The use of ROC and precision-versus-recall curves as performance metrics in unsupervised structural damage classification under changing environment

Valentina Giglioni, Enrique García-Macías, Laura Ierimonti, Ilaria Venanzi, Filippo Ubertini

Department of Civil and Environmental Engineering, University of Perugia, Via G. Duranti 93, Perugia 06125, Italy

Elsevier Inc[a,b], Global Customer Service[b,]*

[a]1600 John F Kennedy Boulevard, Philadelphia
[b]360 Park Avenue South, New York

Abstract

The development of long-term structural health monitoring systems is recently receiving a growing scientific interest in the field of Civil Engineering. In the context of unsupervised learning processes, deviations of dynamic parameters from their normal conditions can allow damage detection. However, due to the fact that modal properties are highly sensitive to environmental and operational factors, it is extremely important to remove such effects in order to obtain suitable damage sensitive features. In this regard, the selection of a proper statistical model for removing environmental effects is not a trivial issue as the distribution of the residuals, the control chart and therefore the damage detection inevitably depend on it. To overcome this problem, an original methodology is developed in the present paper, based on Receiving Operating Characteristic (ROC) curves in combination with Precision-versus-Recall (PR) curves, with the aim to provide a new decision-support tool for the definition of the best environmental effects' removal technique. Specifically, ROC and PR

⋆Fully documented templates are available in the elsarticle package on CTAN.

*Valentina Giglioni
Email address: support@elsevier.com (Global Customer Service)
URL: www.elsevier.com (Elsevier Inc)

curves are computed and compared for a variety of statistical models for data normalization and different damage scenarios. The proposed approach is exemplified by application in two case studies of continuously monitored structures: the Z24 Bridge in Switzerland and the Consoli Palace, a medieval masonry building in Italy. The results highlight that the combined use of both ROC and precision-versus-recall curves represents a suitable tool for defining the most effective data normalization method and the optimal damage threshold, in order to minimize the occurrence of false alarms detection.
Keywords: `Vibration-based SHM, ROC curves, Precision-recall curves, Unsupervised learning, Damage detection`

1. Introduction

In the last decades, Structural Health Monitoring (SHM) has become increasingly popular, gaining a key role in the field of Civil Engineering. One of the reasons lies in the urging need to better manage the large number of ageing structures and infrastructures. Particular attention is paid to cultural heritage buildings, which necessarily require a strategy of maintenance and conservation due to their vulnerability [1, 2, 3] and their exposure to various types of natural hazards, such as earthquakes [4, 5]. In parallel, the whole society is becoming aware of the importance of major structures like bridges, as they represent critical elements in modern transport networks. However, at the same time, they are inevitably subjected to materials' degradation under normal conditions, as well as to extreme events. As an illustrative example, the EU funded BRIME project in 2001 identified that highway bridges in France, Germany and the UK present lacks at a rate of 39%, 30% and 37%, respectively [6]. Furthermore, catastrophic events as the collapse of the Genoa bridge in 2018 or the breakdown of the highway overpass close to Ancona in 2017 have recently brought to the light the urgency to invest in infrastructure management. For all these reasons, it is crucial to develop reliable monitoring tools with a fully non-destructive character, in support to visual inspections and non-destructive evaluations, to

ensure bridges safety [7, 8, 9, 10, 11].

Within this context, the popularity of long-term vibration-based SHM systems [12, 13, 14] has been steadily growing in the last years. Such methods exploit ambient acceleration records under normal operating conditions to extract modal parameter estimates through Operational Modal Analysis (OMA) [15, 16, 17, 18]. Given that modal features depend upon the physical properties of the structure (mass, stiffness and energy dissipation properties), the idea is to detect damage by tracking deviations of dynamic parameters from their normal conditions through statistical pattern recognition [19]. This is what data-driven approaches deal with, providing a base for unsupervised learning processes. Thanks to the advances made in sensing technologies, data migration facilities and data pre/post processing algorithms, data-driven techniques are well suitable for continuous SHM systems, whose goals are manifold: (i) they allow to track any change in the structural behaviour, due to some damage, starting from the variation in time of modal parameters, (ii) they provide real-time information for safety assessment and early-stage damage identification [20], (iii) they permit optimal scheduling of local inspections and non-destructive evaluation tests in view of recovery activities and (iv) they can be used for inverse calibration of numerical models [21, 22].

Despite these advantages, it is worth highlighting that the main drawback of dealing with changes in modal properties is represented by their high sensitivity to alterations in environmental and operational conditions [23, 24]. Regarding this aspect, Cornwell et al. [25] studied the thermal fluctuations in the dynamic features of the Alamosa Canyon Bridge and found about 5% daily changes in the first three natural frequencies. Rohrmann et al. [26] observed that variations of frequencies in the Westend bridge produced by temperature can reach 10% according to monitoring results obtained from 1994 to 1997. This kind of information leads to the concept that alterations in modal properties due to the temperature effects are often more remarkable than those caused by a medium degree of structural damage [27] and operation loads [28], therefore the risk of masking early stage damages is significant. For this reason, numerous works

have been recently devoted to estimating and removing environmental effects from recorded monitoring data in order to define suitable damage-sensitive features [1, 29, 30].

There are different techniques, known in the literature, to cope with the data normalization problem. The basic idea beyond these theories is to create a statistical model able to reproduce the part of variance in frequency estimates that is associated with changes in environmental conditions.

Multiple Linear Regression (MLR) [29, 31, 32] is a statistical tool which exploits linear correlations for predicting values of one or more dependent variables, as can be natural frequencies, starting from a group of independent variables, which are tipically environmental and operational factors.

Another well-established methodology, Principal Component Analysis (PCA) [33, 34, 35, 36], has the aim to convert a set of observations of possibly correlated variables into a group of values of linearly uncorrelated variables called principal components. The main advantage is that PCA does not require to measure the environmental parameters as they are taken into account as embedded variables.

Within the context of bridge monitoring, Comanducci et al. [37] studied the comparison between different statistical tools, while Sohn et al. [38] and Hu et al. [31] applied the multiple regression model to remove temperature effects on the Alamosa Canyon Bridge and on a prestressed-concrete box girder bridge in Berlin, respectively. Other researches carried out by Yan et al. [39] highlighted the effectiveness of PCA in the Z24 bridge case study. Beyond that, the results obtained from the combination of MLR and PCA were discussed by Ubertini et al. [1], who analysed the damage detection of the bell-tower of the Basilica of San Pietro in Perugia.

Once a proper statistical model is defined, the prediction error is conceivably calculated as the difference between the identified modal frequencies and those independently estimated through the statistical model. These quantities, called residuals, should be only minimally affected by environmental effects and therefore suitable for damage detection purposes. Consequently, it is fair to say with

a certain level of confidence that any anomaly in the residuals, in the form of statistical outliers, corresponds to a damage condition. On this basis, Novelty Analysis through control charts [40] is commonly used to infer the presence of damage. Control charts assess certain statistical distances between newly acquired data and a baseline population (training period), which represents the healthy state of the structure.

In this context, the definition of the threshold in the classification (damage or non-damaged) represents one of the trickiest issues, which inevitably leads to a certain number of erroneus predictions, including False Positives and False Negatives. All that may translate into a distorted interpretation of the control chart, in view of a reasonable damage identification. Minimizing the rate of damage detection errors is critical for an effective monitoring. To this aim, it is crucial to select the most appropriate statistical tool for data normalization for the specific case study. Hence, this paper proposes a new method where the statistical model selection problem is optimally solved by considering ROC curves [41] and PR curves [42] that give a complete quantification assessment of monitoring errors. More specifically, they are graphical-based tools for quantifying the performance of a process, varying the threshold definition. ROC curves find applications in several fields [43]. They are used in the medical disciplines for the evaluation of diagnostic tests [44], for damage detection techniques of bridges by using Artificial Neural Networks [45], for the assessment of the damage identification performance of guided wave SHM systems [46] and for the design of the monitoring system of precast reinforced concrete (RC) industrial buildings in seismic hazard zones [47]. However, using a ROC curve with an unbalanced dataset might be deceptive and lead to incorrect interpretations of model's performance. To overcome this issue, which is very common in machine learning field, precision-versus-recall curves turned out to be a valid alternative [42, 48].

Despite the manifold usages of ROC and PR curves in the literature, their use for optimal statistical pettern recognition in the context of damage detection through control charts is still unexplored.

In this regard, it is worth pointing out that the way the environmental effects are removed influences the distribution of the residuals, the control chart and therefore the damage detection. For this reason, comes to the light the need of a strategy aiming at choosing the best data normalization technique for a reliable structural damage detection. Such a purpose is accomplished in this paper by providing an original methodology which uses ROC and PR curves as performance metrics. The basic idea is to supply objective criteria enabling to define (i) which removal technique would ensure the best damage identification and (ii) which cut-off value minimizes the number of false positives.

Preliminarly, residuals obtained by the application of different environmental effects removal techniques are plotted in control charts. Several ROC and PR curves, each one referred to a particular data normalization procedure and to a specific damage class, are afterwards computed. The performance of each curve is assessed by considering a linear combination between two helpful parameters, namely the area under the ROC curve and the area under the PR curve. The procedure's purpose is to maximize an objective function, involving both parameters, to find out the most suitable model. Hence, the developed approach represents a new decision-support system for assessing the effectiveness of any data normalization technique as well as for selecting the optimal damage threshold, leading to a minimization of false alarms detection. Since the outcome of this method highly depends upon the damage scenario, it is important to clearly define, in its practical applications, which is the damage mechanism (or mechanisms) of interest. In those cases where damage data are not available, frequency decays ought to be obtained by non linear FEM simulations.

In order to support the advantages of the proposed approach, two case studies are presented: Palazzo Dei Consoli, a medieval masonry palace located in Gubbio, Italy and the Z24 Bridge, located in the canton Bern near Solothurn, Switzerland. In the firt case, the different damage scenarios have been simulated through a FEM model, while, in the second one, damage data had been collected during field tests.

The paper is organized as follows. Section 2 overviews the theoretical back-

groung of the statistical pattern recognition tools used in this work, Section 3 presents the new methodology aimed at the application of the ROC and PR curves for damage detection and minimization of false alarms under changing environment, Section 4 describes the main characteristics and the monitoring system of the Consoli Palace and the Z24 bridge, Section 5 illustrates the ability of the aforementioned curves to provide reliable performance metrics in unsupervised processes with reference to two case studies. Finally, Section 6 discusses the main conclusions of this paper.

2. Background

Numerous tools have been recently dedicated to estimating and removing environmental effects from monitoring data in order to be able to discern damage-induced changes in the natural frequency time-histories. In this framework, the basic theory of MLR and PCA is briefly described hereafter. In addition, with the aim of extending these approaches for the modelling of non-linear environmental effects, a clustering approach based upon the Gaussian Mixture Model (GMM) is also presented.

2.1. Multivariate Linear Regression (MLR)

Multivariate Regression Models exploit linear correlations between a set of n dependent variables (estimators) and a set of p independent variables (predictors). In this work, dependent variables are the identified modal frequencies, while independent variables can be environmental parameters. The established model is adopted to understand the influence of each predictor on the dependent variables and therefore, to predict future values of the natural frequencies when only the predictors are measured.

The mentioned linear model is characterized by the equation:

$$\mathbf{Y} = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{Z}^{\mathbf{T}} + \mathbf{E} \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{n \times N}$ is the observation matrix with n rows containing the identified frequencies and N columns corresponding to the number of observations, $\mathbf{Z} \in$

$\mathbb{R}^{N\times(p+1)}$ is a matrix that contains a first column of ones and N values of the p independent variables in the remaining p columns, $\boldsymbol{\beta} \in \mathbb{R}^{(p+1)\times n}$ is a matrix with the parameters to be determined that weight the contribution of each independent variable, while $\mathbf{E} \in \mathbb{R}^{n\times N}$ contains the values of the random errors associated to the difference between the observed and fitted variables.

The main goal of the MLR approach is to estimate the coefficients contained in $\boldsymbol{\beta}$, which provide a good fit between the observations and the independent variables predicted by the statistical model. They can be obtained by exploiting the least squares method, whose task is to minimize the sum of the squares of the residuals. Following this approach, the modal frequencies independently estimated are computed as:

$$\hat{\mathbf{Y}} = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{Z}^{\mathbf{T}} \tag{2}$$

2.2. Principal Component Analysis (PCA)

It may sometimes occur that environmental variables are not measured but their effects are merely observed from the variation of the modal parameters. Therefore, in those cases where environmental factors are not known, Principal Component Analysis (PCA) is a suitable solution. The main idea is to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables, called principal components (PCs).

The original data are first projected into the vectorial space generated by the PCs and then moved back to the original space by retaining only some of the PCs. These statistically independent variables constitute an orthogonal basis and yield different contributions to the variance of the original data. The basic concept behind this theory is that the PCs that provide the largest contributions to the variance represent the independent environmental and operational parameters that have to be retained in order to estimate the matrix $\hat{\mathbf{Y}}$.

Operatively, the projection is based on the so-called loading matrix, $\mathbf{T} \in \mathbb{R}^{n\times n}$, computed as

$$\mathbf{T} = \mathbf{U}^{\mathbf{T}} \tag{3}$$

where matrix $\mathbf{U}$ comes from the singular value decomposition (SVD) of the covariance matrix of the original data computed in the training period, namely

$$\mathbf{Y}\mathbf{Y}^{\mathbf{T}} = \mathbf{U}\mathbf{S}^2\mathbf{U}^{\mathbf{T}} \tag{4}$$

It should be noticed that each row of the loading matrix contains the coefficients of a singular PC, while the singular values of the covariance matrix, contained in the diagonal matrix $\mathbf{S}^2$, represent the variance contribution of each PC.

It is possible to obtain a rectangular reduced loading matrix, $\hat{\mathbf{T}} \in \mathbb{R}^{l \times n}$, by considering only the first l columns of matrix $\mathbf{U}$ in Eq. (3), that is, by retaining only the first l PCs. Matrix $\hat{\mathbf{Y}}$ is therefore estimated as follows

$$\hat{\mathbf{Y}} = \hat{\mathbf{T}}^{\mathbf{T}}\hat{\mathbf{T}}\mathbf{Y} \tag{5}$$

which applies the transformation from the space of the PCs to the original one just to the first selected l PCs.

In this regard, the choice of the optimal number of PCs to be retained in the statistical model, represents a key aspect for the reliability of damage detection. In particular, it should be equal to the number of independent variables producing the largest contribution to the variance in the data and whose effects have to be removed. If this number is too small, part of environmental effects are not properly removed; on the contrary, if it is too large, additional effects, including damage, could be removed. A common rule usually consists in retaining the components which capture 70% - 90% of the variation.

2.3. Clustering-Gaussian Mixture Model (GMM)

Clustering tools are quite useful to group the damage sensitive features in the training period into different clusters, in order to take into account the presence of multiple environmental regimes, exhibiting non-linear environmental/operational effects. Gaussian Mixture Model (GMM) is a quite efficient tool for this purpose.

Once defined the m resonant frequencies f_i, $i = 1, ..., m$ as damage sensitive features, the vector $\mathbf{x}_n = [f_{(1,n)}, ..., f_{(m,n)}]$, containing the features at an instant n,

$n = 1, ..., N$ can be introduced. Then, a subset of t_p data samples is selected as the training period, with the aim to statistically describe the healthy state of the structure. Since the data set in the training period is non-normally distributed, it is useful to consider K clusters described by Gaussian distributions and to utilize a linear superposition of them in order to give a better characterization of the data set. Within this context, this approach assumes that the probability density function $p(\mathbf{x})$ of the data set in the training period $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_{t_p}\}$ can be represented as a linear superposition of K Gaussian components as:

$$p\left(\mathbf{x}\right) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{6}$$

which is called a mixture of Gaussians. Each component of the mixture is defined as a Gaussian density $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and has its own mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. The parameters $\boldsymbol{\pi} = [\pi_1, ..., \pi_K]$ are called mixing coefficients. They vary from 0 to 1 ($0 \leq \pi_k \leq 1$) and sum up to 1 ($\sum_{k=1}^{K} \pi_k = 1$). One way to set the model parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and π_k is to minimize the log-likelihood function:

$$\ln p\left(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \sum_{n=1}^{t_p} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{7}$$

where $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K]$ and $\boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_1, ...\boldsymbol{\Sigma}_K]$. The maximum likelihood estimate of the model parameters, namely $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$, is achieved by using the iterative Expectation-Maximization (EM) algorithm. In the expectation (E) step, after a prior initial guess of the parameters, the posterior probability that $\mathbf{x}_n$ is assigned to the *k-th* cluster is evaluated by means of the so-called responsibilities $\gamma(z_{nk})$:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{8}$$

where z_{nk} represents an element of a K-dimensional binary random variable $\mathbf{z}$ with the peculiarity that only one element in $\mathbf{z}$ is equal to 1 and all other

elements are 0.

The posterior probability calculated in the previous E step is afterwards used in the maximization (M) step to re-estimate the means, covariances and mixing coefficients, as indicated in the following equations:

$$\boldsymbol{\mu_k}^{new} = \frac{1}{N_k} \sum_{n=1}^{t_p} \gamma(z_{nk}) \mathbf{x}_n \tag{9}$$

$$\boldsymbol{\Sigma_k}^{new} = \frac{1}{N_k} \sum_{n=1}^{t_p} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu_k}^{new})(\mathbf{x}_n - \boldsymbol{\mu_k}^{new})^T \tag{10}$$

$$\pi_k^{new} = N_k/N, \quad N_k = \sum_{n=1}^{t_p} \gamma(z_{nk}) \tag{11}$$

where N_k can be interpreted as the effective number of points assigned to cluster k. Once the log-likelihood function in Eq. (7) is evaluated, this procedure is iterated, using the updated parameters, to achieve the convergence of the log-likelihood function.

As opposed to k-means, a GMM model is able to provide the probabilities that a given data point belongs to each of the possible clusters. However, on the other hand, it requires significantly more computational efforts, due to the several iterations needed to reach convergence. When new data are acquired, these are assigned to one of the previously obtained clusters K with minimal Mahalanobis distance, whose definition can be found in [49].

2.4. Novelty analysis

The proposed statistical model should be able to reproduce the part of variance in frequency estimates associated with changes in environmental conditions. Consequently, any damage pattern affects only data contained in $\mathbf{Y}$ but not those in $\hat{\mathbf{Y}}$. It follows the importance to take into account the residual error matrix $\mathbf{E}$, which can be straightforwardly calculated as:

$$\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}} \tag{12}$$

Under the assumption that the statistical model is properly defined, quantities in Eq. (12) are only minimally affected by environmental factors and therefore

suitable to be adopted as damage sensitive features. This translates into a conceivable detection of the damage by analyzing anomalies in the distribution of $\mathbf{E}$.

For this purpose, the classical statistical process tool, named Novelty Analysis, is adopted to track the evolution in time of the identified natural frequencies in order to detect any possible outlier, that is an observation which considerably deviates from the data population.

The first step consists in the definition of a reference condition, in which data sets are collected in a training period, so that the system can be able to compare any new data point with the healthy state. It follows that any significant deviation from normal conditions is associated with damage.

A common approach is the use of control charts, which are based on properly defined statistical distances. One of the most common is the T^2-statistic, defined as:

$$T^2 = r \cdot \left(\overline{\mathbf{E}} - \overline{\overline{\mathbf{E}}}\right)^T \cdot \sum^{-1} \cdot \left(\overline{\mathbf{E}} - \overline{\overline{\mathbf{E}}}\right) \tag{13}$$

where r is an integer parameter, named group averaging size, $\overline{\mathbf{E}}$ is the mean of the residuals computed in the subgroup of the last r observations, while $\overline{\overline{\mathbf{E}}}$ and $\sum$ are the mean value and the covariance matrix of the residuals statistically estimated in the training period, respectively.

Once a value of the statistical distance lies above the Upper Control Limit (UCL), it is considered as an outlier. The idea is to compute this threshold so that every portion of the control chart is analysed. However, instead of referring the UCL to the confidence level of the whole monitoring period (with damage included), it is useful to compute the limit based on the dependence on the standard deviation of the control chart in the training period. With this approach, whichever range of T^2 values can be covered. Otherwise, by defining UCL as a confidence level in the training period, the threshold can only assume values between 0 and the maximum statistical distance (T^2) in the training period.

Therefore, if a several number of dots steadily overcome the limit threshold (corresponding to a certain percentage of the standard deviation of the control chart

in the training period), a change in the statistical distribution of the residuals may have been occurred. Thus, damage-induced anomalies not encountered in the training period can be conceivably detected.

3. Methodology

The new approach developed in this work represents a decision-support system for the definition of the most suitable technique to remove environmental and operational effects as well as for the selection of the best damage threshold leading to a minimization of false alarms. This section aims at describing such a novel methodology, based on the combined use of ROC and PR curves, to evaluate and compare the performance of manifold statistical models.

3.1. ROC curves

A Receiver Operating Characteristic (ROC) curve represents a graphical tool which allows to quantify the performance of a process, varying the threshold definition, as well as to enable the statistical evaluation of the errors related to false detection.

Preliminarly, it should be taken into account that the control chart provides four possible outcomes, as reported in Fig. 1:

- if a single data set point prior to damage lies under the UCL, it is counted as True Negative (TN);
- if a single data set point prior to damage lies above the UCL, it is counted as False Positive (FP);
- if a single data set point after the damage lies under the UCL, it is counted as False Negative (FN);
- if a single data set point after the damage lies above the UCL, it is counted as True Positive (TP);

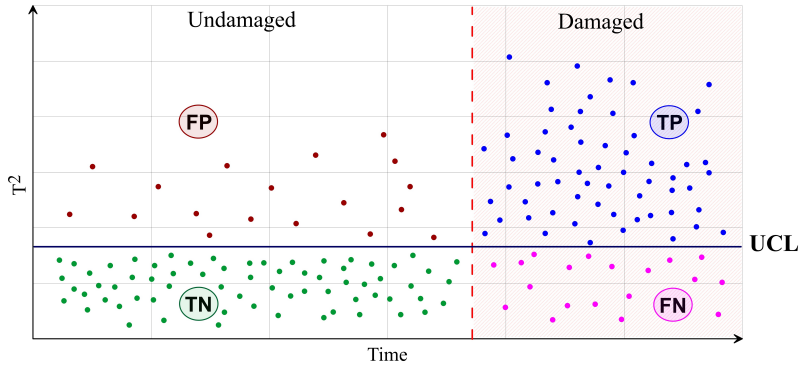

Figure 1: Control chart items' classification: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN).

Given TP, FP, TN and FN, the confusion matrix C can be constructed as follows:

$$C = \begin{bmatrix} \mathrm{TP} & \mathrm{FP} \\ \mathrm{FN} & \mathrm{TN} \end{bmatrix}; \tag{14}$$

Known all the quantities in C, some statistical measures of the performance of a binary classification test are introduced in order to characterize a ROC curve.

The true positive rate TPr, known as *sensitivity* (SE) or *probability of detection*, defines how many correct positive results occur among all positive samples:

$$\mathrm{TPr} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{\text{True Positives}}{\text{All positive cases}} \tag{15}$$

On the other hand the false positive rate FPr, known as *probability of false alarms*, describes how many incorrect positive results occur among all negative samples:

$$\mathrm{FPr} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{SP} = \frac{\text{False Positives}}{\text{All negative cases}} \tag{16}$$

where SP represents the so-called *specificity* or true nagative rate, counting how many correct negative results occur among all negatives samples.

A ROC curve is computed by plotting the true positive rate (TPr) against false positive rate (FPr) at various threshold positions. It follows that each point of the curve corresponds to the different classification thresholds varying from 0 to

$+\infty$. The basic idea is that a very high threshold could never indicate damage, resulting in 0% of false and true positives, whereas a very low threshold classifies more items as positive, producing 100% of false and true positives. It is worth pointing out that the closer the curve comes to the upper-left-hand corner of the ROC space, the more accurate is the model, while the closer the curve comes to the 45° diagonal in the ROC space, the less accurate is the model, as shown in Fig. 2 a).

A common method used to describe the behavior of a ROC curve is to calculate the Area Under the Curve (AUC), indicated in Fig. 2 b). When using normalized units, AUC values are between 0 and 1. Such a parameter is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one [43] and it can be defined as:

$$\mathrm{AUC} = \int_0^1 ROC(f)\mathrm{d}f \tag{17}$$

where f is the false positive rate (FPr), while $ROC(f)$ indicates the corresponding true positive rate (TPr).

3.2. Precision-Recall curves

PR curves have been often used, as an alternative to ROC curves, in Information Retrieval [50] as well as in Machine Learning for assessing binary classification models [42]. In particular, when dealing with highly skewed datasets with an interest in the minority class, they turned out to be a valid tool allowing to improve the statistical evaluation of an algorithm's performance [48]. In PR space, recall against precision is plotted at various threshold positions, as depicted in Fig. 2 d). Recall (RC) is defined the same as the true positive rate (TPr):

$$\mathrm{RC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{18}$$

whereas Precision (PC) measures the fraction of items classified as positive that are truly positives. Therefore, it describes how good a model is at predicting

the positive class. It is computed as follows:

$$\mathrm{PC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{19}$$

In PR space the closer the curve comes to the upper-right-hand corner, the more accurate is the model, while the closer the curve comes to a horizontal line at a low precision, the less accurate is the model, as shown in Fig. 2 c).

Similarly to the ROC curves, it is possible to assess the performance of a Precision-Recall curve by computing the AUC value, which goes from a minimum of 0 to a maximum of 1, as indicated in Fig. 2 d).

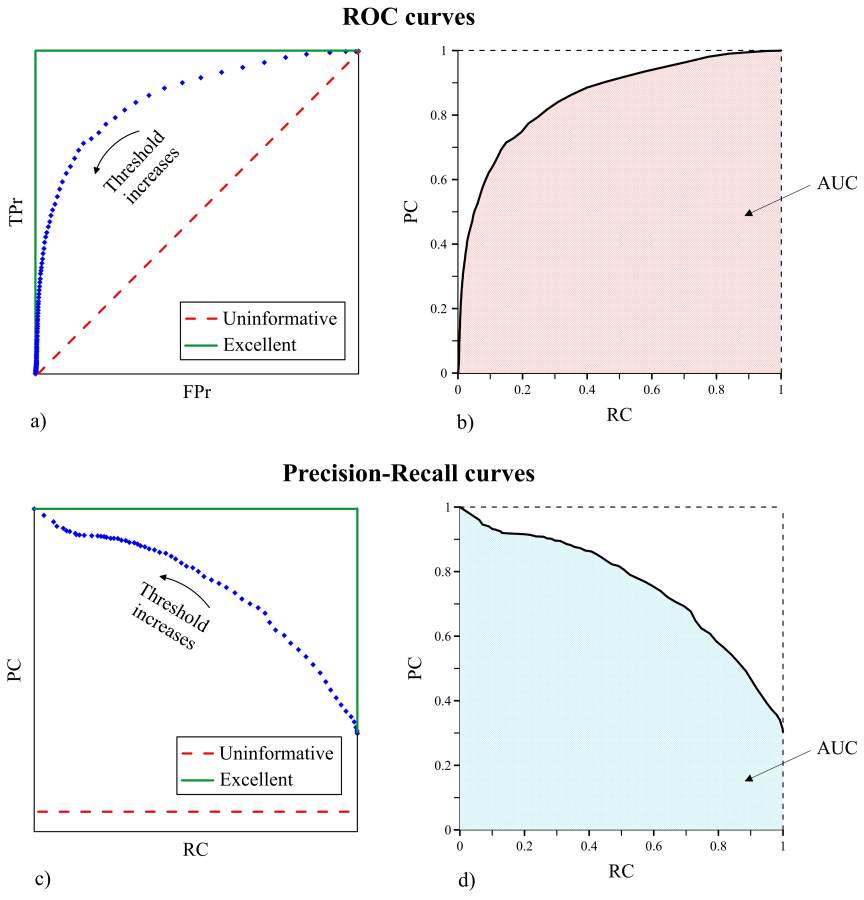

Figure 2: ROC curves: a) Different typologies: - Excellent; - Uninformative; b) Definition of AUC. PR curves: c) Different typologies: - Excellent; - Uninformative; d) Definition of AUC.

The main drawback of ROC curves comes to the light when there is a large skew in the class distribution (Fig. 3). In these situations, they can present an overly optimistic view of an algorithm's perfomance. Conversely, PR curves provide a more informative picture of an unbalanced dataset. If the number of negative samples greatly exceeds the number of positive samples, a large change in the number of false positives can lead to a small change in the false positive rate used in ROC analysis (Eq. (16)). This may translate into an optimistic ROC curve. On the other hand, precision in Eq. (19) is able to capture the effect of the large number of negative samples on the classifier's performance, by comparing false positives to true positives rather than true negatives. Therefore, the use of Precision-Recall is highly recommended when the focus is on the positive class.

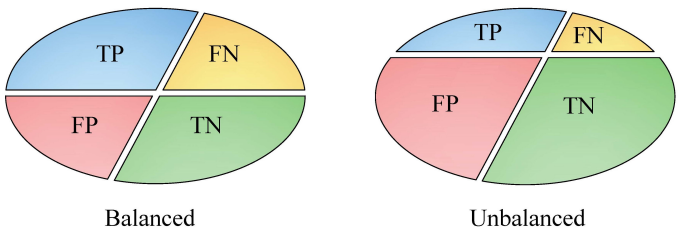

Figure 3: Examples of balanced and unbalanced data

3.3. Proposed approach for the selection of the best data cleansing technique

Numerous tools have been recently dedicated to estimating and removing environmental effects from monitoring data in order to discern damage-induced changes in the natural frequency time-histories. However, the selection of a proper statistical model is not a trivial issue, as it is higly correlated with the distribution of the residuals, the control chart and therefore with the damage detection. For this reason, a novel methodology is proposed in this paper, based on the combined use of ROC and PR curves. It is schematically presented in

Fig. 4 a) and it is detailed below.

Preliminarly, m statistical models are built with the purpose to compare their ability to predict the part of variance in frequency estimates that is associated with changes in environmental conditions. Moreover, d damage scenarios, real or simulated, are taken into account in order to find out how the performance of a statistical model changes, varying the type of damage.

The case of simulated damage scenarios is largely predominant in civil engineering applications. In those cases, a suitable structural model able to predict the variation of modal properties due to different types of damage is needed. Then, such damage-induced variations are artificially introduced into the monitoring data through constant shifts in the time series of identified modal frequencies. Both steps (model prediction of damage-induced effects and assumption of a constant-in-time damage-induced shift in modal frequencies) obviously come with errors and therefore produce approximations. Accounting for such errors is not in the aims of the present paper, where the proposed procedure is deemed effective at selecting the best performing statistical pattern recognition approach, in the understanding that the mentioned errors are sufficiently small (thanks to an accurate modeling) and that they affect all statistical techniques in a similar way.

Residuals are computed by using Eq. (12) and afterwards plotted in $(m \times d)$ control charts. By varying the UCL and by counting the data set points belonging to each class, labelled as FP, TP, FN and TN, $(m \times d)$ ROC and PR curves are obtained straightforwardly. It should be highlighted that each curve is referred to a particular statistical model, or data normalization technique, and to a specific damage scenario. Therefore, a useful metric already mentioned, namely AUC, is introduced in order to be able to make a conceivable judgement about the performance of different curves. According to the theory, the goal is to select the one with the higher AUC, corresponding to a model with an excellent capability in discerning between the positive and negative classes. The basic idea of the proposed methodology is to consider, as a metric performance, a linear combination between the area under the ROC curves and the

area under the PR curves, given by the objective function f_{ij}:

$$f_{ij} = \alpha \cdot AUC_{ij}^{ROC} + (1 - \alpha) \cdot AUC_{ij}^{PR} \tag{20}$$

$$i = 1, 2, ..., m \tag{21}$$

$$j = 1, 2, ..., d \tag{22}$$

where $\alpha \leq 1$ is a weight coefficient, while AUC_{ij}^{ROC} and AUC_{ij}^{PR} represent the areas under the ROC and PR curves, respectively, for the i^{th} statistical model and the j^{th} damage scenario.

Based upon the case study of interest, one could decide to apply a different weight to the two curves, varying the coefficient α.

Thus, in order to select the optimal model, the problem leads back to the maximization of the objective function f_{ij}, as follows:

$$i_{opt} = \arg\max(J_i) \tag{23}$$

$$J_i = \sum_{j=1}^{d} f_{ij} \tag{24}$$

It is worth asserting that a combined use of both curves allows to give a more informative picture of an algorithm's performance. Using a ROC curve with an imbalanced dataset might be deceptive and lead to incorrect interpretations of the model skill. On the other hand, PR curves evaluate the fraction of true positives among positive predictions and hence, they provide an accurate prediction of future classification performance.

3.4. Optimal threshold selection

Once the best performing model is defined, it comes to the light the necessity to select the optimal threshold leading to a minimization of false alarms and false negatives. Two helpful parameters are here proposed for this purpose. The first one is the *Youden index* (Y), representing the vertical distance between a point on the ROC curve and the 45° line. Typically, the higher the *Youden index*, the closer the ROC curve comes to the upper-left-hand corner, resulting

in a high percentage of true positive rate over false positive rate. It can be computed as follows:

$$Y = \mathrm{SE} + \mathrm{SP} - 1 \tag{25}$$

The second parameter is the *F1 score* ($F1$), which combines precision and recall into one metric by calculting the harmonic mean between those two. It reaches its best value at 1 and can be evaluated as:

$$F1 = 2\,\frac{\mathrm{PC} \cdot \mathrm{RC}}{\mathrm{PC} + \mathrm{RC}} \tag{26}$$

The use of both coefficients allows to provide a broader view about the model's performance.

The basic idea is to compute *Youden index* as well as *F1 score*, varying threshold positions, for each j^{th} damage scenario. The maximum values, namely $Y_{max}(j)$ and $F1_{max}(j)$, obtained for every damage case, correspond to a certain threshold, denoted as $T_Y(j)$ and $T_{F1}(j)$, respectively. Thus, in order to reach a unique value of the threshold $T(j)$, the mean between $T_Y(j)$ and $T_{F1}(j)$ is calculated:

$$T(j) = \frac{T_Y(j) + T_{F1}(j)}{2} \tag{27}$$

It should be noticed that this procedure leads to several thresholds, each one related to a specific damage scenario. For this reason, in order to let the system detect even the smallest damage, it is fair to assert that the optimal threshold could be fixed as the minimum among the $T(j)$, that is:

$$T_{opt} = \min\{T(j)\} \tag{28}$$

The developed procedure is schematically presented in Fig. 4 b).

4. Description of the two case studies

4.1. The Z24 Bridge

The Z24 Bridge was an overpass of the highway located between Bern and Zurich, linking the villages of Koppigen and Utzenstorf. It was a post-tensioned concrete box girder bridge with a main span of 30 m and two side spans of 14

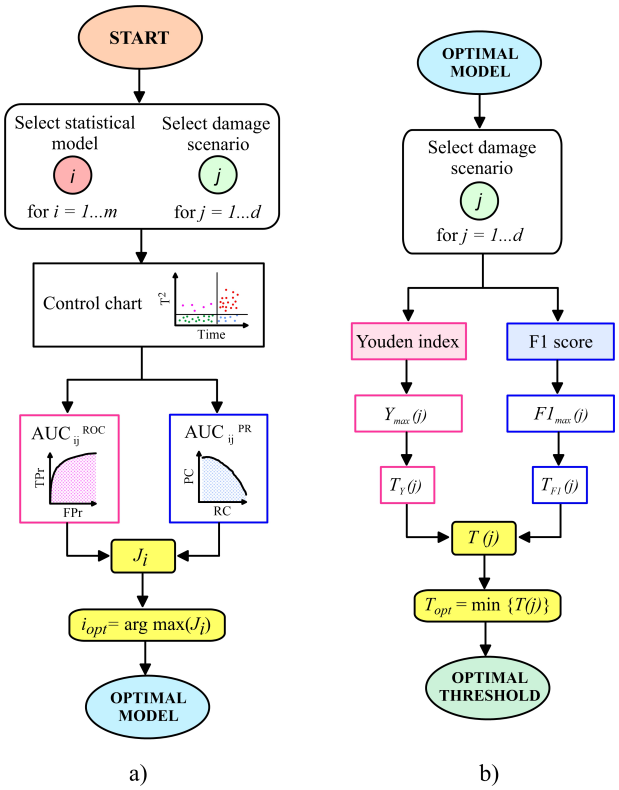

Figure 4: a) Flow chart of the proposed methodology; b) Flow chart of the methodology for the optimal threshold selection.

m, for a global length of 60 m and a width of 8.6 m. The intermediate supports were two concrete piers, clamped into the girders, situated at the end points of the main span (Fig. 5). Both abutments, consisting of triple concrete columns, were connected with concrete hinges to the girder.

The bridge, dated from 1963, was demolished at the end of 1998 to build a larger side span bridge, altough there were no known structural problems.

Before demolition, the Z24 bridge has been continuosly monitored from November 1997 till September 1998 with the aim to provide both environmental and vibration data. Therefore, sensors to measure accelerations as well as environmental parameters like air temperature, humidity, rain, wind speed and wind direction were installed in the bridge. The positions of the 16 accelerom-

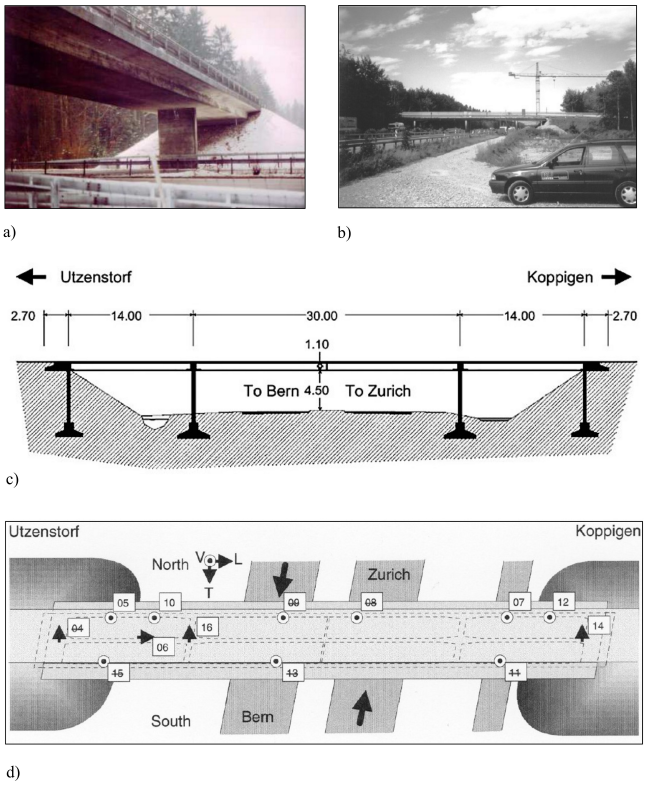

Figure 5: The Z24 Bridge: a,b) General view, c) Longitudinal section, d) Location and orientation of accelerometers. (Maeck et al. [51], Steenackers et al. [52]).

eters, among which only 8 are considered functioning, is visible on Fig. 5 d).

290 The measurements were carried out on an hourly basis, resulting in 24 measurements per day. In particular, environmental parameters were acquired before and after acceleration time histories.

Furthermore, in order to study the influence of different realistic damage scenarios on the dynamic properties, progressive damage tests took place during the

295 summer of 1998, shortly before complete demolition of the bridge. A detailed description of the fifteen induced damage scenarios as well as of the AVT tests results can be found in Kramer et al. [53] and Maeck et al. [51].

4.2. The Consoli Palace

The Consoli Palace (Fig. 6) is located in the historical center of Gubbio, a medieval town in the Central Italy. Considered the most representative monument of the town, Consoli Palace was designed by Angelo da Orvieto and Matteo Gattapone and built in gothic style between 1332 and 1349. The building has hosted the Civic Museum since 1909, with a rich collection af art masterpieces, while in the middle ages it hosted the Consuls who were elected to control both legislative and executive branches of the government.

The Palace is mainly made up of calcareous stone masonry and, in terms of geometry, it has a rectangular plan of about 40×20 m^2 and an elevation of more than 60 m. It is constituted by thick bearing walls and masonry vaults as horizontal elements. Due to the slope of the mountain, the building foundations are placed on two distinct levels with an elevation difference of approximately 10 m. The main façade of Consoli Palace is characterized by round arched windows in the upper part and merlons above, supported by ogival arches. Moreover, it overlooks towards East the central square of Gubbio, where the staircase entrance is positioned.

A long term mixed static-dynamic SHM system has been continuosly recording since July 2017. The SHM system is composed of three accelerometers, two crack meters, two temperature sensors and one data acquisition system with remote connection to a data analysis server located in the Laboratory of Structural Dynamics of University of Perugia. A detailed description of the building's geometry and the monitoring system is provided by Kita et al. [29]. Recently, the monitoring system has been updgraded to comprise a total of twelve accelerometers, four crack meters and four temperature sensors.

For the purpose of natural frequency identification and tracking, data recorded by three high sensitivity uniaxial piezoelectric accelerometers model PCB 393B12 (10 V/g sensitivity and ± 0.5 g measuring range) placed on the roof of the Palace are considered in this paper, as shown in Fig. 6. Accelerations are sampled at 100 Hz and stored in consecutive separate files containing 30 min recordings. Then, the recorded data are sent through the Internet to the remote server,

where they are processed through a specific MATLAB code, called MOSS [20], using each stored 30-minute-long recording file for automated modal analysis and anomaly detection.

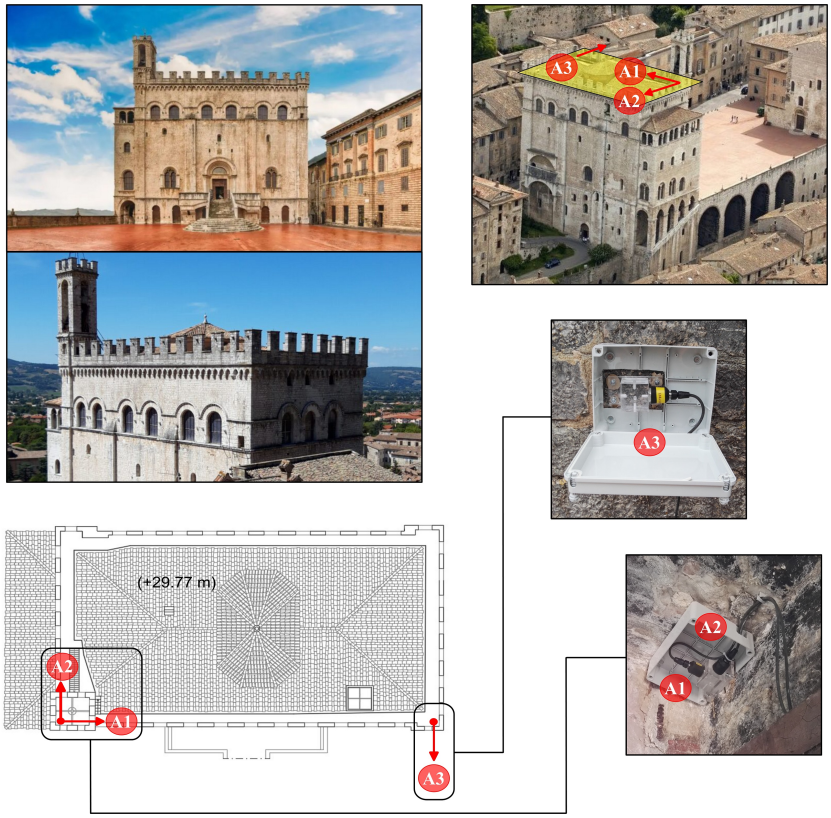

Figure 6: General view of the Consoli Palace. Location and orientation of the three accelerometers A1, A2, A3.

5. Validation of the procedure

Results obtained by application of the proposed methodology in the two selected case studies are presented in this section.

5.1. The Z24 Bridge : numerical results

5.1.1. Definition of the damage scenarios

Based on the continuous monitoring data, six natural frequencies have been identified by means of automated Stochastic Subspace Identification techniques operating on covariance functions (SSI-COV) [8]. The main parameters adopted in the analysis are reported in Table 1, where n and i indicate the model's order and the number of output block rows of the Hankel matrix, respectively.

Parameters	Adopted value
Maximum value of n	120
Minimum value of n	20
Maximum value of i	200
Minimum value of i	140
Step amplitude of n	2
Step amplitude of i	5
Frequency tolerance	0.01
Damping tolerance	0.03
MAC tolerance	0.01
Maximum reasonable damping	0.1
Threshold limit for clustering	0.03

Table 1: Adopted values for numerical parameters used in the automated system identification procedure (SSI-COV).

Then, frequency tracking has been carried out over time to have a broad view about the occurrence of any possible changes (Fig. 7). During the period between August 10^{th} and September 4^{th} 1998, the bridge was subjected to progressive damage tests. In this framework, in order to analyse different types of damage, the following four scenarios ($d = 4$) are taken into account:

- d_1 indicates the whole damage period (10^{th} August - 4^{th} September);
- d_2 indicates the first portion of damage period (10^{th} August - 18^{th} Au-

gust);

- d_3 indicates the second portion of damage period (19^{th} August - 26^{th} August);
- d_4 indicates the third portion of damage period (27^{th} August - 4^{th} September);

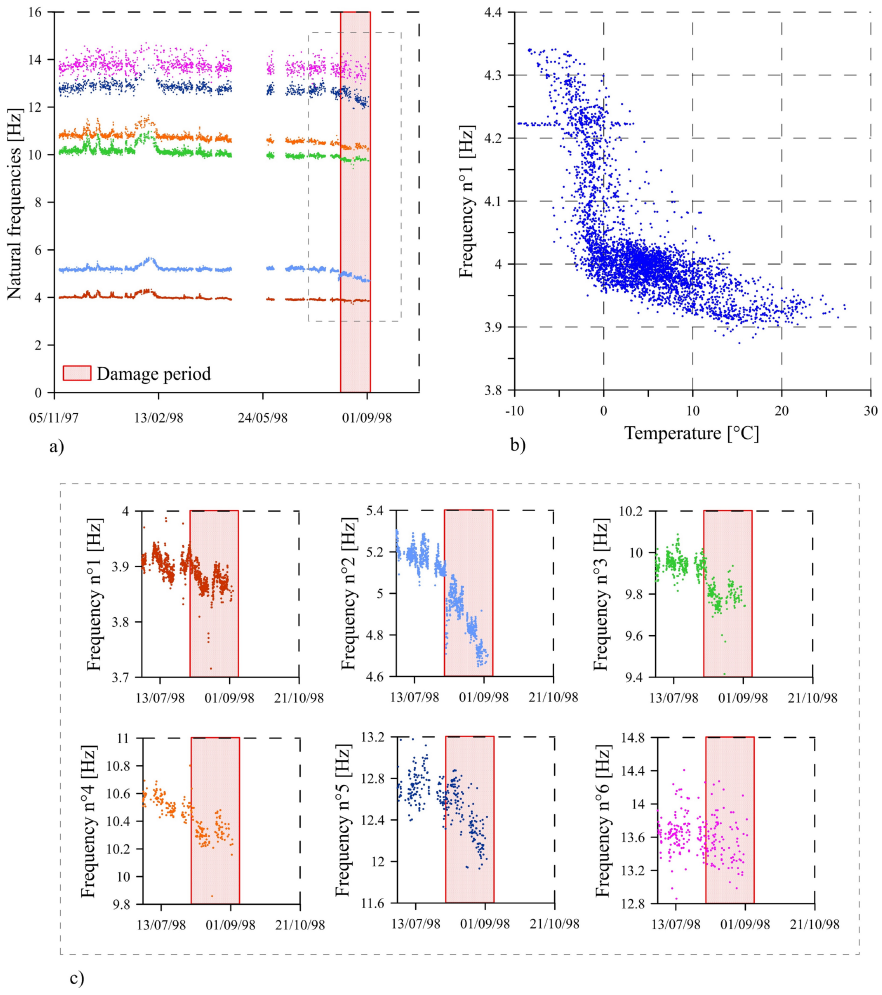

Figure 7: a) Frequency tracking of the Z24 Bridge; b) The relation between the first natural frequency and the temperature; c) Frequency tracking referring to the period close to damage.

5.1.2. Comparison between different statistical models

In order to estimate residuals, four statistical models ($m = 4$) have been adopted: Multiple Linear Regression (MLR), Principal Component Analysis (PCA), Local MLR and Local PCA. In detail, a local model is so called when MLR or PCA are carried out for every cluster identified in the training period by means of a GMM model.

By varying each damage scenario, all the statistical models are applied. As a result, residuals are plotted in different control charts, each one referred to the i^{th} model and the j^{th} damage condition. Following the developed methodology, ROC and PR curves are computed straightforwardly, by varying the threshold position and by counting false positives, true positives, false negatives and true negatives, outgoing from every control chart. As first step, AUC values of both curves are calculated to give a basic idea about models' quality, as reported in Table 2.

d	PCA		Local PCA		MLR		Local MLR	
	AUC_{ROC}	AUC_{PR}	AUC_{ROC}	AUC_{PR}	AUC_{ROC}	AUC_{PR}	AUC_{ROC}	AUC_{PR}
d_1	0.979	0.813	0.994	0.934	0.973	0.750	0.984	0.792
d_2	0.953	0.306	0.988	0.599	0.959	0.477	0.975	0.465
d_3	0.988	0.587	0.996	0.827	0.971	0.392	0.984	0.487
d_4	0.997	0.872	0.999	0.957	0.989	0.674	0.993	0.740

Table 2: AUC_{ROC} and AUC_{PR} values for m statistical models and d damage scenarios.

It is possible to notice that all the statistical models perform better when damage to be identified has a considerable severity (e.g. d_3 or d_4), because the probability to discern true positive items highly increases. Beyond this, the curves exhibit a good behavior even by analysing the whole damage period (d_1), as demonstrated in Fig. 8 a,b). On the contrary, when damage is not so marked (e.g. d_2), AUC_{ROC} and AUC_{PR} values get inevitably worse allowing, though, to clearly distinguish the different performance level associated to the m models (Fig. 8).

As a first result, regardless of damage scenario, Local PCA seems to be the

best technique to remove environmental effects, leading to extremely performing ROC and PR curves.

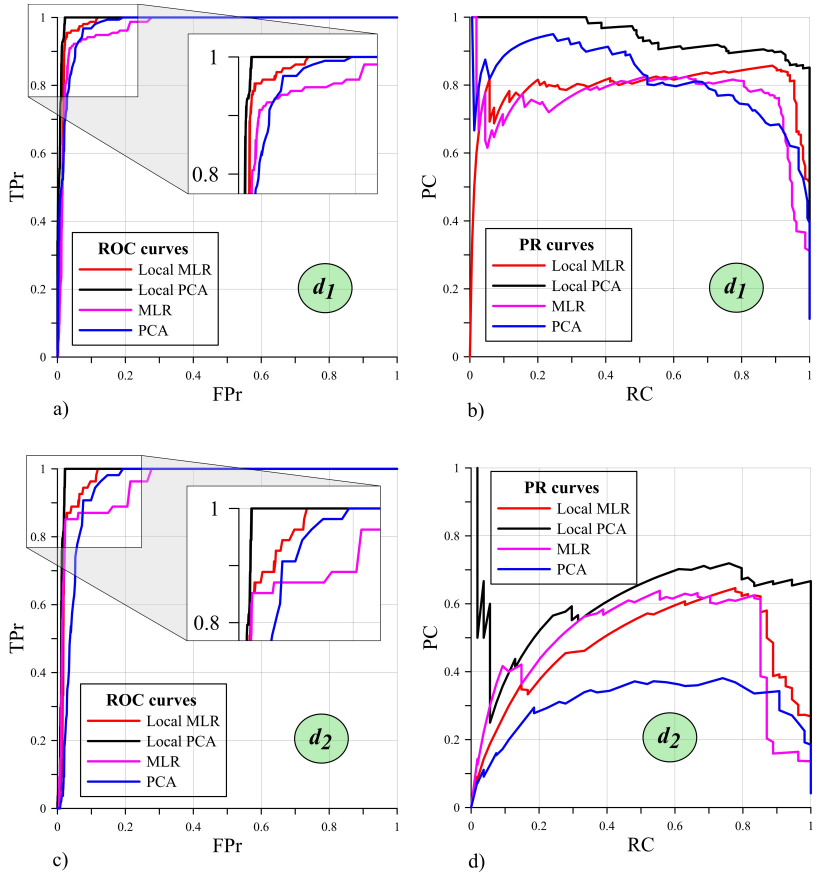

Figure 8: Comparison between the different statistical models in terms of ROC and PR curves. a, b) ROC and PR curves, respectively, computed for the whole damage period (d_1). c, d) ROC and PR curves, respectively, computed for the first portion of damage period (d_2).

It is worth highlighting that, since the damage period is very short, resulting in a large skew in the data distribution, the use of precision-recall curves is highly recommended. In such cases, this statistical tool, strictly focusing on the positive class, is able to provide a more realistic view (less optimistic) about the efficiency of the different models. In fact, by merely observing ROC curves, it could be difficult to discern any difference between all data normalization

techniques (which apparently yield excellent results). Conversely, PR curves allow to underline more visible differences between the models, as shown in Fig. 8 b,d). For this reason, a combined use of both curves should provide more realistic results.

Using Eq. (20) and keeping α fixed at 0.5, the objective function f_{ij} is computed straightforwardly, whose values are reported in Table 3.

d	PCA	Local PCA	MLR	Local MLR
	f	f	f	f
d_1	0.896	0.964	0.861	0.888
d_2	0.629	0.794	0.718	0.720
d_3	0.787	0.911	0.681	0.735
d_4	0.935	0.977	0.831	0.866

Table 3: The objective function f_{ij} computed for m statistical models and d damage scenarios.

The maximum value of the objective function i_{opt} (Eq. (23)) corresponds to the Local PCA model, which represents the best technique to remove environmental and operational effects or, in other words, the statistical model with the highest ability to correctly classify a certain outcome throughout the possible tresholds.

5.1.3. Selection of the optimal threshold

Once identifying Local PCA as the best performing model, the goal is to choose a threshold leading to the higher TPr over the FPr, in order to minimize false alarms and false negatives detection. Hence, *Youden index* and *F1 score* have been computed for every damage scenario, varying the threshold position. Their maximum values, Y_{max} and $F1_{max}$ respectively, are reported in Table 4 which shows, in addition, the corresponding thresholds, namely T_Y and T_{F1}, as well as the mean value T.

By comparing the second and the fourth column of Table 4, it can be noticed that the use of *Youden index* and *F1 score* provides similar results in terms

d	Local PCA				
	Y_{max}	T_Y	$F1_{max}$	T_{F1}	T
d_1	0.978	0.196	0.920	0.196	0.196
d_2	0.978	0.196	0.800	0.197	0.197
d_3	0.990	0.424	0.893	0.440	0.432
d_4	0.993	0.493	0.973	0.574	0.534

Table 4: The maximum values of *Youden index* (Y_{max}) and *F1 score* ($F1_{max}$) for every damage scenario and the corresponding thresholds (T_Y and T_{F1}, respectively).

of optimal threshold, even though this similarity seems to decrease with the growing of damage severity. Moreover, when damage is significant, it is possible to underline an increase of the optimal threshold, which is more remarkable in the case of T_{F1}. Fig. 9 highlights which is the trend of *Youden index* and *F1 score*, varying the threshold. In particular, it is noticeable how the interval containing the maximum values of the two indexes gets larger as the damage increases. Furthermore, by observing the Fig. 9 b,c), it is clear the relation between *F1 score* and the values of recall and precision.

However, this procedure has yielded several thresholds, each one related to a specific damage scenario. Thus, it is necessary to be able to provide a unique value, called as T_{opt}, useful to set the UCL of the control chart. In order to detect the smallest damage, translating in a more performing SHM system, the idea is to fix the optimal threshold as follows:

$$T_{opt} = \min\{T\} = 0.196 \tag{29}$$

The values of true positive rate (TPr), false positive rate (FPr), precision (PC) and specificity (SP) obtained by considering the best statistical model, that is Local PCA, and the optimal threshold in Eq. (29), are reported in Table 5, varying the type of damage.

Fig. 10 shows four control charts reflecting the performance of the different

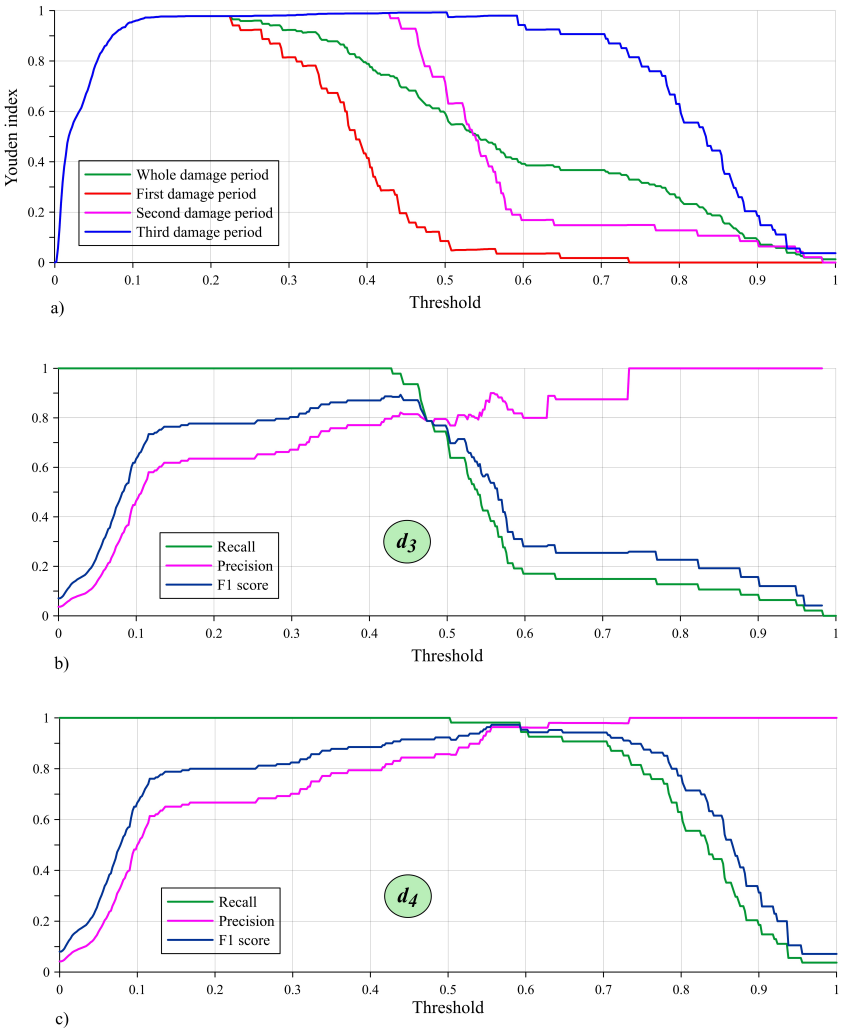

Figure 9: *Youden index* for d damage scenarios, varying the threshold (a) and the relation between recall, precision and *F1 score* for d_3 (b) and d_4 (c).

425 statistical models for one specific damage scenario: d_1. In particular, the optimal threshold for every model has been computed, following the steps in Fig. 4 b), and afterwards adopted to set each control chart. As a result, it is possible to notice that Local PCA represents the method which leads to a perfect classification of the outcomes.

d	TPr	FPr	PC	SP
d_1	1	0.022	0.852	0.978
d_2	1	0.022	0.667	0.978
d_3	1	0.022	0.635	0.978
d_4	1	0.022	0.667	0.978

Table 5: Values of TPr, FPr, PC and SP for every damage scenario d of the Z24 Bridge, fixing Local PCA as statistical model and T_{opt} as threshold.

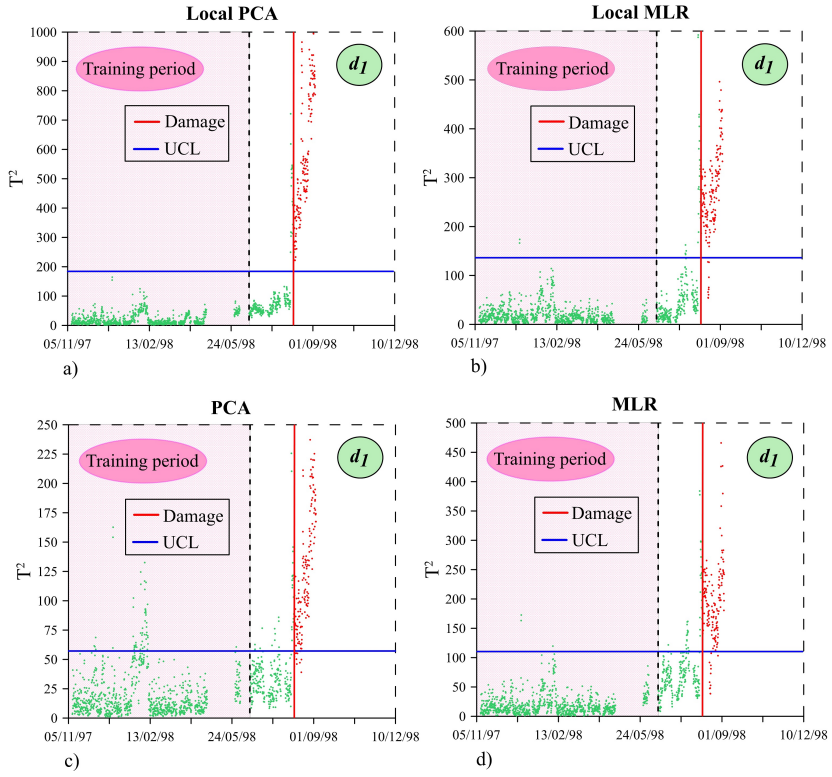

Figure 10: The control charts of the Z24 Bridge obtained for a specific damage scenario (d_1) by considering Local PCA (a), Local MLR (b), PCA (c) and MLR (d).

5.2. The Consoli Palace: numerical results

5.2.1. Definition of the damage scenarios

The natural frequencies of the Consoli Palace, shown in Table 7, have been identified with the SSI method, whose main parameters are reported in Table 6. Specifically, the model's order and the number of output block rows of the Hankel matrix are denoted with n and i, respectively.

Parameters	Adopted value
Maximum value of n	220
Minimum value of n	40
Maximum value of i	200
Minimum value of i	140
Step amplitude of n	4
Step amplitude of i	10
Frequency tolerance	0.01
Damping tolerance	0.03
MAC tolerance	0.01
Maximum reasonable damping	0.1
Threshold limit for clustering	0.03

Table 6: Adopted values for numerical parameters used in the automated system identification procedure (SSI-COV).

Then, all the frequencies have been tracked from July 2017 till August 2019, with the aim to observe any anomalies revealing the presence of damage (Fig. 11 a).

In order to study the structure behaviour for different types of damage, manifold simulations through a finite element model (FEM) have been carried out by applying specific frequencies decays (Δf). In particular, the FEM model has been already developed and calibrated by Kita et al. [29], who provided all the information needed for this paper's purposes.

In this context, two types of damage are taken into account (labelled with 1 and

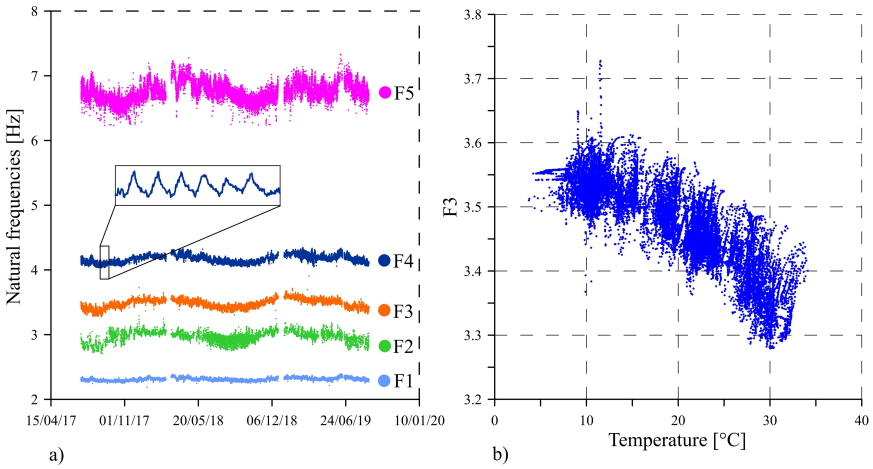

Figure 11: a) Frequency tracking of the Consoli Palace during the monitoring period; b) The relation between the third natural frequency and the temperature.

2), associated to different severity degrees (labelled with a and b), for a total number d of damage cases equal to 4 (Table 7):

- d_{1a} and d_{1b} refer to the seismic damage scenario carried on through pushover analysis along the main direction of the building (Fig. 12 a), where k_{1a} and k_{1b} are the multiplicative coefficient of the elastic modulus of the damaged elements;
- d_{2a} and d_{2b} refer to the bell tower collapse (Fig. 12 b), where k_{2a} and k_{2b} are the multiplicative coefficient of the elastic modulus of the bell tower;

5.2.2. Comparison between different statistical models

In order to estimate the variation of frequencies caused by environmental conditions, five statistical models ($m = 5$) have been compared: PCA, Local PCA, MLR-T (with temperature as predictor) and finally, MLR-T-C and Local MLR (with temperature and cracks as predictors).

By varying damage scenarios, each statistical model is applied in order to compute residuals, whose distribution is plotted, as a result, in different control

Freq. [Hz]	d_{1a}		d_{1b}		d_{2a}		d_{2b}	
	k_{1a}	$\Delta f(\%)$	k_{1b}	$\Delta f(\%)$	k_{2a}	$\Delta f(\%)$	k_{2b}	$\Delta f(\%)$
2.2956	0.5	2.71	0.8	0.8	0.9	0.23	0.94	0.1
2.9147	0.5	0.92	0.8	0.34	0.9	1.95	0.94	1.06
3.7253	0.5	-	0.8	-	0.9	-	0.94	-
4.0983	0.5	1.99	0.8	0.6	0.9	0.86	0.94	0.53
6.9209	0.5	-	0.8	-	0.9	-	0.94	-

Table 7: d damage scenarios simulated through a FEM model with the relative multiplicative coefficients of the elastic modulus and the associated frequencies decays.

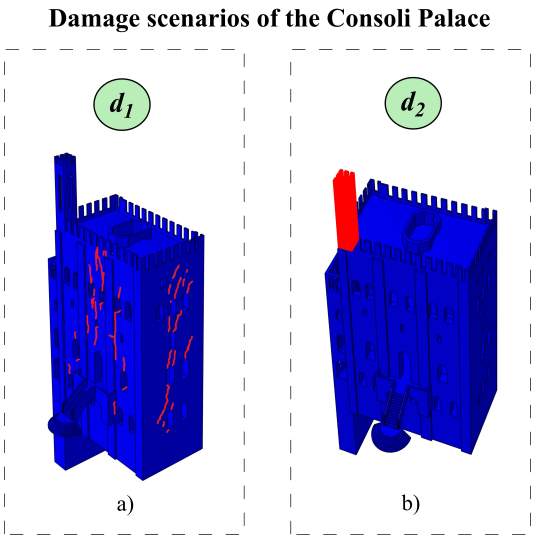

Figure 12: Simulated damage scenarios of the Consoli Palace.

charts. Following the developed methodology, ROC and PR curves are evaluated. Their ability to correctly classify the outcomes of the control chart comes up by observing the Table 8, reporting the AUC values of both curves.

Regarding the pushover analysis used to simulate the first damage scenario, it can be highligthed that low values of the damaged elements' elastic modulus

d	PCA		Local PCA		MLR-T		MLR-T-C		Local MLR	
	AUC_{ROC}	AUC_{PR}	AUC_{ROC}	AUC_{PR}	AUC_{ROC}	AUC_{PR}	AUC_{ROC}	AUC_{PR}	AUC_{ROC}	AUC_{PR}
d_{1a}	0.998	0.985	0.999	0.991	0.999	0.989	0.999	0.990	0.999	0.992
d_{1b}	0.774	0.645	0.856	0.743	0.865	0.743	0.901	0.802	0.939	0.853
d_{2a}	0.785	0.652	0.817	0.674	0.880	0.789	0.881	0.787	0.907	0.812
d_{2b}	0.636	0.473	0.675	0.502	0.765	0.621	0.785	0.643	0.829	0.682

Table 8: Comparison between m statistical models in terms of AUC_{ROC} and AUC_{PR} for every damage scenario.

cause a very important damage and hence, easy to identify. Indeed, the curves associated to all the statistical models show AUC_{ROC} and AUC_{PR} values close to 1, denoting a great capability in obtaining reliable results. On the other hand, if the severity of the first damage scenario decreases (e.g. d_{1b}), the differences between the models are more noticeable, as demonstrated by the Fig. 13 a), b). The different performance level of the involved techniques, varying the type of damage, can be observed in Fig. 13. As a first result, it is possible to underline that linear regression analysis, especially by considering temperature and cracks as predictors (Local MLR and MLR-T-C), leads to the best performing curves. This remark is common for both damage scenarios, even though the use of Local PCA is not recommendable in those data stemming from the simulation of the bell-tower damage (Fig. 13 c,d). Moreover, it is worth pointing out that, in this case study, ROC and PR curves provide the same information about the models' performance. This conclusion appears more clear by considering the first damage scenario (Fig. 13 a,b), while the second one shows PR curves which are less optimistic than ROC ones (Fig. 13 c,d).

Following the Eq. (20) and keeping α fixed at 0.5, the objective function f_{ij} can be evaluated straightforwardly, where i vary from 1 to 5 and j from 1 to 4. All the values are reported in Table 9.

Consequently, the maximization of the function through Eq. (23) leads to the selection of the best statistical model, that is Local MLR.

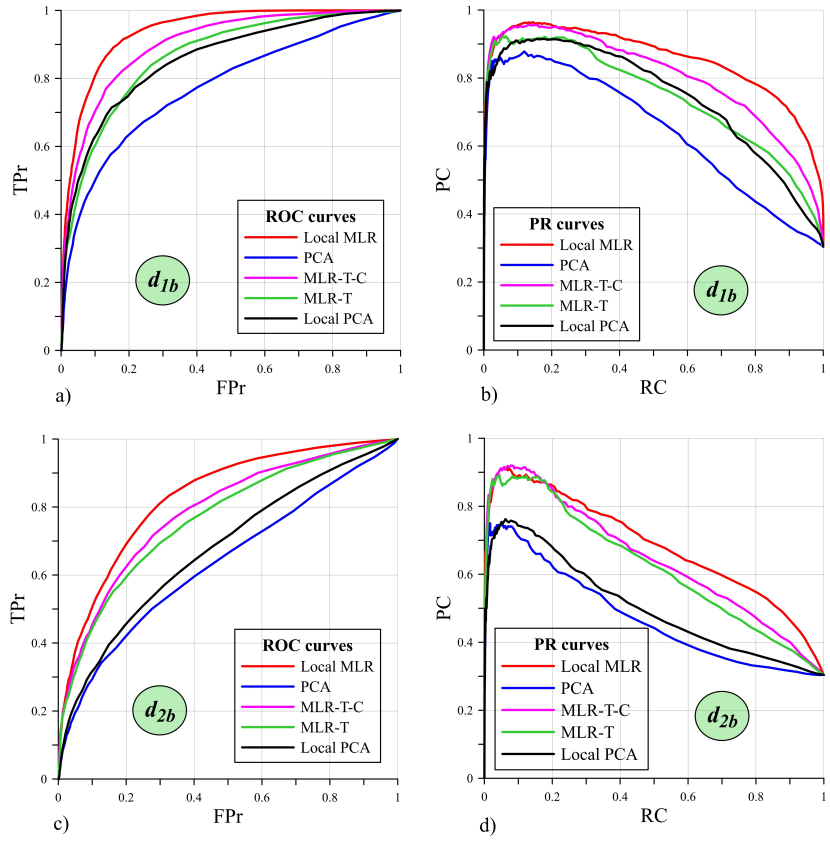

Figure 13: Comparison between m statistical models in terms of ROC and precision-recall curves for the damage scenarios d_{1b} (a,b) and d_{2b} (c,d).

d	f_{ij}				
	PCA	Local PCA	MLR-T	MLR-T-C	Local MLR
d_{1a}	0.992	0.995	0.994	0.995	0.996
d_{1b}	0.710	0.800	0.804	0.852	0.896
d_{2a}	0.719	0.746	0.835	0.834	0.860
d_{2b}	0.555	0.589	0.693	0.714	0.756

Table 9: The objective function f_{ij} computed for d damage scenarios and m statistical models.

5.2.3. Selection of the optimal threshold

Once identifying Local MLR as the best performing model, it is necessary to choose a threshold leading to the minimization of false alarms detection. Thus, *Youden index* and *F1 score* have been computed for every damage scenario, varying threshold position. Table 10 shows their maximum values, Y_{max} and $F1_{max}$ respectively, the corresponding thresholds, namely T_Y and T_{F1} and, in the last column, the mean value T.

d	Local MLR				
	Y_{max}	T_Y	$F1_{max}$	T_{F1}	T
d_{1a}	0.995	0.050	0.996	0.057	0.054
d_{1b}	0.737	0.016	0.798	0.018	0.017
d_{2a}	0.672	0.012	0.755	0.012	0.012
d_{2b}	0.512	0.009	0.651	0.009	0.009

Table 10: The maximum values of *Youden index* (Y_{max}) and *F1 score* ($F1_{max}$) for every damage scenario, the corresponding thresholds (T_Y and T_{F1}, respectively) and the mean value (T).

Focusing on the second and the fourth columns of Table 10, it can be noticed that, for the second damage scenario d_2, the use of *Youden index* and *F1 score* provides the same results in terms of optimal threshold, regardless of the severity degree. On the other hand, slight differences appear between the values of T_Y and T_{F1} when the first damage scenario d_1 is considered. Overall, by observing the mean value T in the last column, it is clear that the general rule is respected, since the optimal threshold increases when damage becomes significant. A graphical demonstration is given by the Fig. 14, representing how the values of *Youden index* and *F1 score* change, varying the threshold position.

With this procedure, several thresholds have been provided, each one associated to a specific damage scenario. Hence, it is important to yield a unique

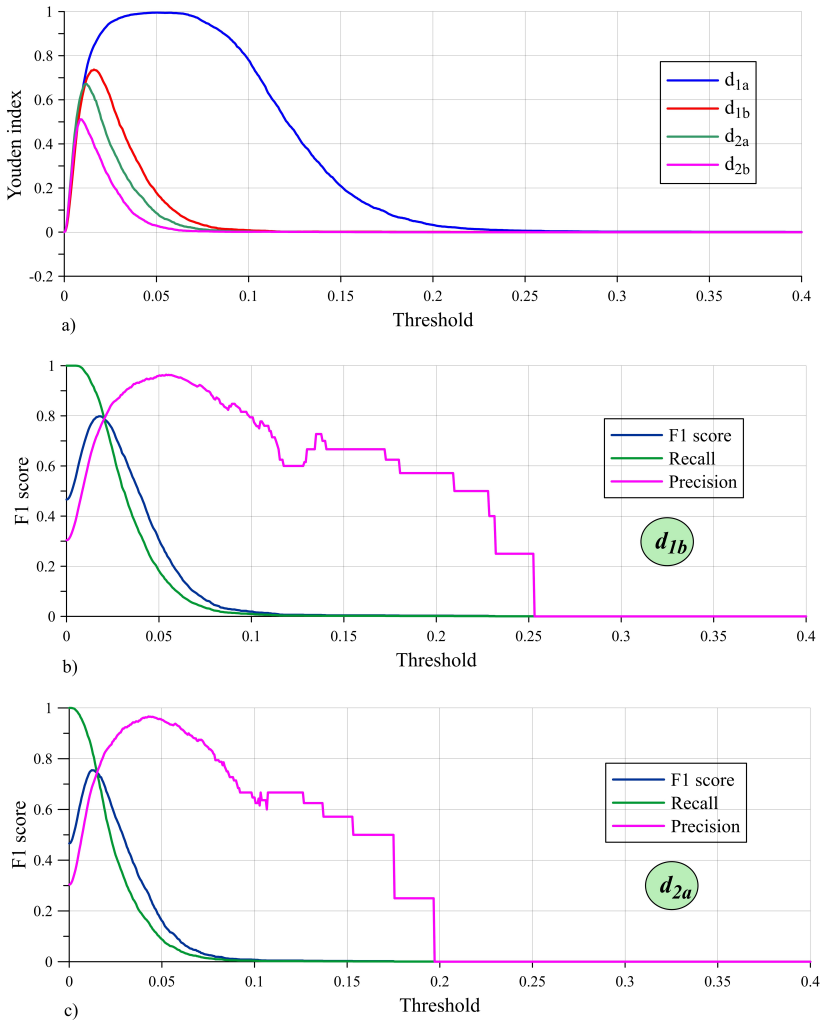

Figure 14: *Youden index* for d damage scenarios, varying the threshold (a) and the relation between recall, precision and *F1 score* for d_{1b} (b) and d_{2a} (c).

505 value, called as T_{opt}, to set the UCL of the control chart. As already mentioned in the previous case study, the goal of the SHM system should be the detection of the smallest damage, thus the optimal threshold can be fixed as follows:

$$T_{opt} = \min\{T\} = 0.009 \tag{30}$$

The values of true positive rate (TPr), false positive rate (FPr), precision

(PC) and specificity (SP) stemming from the selection of the best statistical model, or Local MLR, and the optimal threshold in Eq. (30), are presented in Table 11, varying the type of damage.

d	TPr	FPr	PC	SP
d_{1a}	1	0.362	0.522	0.638
d_{1b}	0.985	0.400	0.518	0.600
d_{2a}	0.906	0.268	0.596	0.733
d_{2b}	0.778	0.268	0.560	0.733

Table 11: Values of TPr, FPr, PC and SP for every damage scenario d, fixing Local MLR as statistical model and T_{opt} as threshold.

Fig. 15 shows four control charts describing the performance of different statistical models, namely Local PCA, PCA, Local MLR and MLR with temperature and cracks as predictors, for one specific damage scenario: d_{1b}. In particular, following the step in Fig. 4 b), the optimal threshold for every model has been computed and utilized to set each control chart.

5.3. Discussion of the results

The developed procedure to find out the best technique to remove environmental and operational effects has produced different results in the two case studies of interest. The reason lies in the type of correlation existing between the identified natural frequencies and the temperature. Regarding the Z24 Bridge, such relation is not linear, as shown in Fig. 7 b). Hence, Principal Component Analysis applied to single clusters (Local PCA) reveals a high capability in minimizing false alarms and false negatives, leading, therefore, to the best classification of the outcomes.

On the contrary, the natural frequencies of the Consoli Palace exhibit a quite linear correlation with temperature, as depicted in Fig. 11 b). Thus, it is reasonable to assert that linear regression models (MLR) perform well in such cases, even though local approaches (with temperature and cracks as predictors) are

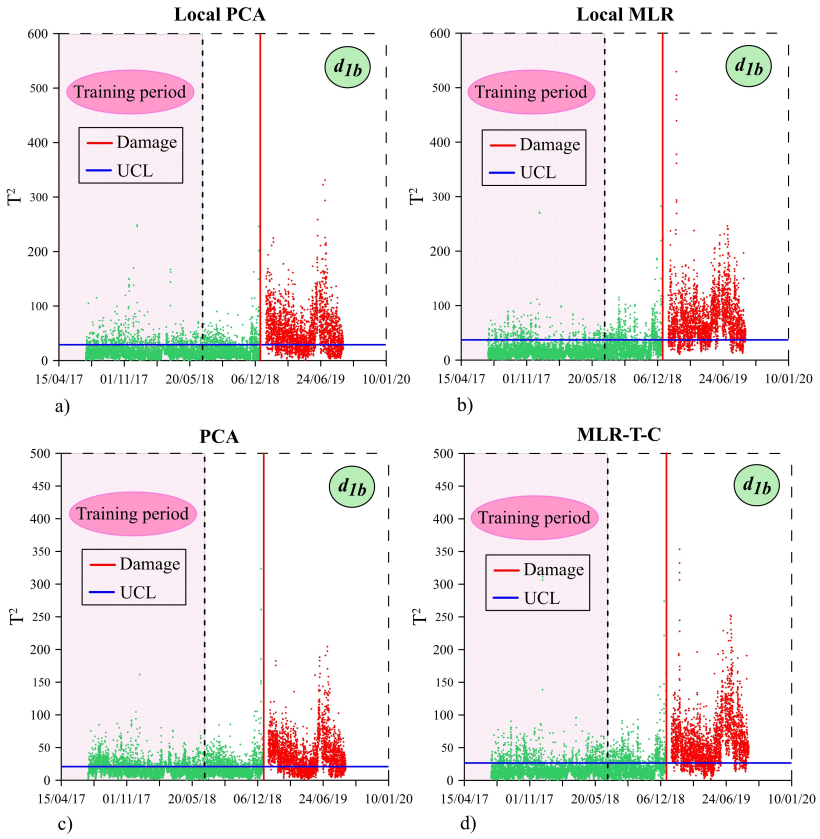

Figure 15: The control charts of the Consoli Palace obtained for a specific damage scenario d_{1b} by considering Local PCA (a), Local MLR (b), PCA (c) and MLR-T-C (d).

able to provide the best results in terms of ROC and PR curves, according to the proposed procedure.

Fig. 16 shows the values of the objective function, varying the weight coefficient α in the Eq. (20), for the damage scenario d_1 of the Z24 Bridge (a) and for the damage scenario d_{1b} of the Consoli Palace (b). In both cases, the comparison between the different statistical models allows to underline that a specific technique (Local PCA and Local MLR, respectively) seems to be better than the other ones, throughout the values of α. In this context, it is worth pointing out the importance to introduce PR curves in the first case study for the model's perfomance evaluation (Fig. 16 a). Indeed, for $\alpha = 1$, there is a clear difficulty

in discerning any difference between the statistical models, due to the fact that ROC curves gives a too optimistic view if data set is strongly unbalanced. On the contrary, fixing $\alpha = 0$, the procedure allows to detect, in a more intelligible manner, which technique leads to the maximization of the objective function.

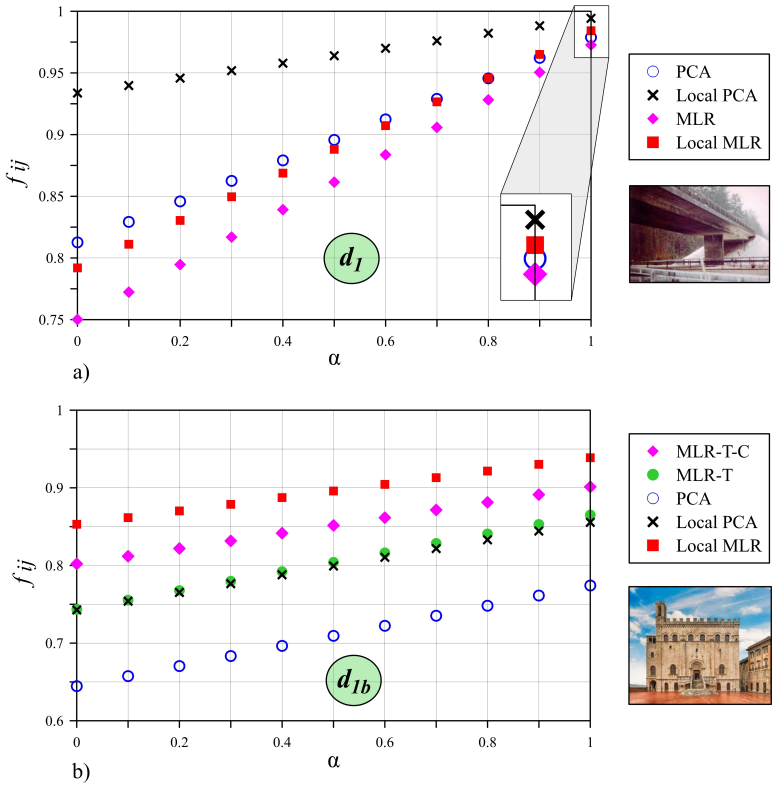

Figure 16: The values of the objective function, varying the weight coefficient α, for the damage scenario d_1 of the Z24 Bridge (a) and for the damage scenario d_{1b} of the Consoli Palace (b).

6. Conclusions

This paper has presented a general methodology with the aim to provide a new decision-support tool for the definition of the best technique to remove environmental effects, as well as for the selection of the optimal threshold leading

to the minimization of false alarms and false negatives detection. The procedure exploits the combination between two statistical tools, namely ROC and PR curves, which are computed and compared for a variety of statistical models and different damage scenarios. In particular, damage can stem from real data or from non linear FEM simulations, especially in those cases where damaged data are not available due to high costs or to practical constraints.

The use of precision-recall curves turned out to be particular meaningfull when a large skew in the class distribution is present. Indeed, primarily focusing on the positive class, they are able to provide a more realistic view of an algorithm's performance. Through the maximization of an objective function, involving both curves' effects, the optimal statistical model is straightforwardly identified.

Then, regarding the selection of the threshold, two coefficients, that are *Youden index* and *F1 score*, have been adopted. After assigning a certain threshold for every damage scenario of interest, the proposed approach adopts the minimum value among threshold values, so that the SHM system is able to detect even the smallest damage.

In order to illustrate the proposed methodology, two case studies have been analysed. With regards to the Z-24 bridge, whose data set is strongly imbalanced, precision-recall curves have been yielded to less optimistic results in comparison with ROC curves, helping to find out the best model in a more reasonable way. However, since damage inferred to the bridge (by means of progressive tests) was significant, all the statistical models managed to clearly detect it. Local PCA, though, showed a higher capability to correctly discern the outocomes of the control chart, that's why it has been considered as the optimal technique to remove environmental effects.

The second case study refers to a medieval masonry building, namely Consoli Palace. Two types of damage scenarios have been taken into account with different severity degrees. The first one stems from pushover analysis, while the second one is associated to a damage in the bell tower. ROC and PR curves, in this case, seem to provide the same information about model's performance.

Moreover, the linear regression analysis leads to more appreciable results in comparison with principal component analysis. Specifically, the use of Local MLR with temperature and cracks as predictors is associated to the best performing curves.

Hence, the proposed methodology represents a valid tool to statistically evaluate any model, by providing reliable performance metrics in unsupervised processes. The procedure is general and can be easily implemented considering different statistical models and different damage scenarios.

Acknowledgements

This work was supported by the Italian Ministry of Education, University and Research (MIUR) through the funded Project of Relevant National Interest "DETECT-AGING - Degradation effects on structural safety of cultural heritage constructions through simulation and health monitoring" (protocol no. 201747Y73L).

References

[1] F. Ubertini, G. Comanducci, N. Cavalagli, Vibration-based structural health monitoring of a historic bell-tower using output-only measurements and multivariate statistical analysis, Structural Health Monitoring 15 (4) (2016) 438–457. doi:10.1177/1475921716643948.

[2] L. Ramos, L. Marques, P. Lourenço, G. De Roeck, A. Campos-Costa, J. Roque, Monitoring historical masonry structures with operational modal analysis: Two case studies, Mechanical Systems and Signal Processing 24 (5) (2010) 1291–1305. doi:10.1016/j.ymssp.2010.01.011.

[3] M.-G. Masciotta, L. Ramos, P. Lourenço, The importance of structural monitoring as a diagnosis and control tool in the restoration process of heritage structures: A case study in portugal, Journal of Cultural Heritage 27 (2017) 36–47. doi:10.1016/j.culher.2017.04.003.

[4] F. Ubertini, N. Cavalagli, A. Kita, G. Comanducci, Assessment of a monumental masonry bell-tower after 2016 central italy seismic sequence by long-term shm, Bulletin of Earthquake Engineering 16 (2) (2018) 775–801. doi:10.1007/s10518-017-0222-7.

[5] A. Saisi, C. Gentile, M. Guidobaldi, Post-earthquake continuous dynamic monitoring of the gabbia tower in mantua, italy, Construction and Building Materials 81 (2015) 101–112. doi:10.1016/j.conbuildmat.2015.02.010.

[6] K. Gkoumas, F. Marques Dos Santos, M. van Balen, A. Tsakalidis, A. Ortega Hortelano, M. Grosso, F. Pekár, Research and innovation in bridge maintenance, inspection and monitoring (no. jrc115319).

[7] J. Ko, Y. Ni, Technology developments in structural health monitoring of large-scale bridges, Engineering Structures 27 (12 SPEC. ISS.) (2005) 1715–1725. doi:10.1016/j.engstruct.2005.02.021.

[8] F. Ubertini, C. Gentile, A. Materazzi, Automated modal identification in operational conditions and its application to bridges, Engineering Structures 46 (2013) 264–278. doi:10.1016/j.engstruct.2012.07.031.

[9] B. Costa, F. Magalhães, T. Cunha, J. Figueiras, Rehabilitation assessment of a centenary steel bridge based on modal analysis, Engineering Structures 56 (2013) 260–272. doi:10.1016/j.engstruct.2013.05.010.

[10] M. Vagnoli, R. Remenyte-Prescott, J. Andrews, Railway bridge structural health monitoring and fault detection: State-of-the-art methods and future challenges, Structural Health Monitoring 17 (4) (2018) 971–1007. doi:10.1177/1475921717721137.

[11] L. Sun, Z. Shang, Y. Xia, S. Bhowmick, S. Nagarajaiah, Review of bridge structural health monitoring aided by big data and artificial intelligence: From condition assessment to damage detection, Journal of Structural Engineering (United States) 146 (5). doi:10.1061/(ASCE)ST.1943-541X.0002535.

[12] C.-P. Fritzen, Vibration-based structural health monitoring - concepts and applications, Key Engineering Materials 293–294 (2005) 3–18. doi:10.4028/0-87849-976-8.3.

[13] J. Brownjohn, A. de Stefano, Y.-L. Xu, H. Wenzel, A. Aktan, Vibration-based monitoring of civil infrastructure: Challenges and successes, Journal of Civil Structural Health Monitoring 1 (3–4) (2011) 79–95. doi:10.1007/s13349-011-0009-5.

[14] O. Avci, O. Abdeljaber, S. Kiranyaz, M. Hussein, M. Gabbouj, D. Inman, A review of vibration-based damage detection in civil structures: From traditional methods to machine learning and deep learning applications, Mechanical Systems and Signal Processing 147. doi:10.1016/j.ymssp.2020.107077.

[15] A. Deraemaeker, E. Reynders, G. De Roeck, J. Kullaa, Vibration-based structural health monitoring using output-only measurements under changing environment, Mechanical Systems and Signal Processing 22 (1) (2008) 34–56. doi:10.1016/j.ymssp.2007.07.004.

[16] F. Magalhães, A. Cunha, Explaining operational modal analysis with data from an arch bridge, Mechanical Systems and Signal Processing 25 (5) (2011) 1431–1450. doi:10.1016/j.ymssp.2010.08.001.

[17] W. Fan, P. Qiao, Vibration-based damage identification methods: A review and comparative study, Structural Health Monitoring 10 (1) (2011) 83–111. doi:10.1177/1475921710365419.

[18] E. Reynders, System identification methods for (operational) modal analysis: Review and comparison, Archives of Computational Methods in Engineering 19 (1) (2012) 51–124. doi:10.1007/s11831-012-9069-x.

[19] A. Alvandi, C. Cremona, Assessment of vibration-based damage identification techniques, Journal of Sound and Vibration 292 (1–2) (2006) 179–202. doi:10.1016/j.jsv.2005.07.036.

[20] E. García-Macías, F. Ubertini, Mova/moss: Two integrated software solutions for comprehensive structural health monitoring of structures, Mechanical Systems and Signal Processing 143. doi:10.1016/j.ymssp.2020.106830.

[21] I. Venanzi, A. Kita, N. Cavalagli, L. Ierimonti, F. Ubertini, Continuous oma for damage detection and localization in the sciri tower in perugia, italy, 8th IOMAC - International Operational Modal Analysis Conference, Proceedings (2019) 127–136.

[22] I. Venanzi, A. Kita, N. Cavalagli, L. Ierimonti, F. Ubertini, Earthquake-induced damage localization in an historic masonry tower through long-term dynamic monitoring and fe model calibration, Bulletin of Earthquake Engineering 18 (5) (2020) 2247–2274. doi:10.1007/s10518-019-00780-4.

[23] B. Peeters, G. De Roeck, One year monitoring of the z24-bridge: Environmental influences versus damage events, Proceedings of the International Modal Analysis Conference - IMAC 2 (2000) 1570–1576. doi:10.1007/s10518-019-00780-4.

[24] Y. Xia, B. Chen, S. Weng, Y.-Q. Ni, Y.-L. Xu, Temperature effect on vibration properties of civil structures: A literature review and case studies, Journal of Civil Structural Health Monitoring 2 (1) (2012) 29–46. doi:10.1007/s13349-011-0015-7.

[25] P. Cornwell, C. Farrar, S. Doebling, H. Sohn, Environmental variability of modal properties, Experimental Techniques 23 (6) (1999) 45–48. doi:10.1111/j.1747-1567.1999.tb01320.x.

[26] R. G. Rohrmann, M. Baessler, S. Said, W. Schmid, W. F. Ruecker, Structural causes of temperature affected modal data of civil structures obtained by long time monitoring, Proceedings of the International Modal Analysis Conference - IMAC 1 (2000) 1–7.

[27] O. Salawu, Detection of structural damage through changes in frequency: A review, Engineering Structures 19 (9) (1997) 718–723. doi:10.1016/S0141-0296(96)00149-6.

[28] Y. Xu, B. Chen, C. Ng, K. Wong, W. Chan, Monitoring temperature effect on a long suspension bridge, Structural Control and Health Monitoring 17 (6) (2010) 632–653. doi:10.1002/stc.340.

[29] A. Kita, N. Cavalagli, F. Ubertini, Temperature effects on static and dynamic behavior of consoli palace in gubbio, italy, Mechanical Systems and Signal Processing 120 (2019) 180–202. doi:10.1016/j.ymssp.2018.10.021.

[30] E. García-Macías, I. Venanzi, F. Ubertini, Metamodel-based pattern recognition approach for real-time identification of earthquake-induced damage in historic masonry structures, Automation in Construction 120. doi:10.1016/j.autcon.2020.103389.

[31] W.-H. Hu, D.-H. Tang, J. Teng, S. Said, R. Rohrmann, Structural health monitoring of a prestressed concrete bridge based on statistical pattern recognition of continuous dynamic measurements over 14 years, Sensors (Basel, Switzerland) 18 (12). doi:10.3390/s18124117.

[32] F. Magalhães, A. Cunha, E. Caetano, Vibration based structural health monitoring of an arch bridge: From automated oma to damage detection, Mechanical Systems and Signal Processing 28 (2012) 212–228. doi:10.1016/j.ymssp.2011.06.011.

[33] N. Kambhatla, T. Leen, Dimension reduction by local principal component analysis, Neural Computation 9 (7) (1997) 1493–1516. doi:10.1162/neco.1997.9.7.1493.

[34] A.-M. Yan, G. Kerschen, P. De Boe, J.-C. Golinval, Structural damage diagnosis under varying environmental conditions - part i: A linear analysis,

Mechanical Systems and Signal Processing 19 (4) (2005) 847–864. doi:10.1016/j.ymssp.2004.12.002.

[35] A. Bellino, A. Fasana, L. Garibaldi, S. Marchesiello, Pca-based detection of damage in time-varying systems, Mechanical Systems and Signal Processing 24 (7) (2010) 2250–2260. doi:10.1016/j.ymssp.2010.04.009.

[36] C. Hanley, D. Kelliher, V. Pakrashi, Principal component analysis for condition monitoring of a network of bridge structures, Journal of Physics: Conference Series 628 (1). doi:10.1088/1742-6596/628/1/012060.

[37] G. Comanducci, F. Magalhães, F. Ubertini, . Cunha, On vibration-based damage detection by multivariate statistical techniques: Application to a long-span arch bridge, Structural Health Monitoring 15 (5) (2016) 505–524. doi:10.1177/1475921716650630.

[38] H. Sohn, M. Dzwonczyk, E. Straser, A. Kiremidjian, K. Law, T. Meng, An experimental study of temperature effect on modal parameters of the alamosa canyon bridge, Earthquake Engineering and Structural Dynamics 28 (7–8) (1999) 879–897. doi:10.1002/(sici)1096-9845(199908)28:8<879::aid-eqe845>3.0.co;2-v.

[39] A.-M. Yan, G. Kerschen, P. De Boe, J.-C. Golinval, Structural damage diagnosis under varying environmental conditions - part ii: Local pca for non-linear cases, Mechanical Systems and Signal Processing 19 (4) (2005) 865–880. doi:10.1016/j.ymssp.2004.12.003.

[40] K. Worden, H. Sohn, C. Farrar, Novelty detection in a changing environment: Regression and inter polation approaches, Journal of Sound and Vibration 258 (4) (2002) 741–761. doi:10.1006/jsvi.2002.5148.

[41] C. Calì, M. Longobardi, Some mathematical properties of the roc curve and their applications, Ricerche di Matematica 64 (2) (2015) 391–402. doi:10.1007/s11587-015-0246-8.

[42] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning 2006 (2006) 233–240.

[43] T. Fawcett, An introduction to roc analysis, Pattern Recognition Letters 27 (8) (2006) 861–874. doi:10.1016/j.patrec.2005.10.010.

[44] J. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, Journal of Thoracic Oncology 5 (9) (2010) 1315–1316. doi:10.1097/JTO.0b013e3181ec173d.

[45] A. Neves, I. González, J. Leander, R. Karoumi, A new approach to damage detection in bridges using machine learning, Lecture Notes in Civil Engineering 5 (2018) 73–84. doi:10.1007/978-3-319-67443-8_5.

[46] C. Liu, J. Dobson, P. Cawley, Efficient generation of receiver operating characteristics for the evaluation of damage detection in practical structural health monitoring applications, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 473 (2199). doi:10.1098/rspa.2016.0736.

[47] L. Ierimonti, I. Venanzi, F. Ubertini, Roc analysis-based optimal design of a spatio-temporal online seismic monitoring system for precast industrial buildings, Bulletin of Earthquake Engineering 19 (3) (2021) 1441–1466. doi:10.1007/s10518-020-01032-6.

[48] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, PLoS ONE 10 (3) (2015) 233–240. doi:10.1371/journal.pone.0118432.

[49] R. De Maesschalck, D. Jouan-Rimbaud, D. Massart, The mahalanobis distance, Chemometrics and Intelligent Laboratory Systems 50 (1) (2000) 1–18. doi:10.1016/S0169-7439(99)00047-7.

[50] A critical investigation of recall and precision as measures of retrieval system performance, ACM Transactions on Information Systems (TOIS) 7 (3) (1989) 205–229. doi:10.1145/65943.65945.

[51] J. Maeck, G. De Roeck, Description of z24 benchmark, Mechanical Systems and Signal Processing 17 (1) (2003) 127–131. doi:10.1006/mssp.2002.1548.

[52] G. Steenackers, P. Guillaume, Structural health monitoring of the z-24 bridge in presence of environmental changes using modal analysis, Conference Proceedings of the Society for Experimental Mechanics Series (2005) 18.

[53] C. Kramer, C. de Smet, G. de Roeck, Z24 bridge damage detection tests, Shock and Vibration Digest 32 (1) (2000) 26.