933	0.9043	0.8884	0.9409	0.9236
4	0.9277	0.9012	0.9363	0.9120	0.9219	0.9044	0.8949	0.8814
5	0.8827	0.8747	0.9060	0.8938	0.9112	0.8929	0.9179	0.8822
6	0.9153	0.9002	0.9239	0.9071	0.9024	0.8990	0.9092	0.8931
7	0.9222	0.9063	0.8937	0.8761	0.9422	0.9269	0.8770	0.8689
8	0.9003	0.8899	0.9074	0.8870	0.8967	0.8800	0.9112	0.8843
9	0.9318	0.9133	0.9221	0.8959	0.9044	0.8845	0.8639	0.8511
11	0.9041	0.8879	0.9114	0.8971	0.8930	0.8782	0.8948	0.8756
12	0.9292	0.9124	0.9216	0.8975	0.9006	0.8888	0.8892	0.8735
13	0.9012	0.8937	0.8800	0.8949	0.9063	0.8944	0.9120	0.8884
14	0.9223	0.9085	0.8768	0.8702	0.9267	0.9025	0.9037	0.8821

To obtain the orders of ARX representation associated with the second stage of the AR-ARX model, account is taken of the constraint $\bar{p} + \bar{r} \leq p$ suggested in [47]. As any selection of ARX orders compliant with this condition is plausible, an efficient way to set them is to use one of the information criteria [54].

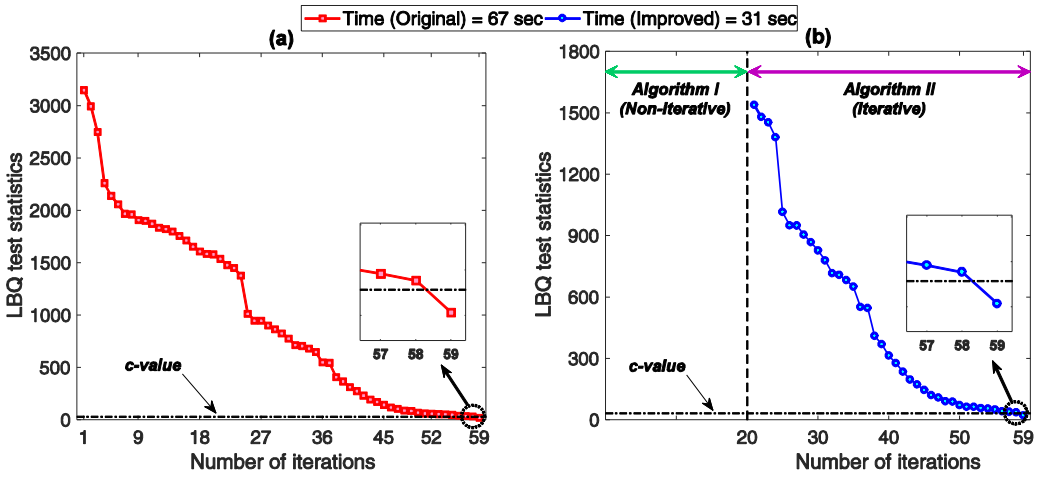


Fig. 8. Evolution of the LBQ test statistics and computational time to determine the AR order of the acceleration response of Sensor 9 at 5:00 PM on Day 1: (a) original iterative-only technique presented in [26], and (b) improved method

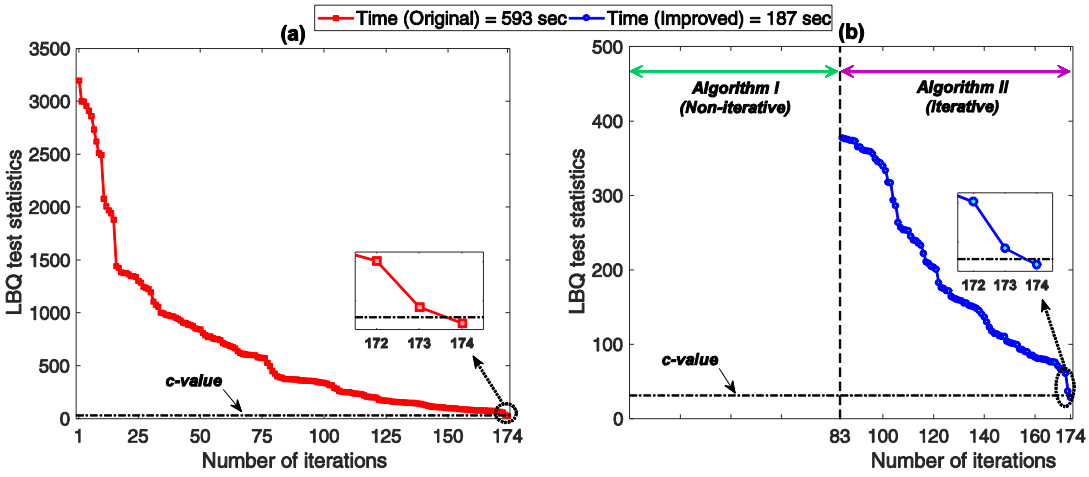


Fig. 9. Evolution of the LBQ test statistics and computational time to determine the AR order of the acceleration response of Sensor 7 at 11:00 AM on Day 1: (a) original iterative-only technique presented in [26], and (b) improved method

Setting $\bar{p} = \bar{r}$ and using the BIC method for the optimization of order selection, one can inspect the effects of the ARX orders by progressively increasing them up to $\bar{p} = \bar{r} = p/2$. Fig. 10 gathers the dependence of the BIC amount on the ARX orders and the finally selected values $\bar{p} = \bar{r}$, for Sensor 8 at 8:00 PM on Days 1, 3, 5, and 7.

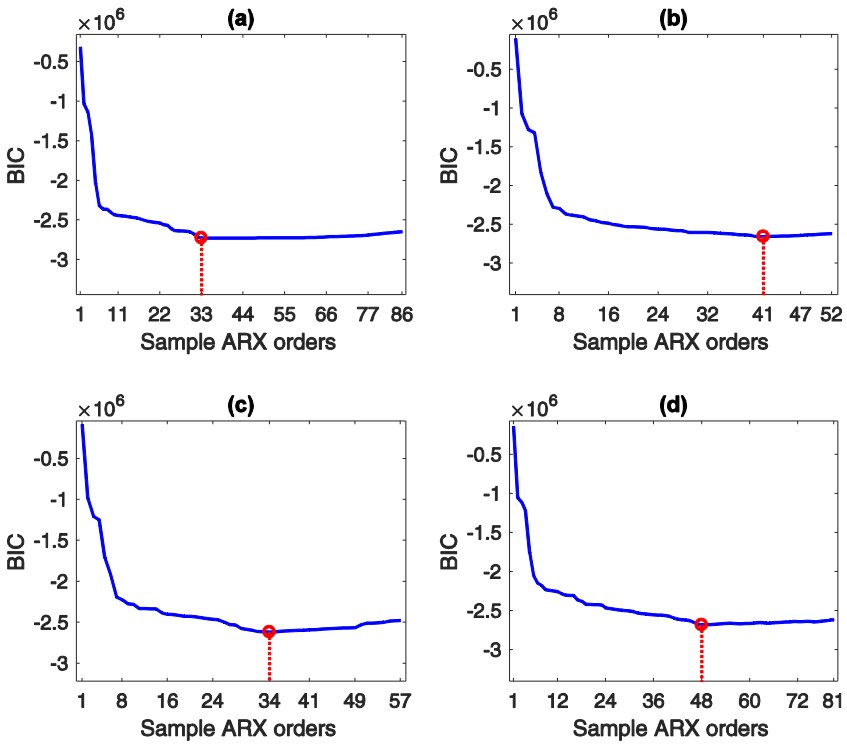


Fig. 10. Effect of the ARX orders $\bar{p} = \bar{r}$ on the BIC value (see Eq. 9) for Sensor 8 at 8:00 PM on (a) Day 1, (b) Day 3, (c) Day 5, and (d) Day 7

5.3. Residual-based feature extraction

Granted that the AR-ARX model turns out to be the most appropriate representation for the present analysis, the process of residual-based feature extraction is briefly discussed next. Having obtained the optimal AR orders in the baseline phase, the model coefficients are estimated by the least-squares technique. The residuals of the AR model for the conditions in the baseline and inspection phases are then extracted, to be used as the input data in the ARX representation. These residuals on Days 1-7 are finally extracted by adopting the obtained model orders $\bar{p}$ and $\bar{r}$, and the estimated coefficients $\bar{\boldsymbol{\Theta}} = [\bar{\theta}_1 \dots \bar{\theta}_{\bar{p}}]$ and $\bar{\boldsymbol{\Phi}} = [\bar{\varphi}_1 \dots \bar{\varphi}_{\bar{r}}]$. The

extracted residual datasets are used as the damage-sensitive features in the baseline phase. As explained, the same orders and coefficients associated with Days 1-7 are adopted to extract the residual samples on Days 8 and 9; the relevant residual datasets are used as the damage-sensitive features in the inspection phase. Since the numbers of measurements and sensor locations on any day are the same, each residual set for each of the structural states in the baseline and inspection phases (respectively Days 1-7 and Days 8-9) is a three-dimensional matrix containing 360,000 data points obtained from 13 sensors and 24-h measurements. For each of the seven individual residual datasets (matrices) for Days 1-7, the distance is computed via the PKLD to provide the training data $\mathbf{D}_{Tr}$. These matrices and the individual residual datasets for Days 8-9 are then used to set the testing set $\mathbf{D}_{Te}$.

5.4. Early damage detection

To detect damage with the proposed PKLD-MSD methodology, it is necessary to first define the training and testing datasets. As explained above, by considering the aforementioned n=360,000 residual sequences, the number of samples $\sigma=\sqrt{n}$ in each partition of the arranged set $\mathbf{E}_\mathbf{y}$ is 600; the number of partitions $c=n/\sigma$ therefore turns out to be 600 as well. In the present analysis, the reference residual vector is partitioned by means of the maximum entropy approach: given the c=600 partitions, Fig. 11 displays the number of samples in each partition of $\mathbf{E}_\mathbf{x}$ for Sensor 7 at 3:00 PM on Days 1, 3, 5, and 7.

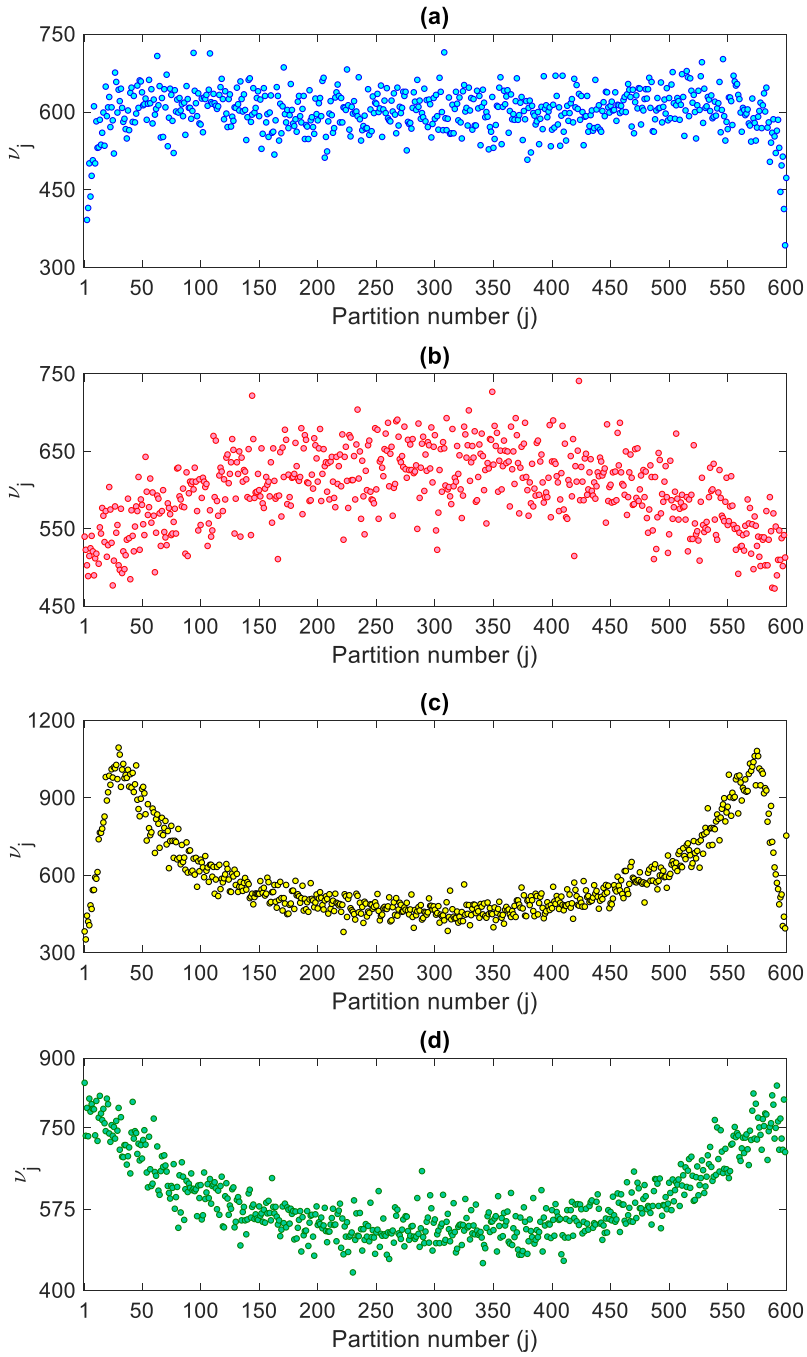


Fig. 11. Partitioning of the reference residual vector for Sensor 7 at 3:00 PM: (a) Day 1, (b) Day 3, (c) Day 5, (d) Day 7

To compute the PKLD of the residual sets of Days 1-7, the training data read $\mathbf{D}_{Tr} \in \mathbb{R}^{13\times1176}$ as n_L=7 and n_T=24. The testing matrix instead reads $\mathbf{D}_{Te} \in \mathbb{R}^{13\times2352}$ as, besides the values of n_L and n_T already reported, n_U=2; out of the 2352 vectors, the first and second halves respectively refer to Days 8 and 9. In this regard, Fig. 12 provides an illustrative comparison

between the variations in the PKLD values on Days 1-7 and those of either Day 8 or 9. It can be seen that the values relevant to Days 1-8 are roughly in the same range; at variance, one can observe a clear increase in the distance quantities on Day 9. Such a difference between Days 1-7 and Day 9 is indicative of damage occurrence.

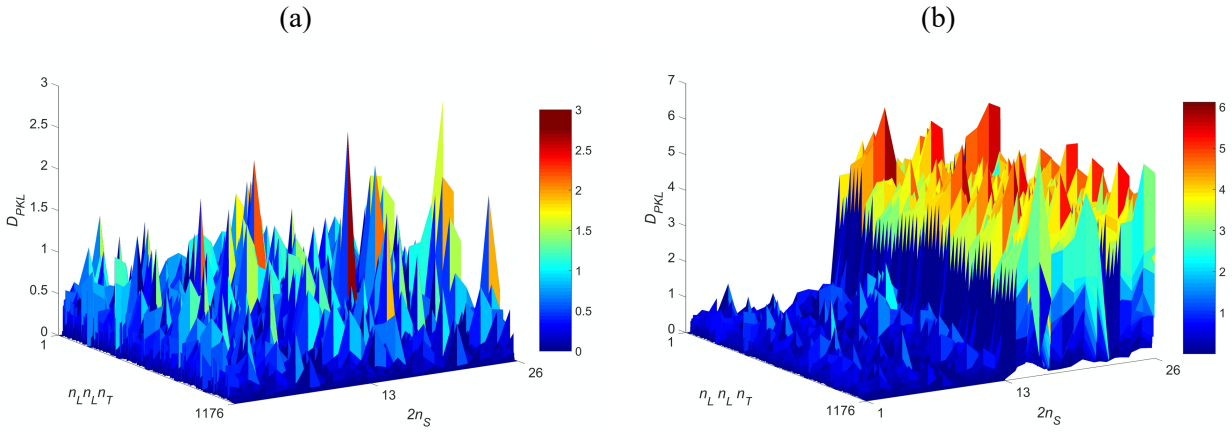


Fig. 12. Variations in the PKLD values: (a) Days 1-7 and Day 8, (b) Days 1-7 and Day 9

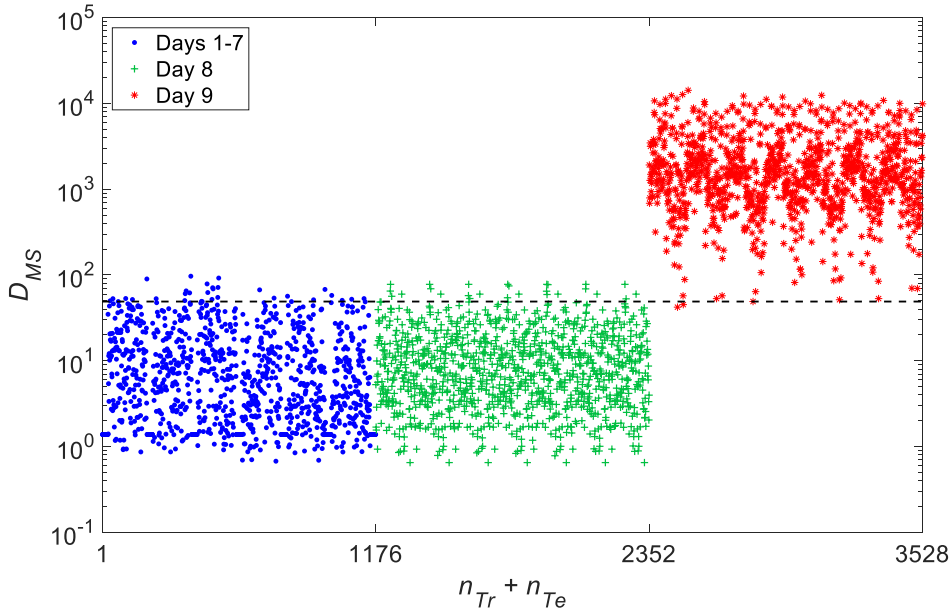


Fig. 13. Early damage detection by the proposed PKLD-MSD method, using information relevant to Days 1-7 in the baseline phase and Days 8-9 in the inspection stage: evolution of the D_{MS} values, see Eq. (15)

However, this comparison of the PKLD values may fail to be always useful, particularly when Big Data have to be processed if the number of current states increases. Using the training and testing datasets, the MSD formula can then be adopted to provide the final scores for early damage detection, as shown in Fig. 13. Here, the horizontal dashed line stands for the threshold limit estimated through the 95% confidence interval of the MSD values related to the training datasets. As can be seen, the values of D_{MS} associated to Days 1-7 and Day 8 are again in the same range and testify that the bridge on Day 8 is still undamaged. Further to that, the majority of MSD quantities lies under the threshold limit, which implies a good detection capability on Day 8 in the inspection phase. In contrast, it can be observed that almost all the MSD values regarding Day 9 exceed the threshold limit, which implies that the state of the bridge is now damaged. Even without referring to a specific threshold limit, it is evident that there exist obvious differences between the MSD quantities for Days 1-8, and Day 9. It can be also seen that the number and percentage of false positive and false negative errors, and misclassification rate are limited: they respectively amount to 55 (2.33%), 2 (0.17%), and 57 (1.61%). The proposed PKLD-MSD method thus succeeded in detecting early damage, even in the presence of potential environmental or loading variability.

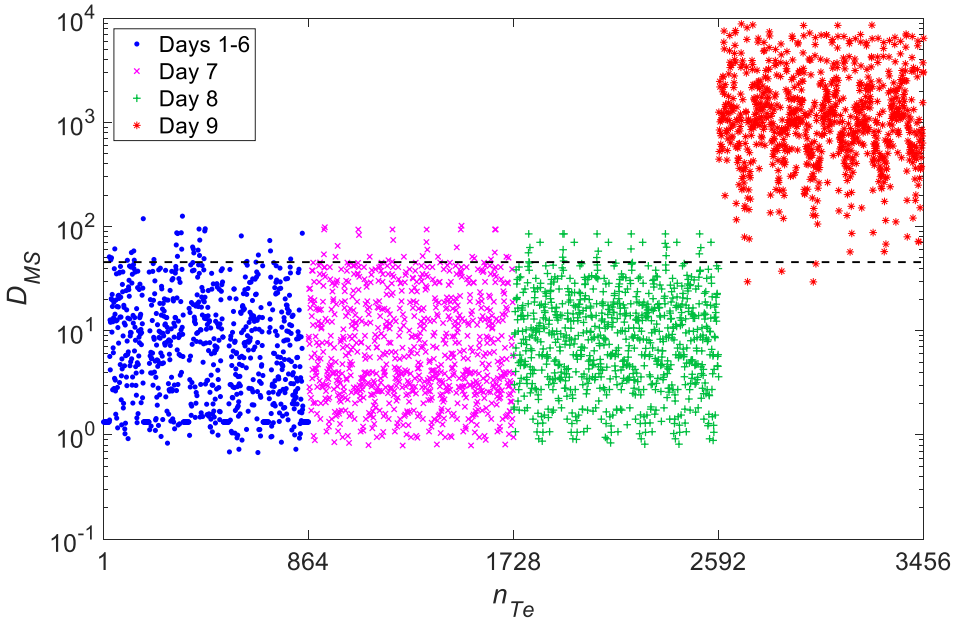


Fig. 14. Early damage detection by the proposed PKLD-MSD method using information relevant to Days 1-6 in the baseline phase and Days 7-9 in the inspection stage: evolution of the D_{MS} values

In order to assess the effect on the outcomes of the number of normal conditions and the size of the training data, the procedures of feature extraction (via AR-ARX modeling) and feature classification (via the PKLD-MSD method) are repeated by accounting for Days 1-6 as normal conditions in the baseline phase (n_L=6) and Days 7-9 as the current states in the inspection phase (n_U=3). In this case, the training and testing matrices become $\mathbf{D}_{Tr} \in \mathbb{R}^{13\times 864}$ and $\mathbf{D}_{Te} \in \mathbb{R}^{13\times 2592}$. Results in terms of early damage detection for this scenario are shown in Fig. 14, where the horizontal line still represents the threshold limit. Also, in this case, the MSD values for Days 1-8 are similar and most of them fall under the threshold; on the other hand, the majority of MSD values for Day 9 exceed the threshold. The number and percentage of false positive and negative errors, and the misclassification rate are now respectively equal to 90 (3.47%), 4 (0.46%) and 94 (2.72%). Comparing these amounts with those related to Fig. 13, it is shown that the reduction of training duration increases the rates of false positive and negative, and the misclassification errors.

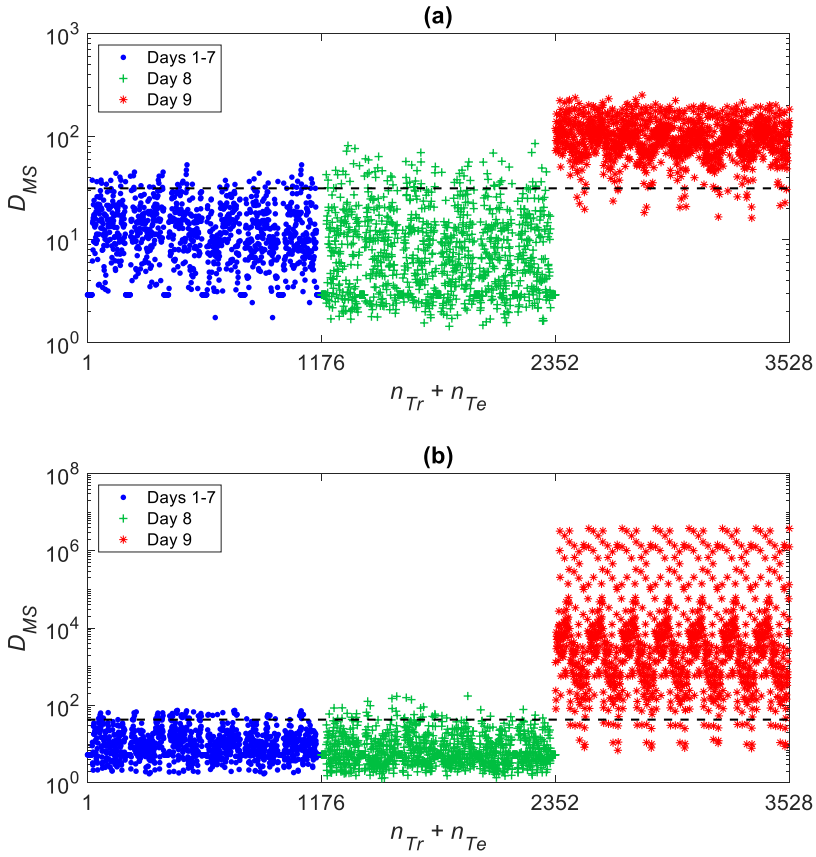


Fig. 15. Early damage detection using information relevant to Days 1-7 in the baseline phase and Days 8-9 in the inspection stage: evolution of the D_{MS} values obtained via (a) KSD-MSD, (b) ESD-MSD

To further assess the performance of the PKLD method, training and testing datasets are processed via the well-known Euclidean-square distance (ESD) and Kolmogorov-Smirnov distance (KSD) procedures [34]. For this comparison, Days 1-7 and Days 8-9 are respectively handled as the normal and current states. The results related to early damage detection via the KSD-MSD and ESD-MSD techniques are reported in Fig. 15: similarly to the present method results, the values of D_{MS} concerning Days 1-7 and Day 8 are roughly the same and mostly fall under the threshold limits; values for Day 9 are instead larger than the threshold, to represent that the current state on this day is damaged. Although all the three novelty detection methods attain the same conclusion as far as early damage assessment is concerned, the proposed PKLD-MSD method outperforms the other techniques as testified by the smaller numbers of false positive and false negative errors. To quantitatively assess this issue, Table 3

lists the number and percentage of these errors, as well as the misclassification rate. Furthermore, Fig. 16 provides the receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC) values for the three novelty detection methods; here, TPR and FPR respectively refer to the true positive and the false positive rates. Both the KSD-MSD and ESD-MSD techniques provide larger false positive, false negative and misclassification errors in comparison with the proposed PKLD-MSD method. This results clearly testifies a better performance of the newly proposed method, with an AUC value closer to one.

Table 3. Comparison of the three novelty detection methods in terms of the false positive and false negative errors, and misclassification rate

Method	False positive	False negative	Misclassification rate
PKLD-MSD	55 (2.33%)	2 (0.17%)	57 (1.61%)
KSD-MSD	137 (5.82%)	26 (2.21%)	163 (4.62%)
ESD-MSD	100 (4.25%)	49 (4.16%)	149 (4.22%)

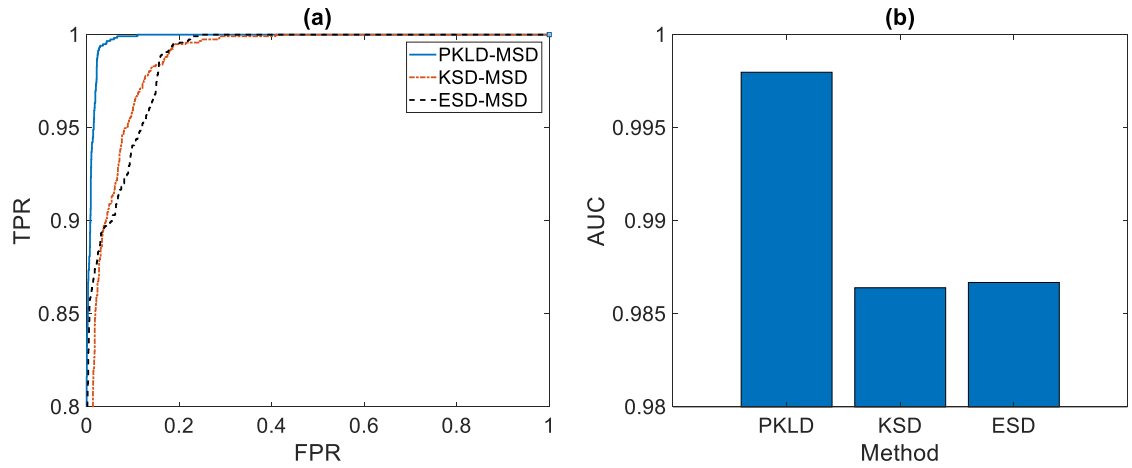


Fig. 16. Comparison of the performances of PKLD, KSD, and ESD for detecting damage: (a) ROC curves, and (b) AUC values

All the previous results have been based on the use of all accelerometers distributed on the bridge. Due to the importance of the sensor network density, a comparison is carried out now by assessing the performance of the proposed methods using a varying number of

sensors, or datasets. For this comparison, the rates of false positive, false negative, and misclassification errors are listed in Table 4 for five difference cases: Case 1 refers to the handling of all the 13 sensors (i.e. the accelerometers 1-9 and 11-14); in the other cases, the number of accelerometers for response modeling and to provide the damage-sensitive features is decreased gradually. As the data in Table 4 reveal, the best performance, in terms of values of false positive, false negative and misclassification errors, is obtained by adopting the complete set of the accelerometers of Case 1. In contrast, the worst performance is given by Case 5, for which the vibration responses and the features of the four accelerometers at the midspan of the bridge have been considered to detect damage. Moreover, the rates of the three mentioned errors all increase by reducing the number of sensors. It is seen that, although the reduction in the number of sensors decreases the size of the damage-sensitive features (i.e. the rows of the matrices $\mathbf{D}_{Tr}$ and $\mathbf{D}_{Te}$), it is preferable to exploit a relatively dense sensor network to capture all the possible features sensitive to damage.

Table 4. Effects of the number of sensors used for SHM on the false positive, false negative errors and misclassification rate.

Case no.	Sensor labels	False positive	False negative	Misclassification
1	1-9 & 11-14	55 (2.33%)	2 (0.17%)	57 (1.61%)
2	1,2,5,6,9,13,14	101 (4.29%)	39 (3.31%)	140 (3.96%)
3	1,3,5,7,9,11,13	110 (4.67%)	36 (3.06%)	146 (4.13%)
4	5,6,7,8,9	158 (6.71%)	61 (5.18%)	219 (6.20%)
5	4,6,8,12	184 (7.82%)	85 (7.22%)	269 (7.62%)

6. Conclusions

In this paper, we have proposed innovative statistical pattern recognition methods based on time series modeling and novelty detection for early damage assessment in the

presence of big data to process. These methods consist of an automatic, non-graphical algorithm for model identification, an improved order determination approach, and a distance-based hybrid novelty detection termed PKLD-MSD. The performance and effectiveness of the proposed methods have been investigated by using the experimental datasets from vibration measurements of the Tianjin-Yonghe bridge: for 24-h test measurements at 13 sensor locations on nine days, the experimental datasets to handle are characterized by substantially large volumes of random high-dimensional feature samples.

The main achievements of this work can be summarized as follows. Among all the polynomial model classes, the AR-ARX representation has been shown to be the most appropriate one based on engineering and statistical aspects. Model identification by the ARMAsel algorithm has allowed to automatically select the best time series model without any user expertise required; a comparison with the graphical Box-Jenkins methodology has shown that the ARMAsel algorithm is more efficient and faster. The proposed improved approach to model order determination has been successful in obtaining optimal orders of the AR models, extracting uncorrelated residuals and simultaneously preventing overfitting. A comparison with a former iterative-only order determination technique has proved that the approach works faster. The method therefore stands as an appropriate and efficient tool for the model order determination in the presence of Big Data.

The proposed PKLD-MSD method has been successfully adopted to detect damage in a cable-stayed bridge, using different training datasets and handling high-dimensional features. It has been reported that large training samples must be exploited to attain appropriate performances in terms of early damage detection. At any rate, a comparison of PKLD with the state-of-the-art distance measures ESD and KSD has revealed that the adopted method is more reliable and more performative to avoid false positive, false negative and misclassification errors. Finally, it has been observed that the use of a relatively dense sensor

network can guarantee a good performance of the proposed methods for early damage detection.

Even by accounting for the reliable and accurate results obtained, some issues may be considered to further assess the performance and robustness of the proposed methods. Due to the importance of a continuous long-term SHM approach, the effectiveness in case of long-term data will be evaluated; the sensitivity to measurement noise of the results accuracy will be evaluated as well.

Acknowledgment

The authors are indebted to the SMC Group at the Harbin Institute of Technology in China, for granting the access to their experimental datasets. This research was partially supported by Grant No. 96007230 from the Iran National Science Foundation (INSF).

References

[1] Farrar CR, Worden K. Structural Health Monitoring: A Machine Learning Perspective Chichester, United Kingdom: John Wiley & Sons Ltd; 2013.
[2] Liang D, Yuan S. Structural health monitoring system based on multi-agent coordination and fusion for large structure. Advances in Engineering Software. 2015;86:1-12.
[3] Santos JP, Crémona C, Calado L, Silveira P, Orcesi AD. On-line unsupervised detection of early damage. Structural Control and Health Monitoring. 2016;23:1047-69.
[4] Ubertini F, Cavalagli N, Kita A, Comanducci G. Assessment of a monumental masonry bell-tower after 2016 Central Italy seismic sequence by long-term SHM. Bulletin of Earthquake Engineering. 2018;16:775-801.
[5] Sarmadi H, Karamodin A, Entezami A. A new iterative model updating technique based on least squares minimal residual method using measured modal data. Applied Mathematical Modelling. 2016;40:10323-41.
[6] Rezaiee-Pajand M, Entezami A, Sarmadi H. A sensitivity-based finite element model updating based on unconstrained optimization problem and regularized solution methods. Structural Control and Health Monitoring. 2020;27:e2481.
[7] Sanayei M, Rohela P. Automated finite element model updating of full-scale structures with PARameter Identification System (PARIS). Advances in Engineering Software. 2014;67:99-110.
[8] Azam SE, Mariani S. Online damage detection in structural systems via dynamic inverse analysis: A recursive Bayesian approach. Engineering Structures. 2018;159:28-45.

[9] Azam SE, Mariani S, Attari N. Online damage detection via a synergy of proper orthogonal decomposition and recursive Bayesian filters. Nonlinear Dynamics. 2017;89:1489-511.
[10] Capellari G, Chatzi E, Mariani S. Structural health monitoring sensor network optimization through Bayesian experimental design. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering. 2018;4:04018016.
[11] Capellari G, Eftekhar Azam S, Mariani S. Damage detection in flexible plates through reduced-order modeling and hybrid particle-Kalman filtering. Sensors. 2016;16:2.
[12] Fan W, Qiao P. Vibration-based damage identification methods: a review and comparative study. Structural Health Monitoring. 2011;10:83-111.
[13] Entezami A, Shariatmadar H, Sarmadi H. Structural damage detection by a new iterative regularization method and an improved sensitivity function. Journal of Sound and Vibration. 2017;399:285-307.
[14] Kaveh A, Zolghadr A. An improved CSS for damage detection of truss structures using changes in natural frequencies and mode shapes. Advances in Engineering Software. 2015;80:93-100.
[15] Nhamage IA, Lopez RH, Miguel LFF. An improved hybrid optimization algorithm for vibration based-damage detection. Advances in Engineering Software. 2016;93:47-64.
[16] Dinh-Cong D, Vo-Duy T, Ho-Huu V, Dang-Trung H, Nguyen-Thoi T. An efficient multi-stage optimization approach for damage detection in plate structures. Advances in Engineering Software. 2017;112:76-87.
[17] Entezami A, Shariatmadar H, Ghalehnovi M. Damage detection by updating structural models based on linear objective functions. Journal of Civil Structural Health Monitoring. 2014;4:165-76.
[18] Entezami A, Shariatmadar H, Mariani S. Fast unsupervised learning methods for structural health monitoring with large vibration data from dense sensor networks. Structural Health Monitoring. in press.
[19] Sarmadi H, Entezami A, Ghalehnovi M. On model-based damage detection by an enhanced sensitivity function of modal flexibility and LSMR-Tikhonov method under incomplete noisy modal data. Engineering with Computers. 2020.
[20] Rageh A, EftekharAzam S, Linzell DG. Steel railway bridge fatigue damage detection using numerical models and machine learning: Mitigating influence of modeling uncertainty. International Journal of Fatigue. 2020;134:105458.
[21] Gul M, Necati Catbas F. Statistical pattern recognition for Structural Health Monitoring using time series modeling: Theory and experimental verifications. Mechanical Systems and Signal Processing. 2009;23:2192-204.
[22] Jiang S-F, Fu C, Zhang C. A hybrid data-fusion system using modal data and probabilistic neural network for damage detection. Advances in Engineering Software. 2011;42:368-74.
[23] Entezami A, Shariatmadar H. Structural health monitoring by a new hybrid feature extraction and dynamic time warping methods under ambient vibration and non-stationary signals. Measurement. 2019;134:548-68.
[24] Entezami A, Shariatmadar H. Damage localization under ambient excitations and non-stationary vibration signals by a new hybrid algorithm for feature extraction and multivariate distance correlation methods. Structural Health Monitoring. 2019;18:347-75.
[25] Ghiasi R, Ghasemi MR, Noori M. Comparative studies of metamodeling and AI-Based techniques in damage detection of structures. Advances in Engineering Software. 2018;125:101-12.

[26] Entezami A, Shariatmadar H. An unsupervised learning approach by novel damage indices in structural health monitoring for damage localization and quantification. Structural Health Monitoring. 2018;17:325-45.
[27] Sarmadi H, Entezami A, Daneshvar Khorram M. Energy-based damage localization under ambient vibration and non-stationary signals by ensemble empirical mode decomposition and Mahalanobis-squared distance. Journal of Vibration and Control. 2020;26:1012-27.
[28] Entezami A, Sarmadi H, Saeedi Razavi B. An innovative hybrid strategy for structural health monitoring by modal flexibility and clustering methods. Journal of Civil Structural Health Monitoring. 2020.
[29] Argyris C, Chowdhury S, Zabel V, Papadimitriou C. Bayesian optimal sensor placement for crack identification in structures using strain measurements. Structural Control and Health Monitoring. 2018;25:e2137.
[30] Amezquita-Sanchez JP, Adeli H. Signal Processing Techniques for Vibration-Based Health Monitoring of Smart Structures. Archives of Computational Methods in Engineering. 2016;23:1-15.
[31] Entezami A, Sarmadi H, Behkamal B, Mariani S. Big data analytics and structural health monitoring: A statistical pattern recognition-based approach. Sensors. 2020;20:2328.
[32] Kopsaftopoulos F, Fassois S. Vibration based health monitoring for a lightweight truss structure: experimental assessment of several statistical time series methods. Mechanical Systems and Signal Processing. 2010;24:1977-97.
[33] Entezami A, Shariatmadar H, Karamodin A. Data-driven damage diagnosis under environmental and operational variability by novel statistical pattern recognition methods. Structural Health Monitoring. 2019;18:1416-43.
[34] Roy K, Bhattacharya B, Ray-Chaudhuri S. ARX model-based damage sensitive features for structural damage localization using output-only measurements. Journal of Sound and Vibration. 2015;349:99-122.
[35] Mei L, Mita A, Zhou J. An improved substructural damage detection approach of shear structure based on ARMAX model residual. Structural Control and Health Monitoring. 2016;23:218–36.
[36] Kim YJ, Queiroz LB. Big Data for condition evaluation of constructed bridges. Engineering Structures. 2017;141:217-27.
[37] Gulgec NS, Shahidi GS, Matarazzo TJ, Pakzad SN. Current Challenges with BIGDATA Analytics in Structural Health Monitoring. Structural Health Monitoring & Damage Detection, Volume 7: Springer; 2017. p. 79-84.
[38] Rezaiee-Pajand M, Entezami A, Shariatmadar H. An iterative order determination method for time-series modeling in structural health monitoring. Advances in Structural Engineering. 2017;21:300-14
[39] Figueiredo E, Figueiras J, Park G, Farrar CR, Worden K. Influence of the Autoregressive Model Order on Damage Detection. Computer-Aided Civil and Infrastructure Engineering. 2011;26:225-38.
[40] Box GE, Jenkins GM, Reinsel GC, Ljung GM. Time Series Analysis: Forecasting and Control. 5th Edition: John Wiley & Sons; 2015.
[41] Gomez V, Maravall A. Automatic modeling methods for univariate series: Wiley Online Library; 1998.
[42] Eftekhar Azam S, Rageh A, Linzell D. Damage detection in structural systems utilizing artificial neural networks and proper orthogonal decomposition. Structural Control and Health Monitoring. 2019;26:e2288.

[43] Sarmadi H, Karamodin A. A novel anomaly detection method based on adaptive Mahalanobis-squared distance and one-class kNN rule for structural health monitoring under environmental effects. Mechanical Systems and Signal Processing. 2020;140:106495.
[44] Gobbato M, Kosmatka J, Conte J. Developing an integrated structural health monitoring and damage prognosis (SHM-DP) framework for predicting the fatigue life of adhesively-bonded composite joints. Fatigue and Fracture of Adhesively-Bonded Composite Joints: Elsevier; 2015. p. 493-526.
[45] Catbas FN, Malekzadeh M. A machine learning-based algorithm for processing massive data collected from the mechanical components of movable bridges. Automation in Construction. 2016;72:269-78.
[46] Li S, Li H, Liu Y, Lan C, Zhou W, Ou J. SMC structural health monitoring benchmark problem using monitored data from an actual cable-stayed bridge. Structural Control and Health Monitoring. 2014;21:156-72.
[47] Ljung L. System Identification: Theory for the User. 2nd Edition Upper Saddle River, NJ: Prentice-Hall; 1999.
[48] Broersen PM. Automatic spectral analysis with time series models. IEEE Transactions on Instrumentation and Measurement. 2002;51:211-6.
[49] Broersen PM, Wensink HE. On the penalty factor for autoregressive order selection in finite samples. IEEE Transactions on Signal Processing. 1996;44:748-52.
[50] Montgomery DC, Jennings CL, Kulahci M. Introduction to time series analysis and forecasting: John Wiley & Sons; 2015.
[51] Wang X, Mueen A, Ding H, Trajcevski G, Scheuermann P, Keogh E. Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery. 2013;26:275–309.
[52] Zhou W, Li S, Li H. Damage detection for SMC benchmark problem: A subspace-based approach. International Journal of Structural Stability and Dynamics. 2016;16:1640025.
[53] Nguyen T, Chan TH, Thambiratnam DP. Field validation of controlled Monte Carlo data generation for statistical damage identification employing Mahalanobis squared distance. Structural Health Monitoring. 2014;13:473-88.
[54] Wei WWS. Time Series Analysis: Univariate and Multivariate Methods. 2nd Edition: Pearson Addison Wesley; 2006.