Elsevier Editorial System(tm) for Engineering Structures
Manuscript Draft

Manuscript Number: ENGSTRUCT-D-11-01111R3

Title: Methodologies for predicting natural frequency variation of a suspension bridge

Article Type: Research Paper

Keywords: Environmental effect; artificial neural network; support vector regression; regression tree; random forest; variable importance; suspension bridge

Corresponding Author: Dr. Irwanda Laory, Ph.D

Corresponding Author's Institution: University of Warwick

First Author: Irwanda Laory, Ph.D

Order of Authors: Irwanda Laory, Ph.D; Thanh Ngoc Trinh, Ph.D.; Ian Smith, Ph.D.; James Brownjohn, Ph.D.

Abstract: In vibration-based structural health monitoring, changes in the natural frequency of a structure are used to identify changes in the structural conditions due to damage and deterioration. However, natural frequency values also vary with changes in environmental factors such as temperature and wind. Therefore, it is important to differentiate between the effects due to environmental variations and those resulting from structural damage. In this paper, this task is accomplished by predicting the natural frequency of a structure using measurements of environmental conditions. Five methodologies - multiple linear regression, artificial neural networks, support vector regression, regression tree and random forest - are implemented to predict the natural frequencies of the Tamar Suspension Bridge (UK) using measurements taken from three years of continuous monitoring. The effects of environmental factors and traffic loading on natural frequencies are also evaluated by measuring the relative importance of input variables in regression analysis. Results show that support vector regression and random forest are the most suitable methods for predicting variations in natural frequencies. In addition, traffic loading and temperature are found to be two important parameters that need to be measured. Results show potential for application to continuously monitored structures that have complex relationships between natural frequencies and parameters such as loading and environmental factors.

Highlights

- We compare five regression methods to predict natural frequencies of a bridge.
- Random forest and support vector regression are the most appropriate methods.
- The relative importance of variables is useful to evaluate environmental effects.
- Traffic loading and temperature are the most influential parameters.
- Obtaining these parameters should be a priority.

Abstract

In vibration-based structural health monitoring, changes in the natural frequency of a structure are used to identify changes in the structural conditions due to damage and deterioration. However, natural frequency values also vary with changes in environmental factors such as temperature and wind. Therefore, it is important to differentiate between the effects due to environmental variations and those resulting from structural damage. In this paper, this task is accomplished by predicting the natural frequency of a structure using measurements of environmental conditions. Five methodologies - multiple linear regression, artificial neural networks, support vector regression, regression tree and random forest - are implemented to predict the natural frequencies of the Tamar Suspension Bridge (UK) using measurements taken from three years of continuous monitoring. The effects of environmental factors and traffic loading on natural frequencies are also evaluated by measuring the relative importance of input variables in regression analysis. Results show that support vector regression and random forest are the most suitable methods for predicting variations in natural frequencies. In addition, traffic loading and temperature are found to be two important parameters that need to be measured. Results show potential for application to continuously monitored structures that have complex relationships between natural frequencies and parameters such as loading and environmental factors.

Methodologies for predicting natural frequency variation of a suspension bridge

Irwanda Laory[1], Thanh N. Trinh[2], Ian F. C. Smith[2] and James M. W. Brownjohn[3]

[1]*School of Engineering, University of Warwick, Coventry, United Kingdom*
Email: I.Laory@warwick.ac.uk
[2]*Applied Applied Computing and Mechanic Laboratory, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland*
[3]*Vibration Engineering Section, College of Engineering, Mathematics and Physical Sciences, The University of Exeter, Exeter, United Kingdom*

Abstract

In vibration-based structural health monitoring, changes in the natural frequency of a structure are used to identify changes in the structural conditions due to damage and deterioration. However, natural frequency values also vary with changes in environmental factors such as temperature and wind. Therefore, it is important to differentiate between the effects due to environmental variations and those resulting from structural damage. In this paper, this task is accomplished by predicting the natural frequency of a structure using measurements of environmental conditions. Five methodologies - multiple linear regression, artificial neural networks, support vector regression, regression tree and random forest - are implemented to predict the natural frequencies of the Tamar Suspension Bridge (UK) using measurements taken from three years of continuous monitoring. The effects of environmental factors and traffic loading on natural frequencies are also evaluated by measuring the relative importance of input variables in regression analysis. Results show that support vector regression and random forest are the most suitable methods for predicting variations in natural frequencies. In addition, traffic loading and temperature are found to be two important parameters that need to be measured. Results show potential for application to continuously monitored structures that have complex relationships between natural frequencies and parameters such as loading and environmental factors.

KEY WORDS: *Environmental effect, artificial neural network, support vector regression, regression tree, random forest, variable importance, suspension bridge.*

1. Introduction

Many vibration-based approaches in structural health monitoring have been designed to identify changes in natural frequency values for the purpose of detecting changes in structural conditions that may indicate structural damage and degradation. In reality, however, civil engineering structures are subject to environment and operating effects caused by changes in temperature, traffic, wind, humidity and solar-radiation [1-5]. Such environmental effects also change natural frequency values, hence concealing changes due to structural damage [6-10]. Therefore, it is important to distinguish between changes due to structural damage and changes resulting from environmental effects. This task is managed observing then modeling dependencies of natural frequencies on environmental parameters [11]. The prediction of natural frequencies of structures under environmental changes has been studied using methods such as linear regression analysis, artificial neural networks and support vector regression.

Multiple linear regression (MLR) was employed to predict changes in natural frequencies of the Alamosa Canyon Bridge (USA) due to environmental temperature variation [9] with natural frequencies formulated as a linear function of temperature data. It was found that the changes in the frequencies were linearly correlated with temperature taken from different locations on the bridge. Peeters et al. [12] conducted a one-year monitoring study for the Z24-Bridge (Switzerland) before it was deliberately damaged, applying a linear regression analysis to distinguish normal frequency changes from abnormal changes due to damage. Also, for this concrete box girder bridge, Peeters and Roeck [13] applied an autoregressive method with exogeneous inputs (ARX) to predict the bridge natural frequencies, where no relationship was found between natural frequencies and wind, rainfall and humidity. Liu and

Dewolf [3] simulated the varying natural frequencies under temperature changes using a linear regression analysis, concluding that the long-term variations of natural frequencies are closely related to the variation in in-situ concrete temperature for the three frequencies they measured. The MLR method has also been used to predict natural frequencies of suspension bridges and a footbridge using long-term monitoring data [11, 14].

Artificial neural networks (ANNs) have been successfully applied in fields such as pattern recognition [15], artificial intelligence [16] and civil engineering [17-20]. For long-term monitoring of structures, ANNs have been employed to predict time-dependent natural frequencies of a structure in order to eliminate the environmental effects on vibration-based damage detection procedures. For example, Ni et al. [21] applied an ANN to formulate the correlation between the natural frequencies and environmental temperatures taken from the cable-stayed Ting Kau Bridge (Hong Kong). Zhou et al [22] further investigated the performance of the ANNs formulated using the early stopping technique by constructing three different kinds of input, including mean temperatures, effective temperatures and principle components (PCs) of temperatures. The results indicated that when a sufficient number of PCs were taken into account, the ANN using temperature PCs as inputs predicted natural frequencies more accurately than that when using the mean temperatures. More studies on ANNs for the prediction of structural responses are found in references [22-25].

Support vector regression (SVR) is an application form of support vector machines that is a learning system using a high dimension feature space [26-27]. An attractive characteristic of SVR is that instead of minimizing the observed training error such as with MLR and ANNs, SVR involves minimizing the generalized error bound in order to achieve good performance. The generalized error bound is the combination of the training error and a regularization term that controls the complexity of prediction functions. A good overview of SVR is given in [28-29]. SVR has been successfully employed in fields such as text categorization and pattern recognition as well as structural health monitoring [27, 30]. Ni et al. [31] applied SVR to predict natural frequencies of the cable-stayed Ting Kau Bridge (Hong Kong) subjected to temperature variations taken from one-year measurement data, the method exhibiting better prediction capability than the MLR method. Also using measurement data of this bridge, Hua et al. [32] combined principle component analysis (PCA) and SVR to simulate temperature-frequency correlations. It was found that the SVR method trained using the PCs of measured temperature data outperformed that trained using measured temperature data directly.

The methodologies used above are based on parametric functions that specify the form of the relationship between inputs and a response (output) but in many cases, the form of the relationship is unknown. Regression tree (R_Tree) methods offer a non-parametric alternative [33] that has been used extensively in a variety of fields. The method has been found to be especially useful in biomedical and genetic research, speech recognition and other applied sciences [34]. Recent studies in the machine-learning field found that significant improvements in prediction accuracy have resulted from growing an ensemble of trees in a random way, a methodology called *random forest* (RF) [35]. It has been demonstrated that RF has improved prediction accuracy in comparison to other regression methods [36] but additionally provides measures of variable importance for each input variable [37-38]. This method has not been evaluated for its applicability to structural health monitoring, so this paper investigates the performance of RF on predicting natural frequencies through a case study of a suspension bridge.

The studies mentioned above have proposed methodologies for predicting the dynamic responses of bridges, but none has compared methodologies for prediction accuracy. This paper compares five methodologies – multiple linear regression, artificial neural networks, support vector regression, regression tree and random forest – in terms of their ability to predict natural frequencies of a suspension bridge. Confidence intervals are then defined for the best method to differentiate the effects due to environmental changes from those caused by structural damage. Furthermore, the individual effects of temperature, wind and traffic loading on the natural frequency responses of the bridge are evaluated using the variable importance metric in regression analysis.

2. Methodologies for predicting natural frequencies of the bridge

2.1. Multiple linear regression (MLR)

Assuming that a response variable y (for example natural frequency) is linearly related to the p input variables (for example temperature, wind and traffic loading) $x_1,...x_p$ so that

$$y=\beta_0+\sum_{i=1}^{p}\beta_i x_i+e. \tag{1}$$

This relationship is known as a linear regression analysis, where β_i is the regression coefficient associated with the i^{th} input variable x_i and e the random error with mean zero and variance σ^2. Using the dataset of n observations in measurement time series, the unknown coefficients β_i are determined using the least-squares method.

2.2. Artificial neural networks (ANNs)

Artificial neural networks can be used as a nonlinear regression method to predict the natural frequency of a bridge. ANN is a two-stage regression in which the first stage is to create derived features Z_m, represented by hidden layer, from linear combinations of the inputs and the second stage is to model the output Y_m as a function of linear combinations of the Z_m. Z_m could be considered as a basis expansion of the original input X.

$$
\begin{aligned}
&Z_m = \phi\left(\alpha_{0m} + \alpha_m^T X\right),\ m = 1,\ldots,M,\\
&T_k = \beta_{0k} + \beta_k^T Z\ ,k = 1,\ldots,K,\\
&f_k\left(X\right) = T_k + e, k = 1,\ldots,K,
\end{aligned}
\tag{2}
$$

where $Z = \left(Z_1, Z_2, \ldots, Z_M\right)$, $\phi(v)$ is the activation function which is usually chosen to be the sigmoid $\phi(v) = 1/(1+e^{-v})$, e the random error, α_i and β_i are unknown parameters. Given a training set $\{x_i, y_i\}$ $\left(i = 1,\ldots,N\right)$, the ANN regression model is formulated by searching these unknowns so that the sum-of-squared errors as a measure of fit reaches a minimum value.

$$
R\left(\alpha,\beta\right) = \sum_{k=1}^{K}\sum_{i=1}^{N}\left(y_{ik} - f_k\left(x_i\right)\right)^2 \tag{3}
$$

The generic approach to minimizing, $R(\alpha,\beta)$, is by gradient descent, called back-propagation. A two-layer back-propagation neural network (BPNN) is employed to predict the natural frequencies of a structure. BPNN is first trained using the training set in order to formulate the relationship between the natural frequencies and environmental factors including direct loading such as traffic. BPNN is composed of one hidden layer and one output layer with a tan-sigmoid transfer function in the hidden layer and a linear transfer function in the output layer. The tan-sigmoid transfer function is capable of capturing the nonlinear relationship between input variables (in our example three of them) and output variables (in our example individual natural frequencies).

An important parameter to be determined when using BPNN for prediction tasks is the optimal number of hidden nodes in the hidden layer. A network with too few hidden nodes might not have enough flexibility to capture the nonlinearities in the relationship while a network with too many hidden nodes may have a tendency to overfit the training data.

2.3. Support vector regression (SVR)

The strategy of SVR is to transform nonlinear relationships from the original space into linear relationships in a new space (or feature space) defined using a kernel function so as to discover relationships more easily [27, 36]. The linear function in the new space is given by

$$y(x) = w^T \varphi(x) + b + e \tag{4}$$

where w is the weight vector; b is the bias constant and $\varphi(x)$ is the mapping function that transfers the input vector x into the new space. Given a training set $\{x_i, y_i\}$ $(i = 1,...,N)$, a SVR model is obtained by minimizing the following objective function [39]

$$\begin{aligned} &\min_{w,b,e} J(w,e) = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} e_i^2 \\ &\text{subject to } y_i = w^T \varphi(x_i) + b + e_i, \quad i = 1,...,N. \end{aligned} \tag{5}$$

where γ is the regularization parameter and e_i is the error. Such optimization that is subject to a condition is solved using the Lagrangian function

$$L(w,b,e,\alpha) = J(w,e) - \sum_{i=1}^{N} \alpha_i \left\{ w^T \varphi(x_i) + b + e_i - y_i \right\} \tag{6}$$

where α_i is a Lagrange multiplier. The conditions for optimality are given by

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i \varphi\left(x_i\right) \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i = 0 \\ \dfrac{\partial L}{\partial e} = 0 \rightarrow \alpha_i = \gamma e_i, \qquad\qquad i = 1, \ldots, N. \\ \dfrac{\partial L}{\partial \alpha} = 0 \rightarrow w^T \varphi\left(x_i\right) + b + e_i - y_i = 0, \ i = 1, \ldots, N. \end{cases} \tag{7}$$

Elimination of w and e yields a set of linear equations that are written in the matrix form

$$\begin{bmatrix} 0 & 1_N^T \\ 1_N & \Omega + \gamma^{-1} I_N \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \tag{8}$$

where $Y = \left[y_1, \ldots, y_N\right]^T$, $1_N = \left[1, \ldots, 1\right]^T$ and $\alpha = \left[\alpha_1, \ldots, \alpha_N\right]^T$. I_N is an $N \times N$ identity matrix and Ω is a $N \times N$ kernel matrix defined by a kernel function as

$$\Omega_{ij} = \varphi\left(x_i\right)^T \varphi\left(x_j\right) = K\left(x_i, x_j\right). \tag{9}$$

The kernel function is designed to compute inner-products in the new space using only the original input data. The choice of K implicitly determines φ and the new space. Thus, the advantage of kernel functions is that if a kernel function K is given, it is not necessary to know the explicit form of the mapping function $\varphi\left(x\right)$. The selection of the kernel function generally depends on the application domain. It has been shown that Gaussian radial-basis function (RBF) is a reasonable first choice of kernel functions since it has only a single parameter (standard deviation, σ) to be determined [27, 40]. The Gaussian RBF is expressed as

$$K\left(x_i, x_j\right) = e^{-\left\|x_i - x_j\right\|^2 / 2\sigma^2} . \tag{10}$$

Solving Equation (8) identifies the values of α and b. Then, substituting these values into Equation (4) leads to the prediction

$$y\left(x\right) = \sum_{i=1}^{N} \alpha_i K\left(x, x_i\right) + b . \tag{11}$$

There are only two tuning parameters, γ and σ, that need to be determined when using the RBF kernel function and their optimal values are determined using the grid search method . Possible intervals for the two parameters are first defined. Then all grid points are tried to find the one giving the best accuracy. For each combination of the two parameters, SVR is trained using the training data and their performance is evaluated by a ten-fold cross-validation scheme.

2.4. Regression tree (R_Tree)

Regression tree is a nonparametric statistical method [33] that offers an alternative to parametric regression methods which usually require assumptions and simplifications to form the relationship. A regression tree is built by recursively partitioning the entire dataset, represented by a *root node*, into more homogeneous groups with each to be represented by a node. When the splitting process terminates, each resulting group is referred to as a terminal node. Splitting at each node is based on one value of an input variable that leads to the most homogeneous resulting nodes. Assuming that we have a partition into M regions R_1, R_2, ..., R_M the system model is identified as

$$y(x) = ave(y_j \left| x_j \in R_m \right.) + e \tag{11}$$

Where y_j and x_j represent the response and input variables at j^{th} observation respectively. Equation 10 shows that the predicted response is the average of y_j in region R_m with the error e.

A simple regression tree is built with two input variables x_1 and x_2 and a response y by considering a recursive partition as shown in Figure 1(a). First, we select the *splitting variable* (for example, x_1) and the *split point* (for example s_1) in order to achieve the most homogeneous splitting groups and split the space of the dataset into two groups. The selected variable and point solve

$$\min_{j,s}\left[\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2 + \sum_{x_i\in R_2(j,s)}(y_i-c_2)^2 +\right] \tag{11}$$

c_1 and c_2 are the mean value of all the responses in the corresponding groups. Then, each of these groups is further split into two more groups. As shown in Figure 1(a), the group $x_1 \le s_1$ is split at $x_2 \le s_2$ and finally the group $x_1 > s_1$ is split at $x_1 = s_3$. The process results in four groups R_1, …, R_4. This process can be represented by the binary tree (Figure 1(b)). The entire dataset sits at the top of the tree, as a so-called root node. Observations (data points) satisfying the condition at each node are assigned to the left branch, and the others to the right branch. The terminal nodes of the tree correspond to the groups, R_1, .., R_4. Once a tree has been built, the response for any new observation can be predicted by following the path from the root node down to the appropriate terminal node of the tree, based on the observed values of the splitting variables.

When determining tree size, note that a small tree may not capture a nonlinear relationship that may exist while a very large tree may over-fit the data. Therefore, tree size is a tuning parameter and the optimal tree size should be adaptively chosen from the data. The preferred strategy is to gradually increase the tree size and evaluate the accuracy of each tree size until each node contains fewer than a given number of observations (for example, 5). Then this large tree is pruned by sequentially cutting off branches that add the smallest capability to predictive performance of the tree according to a specified pruning criterion.

2.5. Random forest (RF)

A random forest is a combination of regression trees that are grown in random ways [35]. The idea behind the random forest method is to generate an ensemble of low-correlated regression trees and average results in order to reduce variance. The low-correlated trees are generated by adding randomization in two steps: (i) each tree is grown using a random sub-dataset of observations and (ii) each node of a tree is split using a random subset of input variables. Figure 2 shows the layout of the random forest method.

The first step is to generate B sub-datasets of observations by randomly copying observations from the original training set L until each sub-dataset has the same number of observations N as the original training set. Some observations can be chosen several times for each sub-dataset, whereas others are not chosen at all. It has been proved that about 37% of the observations in the original training set are not chosen for each sub-dataset [38, 41]. The collection of non-chosen observations corresponding to each sub-dataset functions as a validation set. Each sub-dataset is denoted L_b where $b = 1, 2, ..., B$.

The second step involves growing a regression tree (T_b) using a sub-dataset (L_b). This step is to reduce further the correlation between the regression trees that enter into the averaging step later. This is achieved during the tree-growing process by randomly selecting a subset of m input variables from all p input variables $(m \le p)$ before splitting each node. A regression tree is grown by recursively repeating the following three sub-steps for each node until the specified number of observations within each node is reached.

- Randomly select a subset of m variables from all p variables.
- Find the best split among the m variables.
- Split the selected node into two resulting nodes.

After B regression trees are grown from B sub-datasets, an ensemble of these B trees is called a random forest. The random forest makes a prediction for a new observation x by using each regression tree T_b in the forest to obtain a prediction $y_b(x)$ and then averaging B prediction values from the B trees:

$$y(x) = \frac{1}{B}\sum_{b=1}^{B} y_b(x) + e \tag{12}$$

3. Case study subject: The Tamar suspension bridge

The Tamar Suspension Bridge, as shown in Figure 3, is a road bridge connecting Saltash to Plymouth in southwest England. The original bridge was designed as a conventional suspension bridge with symmetrical geometry and was first opened in 1961. The total length is 642 m with a main span of 335 m and side spans of 114 m and the tower height is 73 m. Trusses are 4.9 m deep with chords of welded hollow box structures. To meet the requirement that bridges should be capable of carrying lorries up to 40 tons, the Tamar Bridge was strengthened and widened in March 1999 and the upgrading was completed in December 2001[42-43]. The upgrading included replacing the original composite main deck by a three-lane orthotropic steel deck, adding single lane cantilevers at each side of the truss and installing sixteen new cables acting as additional stays to carry the additional dead load of new cantilever lanes and associated temporary works. Figure 2 shows the layout of one of the truss sections with the main orthotropic deck and two cantilever lanes.

Many types of sensors were installed during and subsequent to the strengthening and widening to monitor the behavior of the bridge [44]. They included anemometers, displacement sensors, thermometers, load cells and accelerometers. Most recently, a robotic total station was added to monitor the deflection of the bridge deck and a pair of extensometers installed to track relative movement across the single expansion joint located around the Saltash Tower [45]. Measurement data have been collected continuously since February 2007.

These data used in this study include air temperature, wind velocity, the measured natural frequencies of the bridge and the number of vehicles crossing the bridge every hour. Vehicle crossing data were available from the bridge toll reports, temperature and wind values are 30-minute averages of data sampled at either 1Hz from four thermistors on the cable and deck and an anemometer close to midspan, while frequencies are derived from modal analysis of 64-Hz sampled acceleration signals from a pair of accelerometers located near mid span. Locations of these sensors are shown in Figure 3. The covariance-driven stochastic subspace identification (SSI-COV) procedure operated automatically on the acceleration data after 8-fold decimation, reporting frequency and damping estimates at 30 minute intervals.

Figure 6 shows the time history of air temperature for three years, including daily and seasonal temperature variations. The temperature ranges from -5 °C to 25 °C between winter and summer. The first five natural frequencies of the bridge are summarized in Table 1.

4. Results

The five regression methodologies presented in the previous section are applied to predict the natural frequency variation of the Tamar Bridge based on environmental factors as well as traffic loading. The prediction is performed for each natural frequency separately. The measurement data taken from July 2007 to January 2010 on the Tamar Bridge are divided into two non-overlapping and independent data sets: a training set of 70% and a test set of 30%. While the training set (from July 2007 to May 2009) is used for regression analysis to predict natural frequencies, the test set is used for assessing prediction accuracy (from June 2009 to January 2010).

4.1. *Multiple linear regression*

The relationship between natural frequency responses and air temperature, wind and traffic loading are first formulated for each frequency using the least square method. Figure 7 shows the prediction of the 10-day time histories (from 20 to 30 July 2009) of the third frequency that is compared with the measured frequency in the test phase. This figure indicates that the predicted frequency is unable to capture the high variation in the natural frequency. Table 2 presents the mean square error (MSE) values in the training and test sets, where the error is the difference between the measured frequency value and its corresponding predicted value. For frequency 1 and 3, MSE values in the training set are somewhat larger than those in the test set. This is because there are more outliers in the training data than in the test set.

4.2. *Artificial neural networks*

The optimal number of hidden nodes in the hidden layer is determined so that the validation error reaches the minimum value. To do this, a set of neural networks with respect to the increasing number of hidden nodes from 1 to 50 are trained using training data. The number of hidden nodes of the neural network that gives the minimum error is taken as the optimal number.

Table 3 presents the optimal numbers of hidden nodes for five natural frequencies, together with MSE values of the training set and the test set. The optimal number of hidden nodes for five frequencies ranges from 10 to 33 nodes. These values are close to the optimal value (19 hidden nodes) for the first natural frequency of the Ting Kau cable-stayed bridge [46]. Figure 8 shows the predicted natural frequency along with the measured frequency. Comparing Figure 7 and Figure 8 indicates that an artificial neural network achieves a better prediction value than multiple linear regression. Comparing the prediction capability of ANN with other methods is further discussed in Section 4.6.

4.3. Support vector regression

Table 4 presents the optimal values of γ and α that give the best performance (lowest MSE) of SVR for five natural frequencies. The corresponding MSEs are also listed in this table. Comparing the MSE values in Table 4 with Table 2 and Table 3 indicates that the SVR method has a better performance than the MLR and ANN methods in both the training set and the test set. For example, the prediction error (in the test set) for frequency 5 using SVR is reduced by 20% when compared with the prediction error using MLR. Figure 9 shows the predicted and measured time histories of frequency 3 from July 20 to 30, 2009. It is shown that the predicted frequencies closely match the measured ones.

4.4. Regression tree

Figure 10 shows the mean squared error with respect to the increasing number of terminal nodes (i.e. tree size) of the pruned tree for the first natural frequency. The optimal tree size (i.e. the point where increasing tree size only leads to minor decrease of MSE) for this frequency is composed of 31 terminal nodes. The optimal tree sizes for the second and third frequencies are 44 and 35 terminal nodes respectively (Table 5). It is observed that the higher frequency requires more terminal nodes, leading to a larger tree size, i.e. higher tree complexity. Table 5 also presents the mean squared errors in the training and test sets. The prediction of the R_Tree method is better than that of the ANN and MLR methods but it is not as good as that of the SVR method.

Figure 11 shows the 10-day time histories of measured and predicted frequencies from July 20 to July 30, 2009. For both sets, the predicted frequency time history reasonably matches the measured one. Flatness exists at some peaks of the predicted time history. This is because the observations at the peaks fall into the same groups where the predicted responses are equal to the mean of measured responses within the corresponding group.

4.5. Random forest

When applying the random forest method for regression analysis, three parameters need to be determined: (i) the sufficient number B of trees, (ii) the optimal number of observations in each terminal node and (iii) the number m of input variables randomly chosen as candidates for splitting at each node. For the Tamar Bridge, there are three input variables (i.e. $p = 3$) including temperature, wind and traffic. For this case study, to reduce the correlation between regression tress, the number of input variables chosen for splitting at each node is 2 (i.e. $m = 2$).

Figure 12 shows that when the number of trees increases, the mean squared error computed from the validation set decreases. The prediction is stable at about 100 trees for both cases of 1 and 50 observations in each terminal node. It is seen that the tree with 50 observations in terminal nodes performs better than that with only one observation in terminal nodes. This is attributable to the over-fitting situation when growing a tree to its maximum size (i.e. one observation in terminal nodes).

Figure 13 shows the change in the normalised mean squared error with respect to the increase in the number of observations in terminal nodes for 5 frequencies. For each frequency, the normalised MSE is calculated by dividing the MSE by the difference between maximum and minimum MSE. When the number of observations in terminal nodes starts increasing, initially the normalised MSE of all three frequencies drops dramatically to a minimum and then it increases gradually. The optimal number of observations in terminal nodes ranges from 10 to 50 observations. Table 6 presents the optimal number of observations for each frequency together with its mean squared errors computed from the training set and the test set. The results show that the random forest method has the smallest errors as compared with those from the previous four methods. Figure 14 compares the predicted natural frequency with the measured frequency. The predicted frequency closely matches the measured one.

4.6. Performance comparisons and discussions

In order to find a suitable method to predict the natural frequency responses from environmental measurement data for a suspension bridge, the prediction capability of five regression methods are compared. The result of a regression method can have a very good fit to the training data; however, it may poorly predict the response for a new observation. Thus, the prediction capability of these methods is evaluated through prediction error that is defined as the mean squared error from the test set, with a smaller prediction error indicating a better prediction capability. When comparing the prediction error of the five regression methods from Table 2 to Table 6, it can be seen that the four nonlinear regression methods (ANNs, SVR, R_Tree and RF) predict frequencies more accurately than the MLR method. Table 7 presents the reduction in the prediction error for these methods when using the prediction error of the MLR method as a basis. For frequency 5, SVR and RF can reduce the prediction error up to 20% when compared with MLR. The good performances of SVR and RF indicate the possibility of existence of non-linear correlations between natural frequency responses and environmental factors as well as traffic loading for the Tamar Suspension Bridge. In addition, comparing Figure 11 and Figure 14 demonstrates that RF employing multiple trees, which are grown in a random way, can lead to better predictions than the R_Tree method that employs a single tree. RF is able to capture the high variations at peaks of frequency time histories.

The performance of SVR and RF are further assessed through a normality test [47]. From a statistical point of view the error, which is the difference between the predicted value and the corresponding measured frequency value, complies with a normal distribution with zero mean [32]. Figure 15 compares the observed probability density functions of the error with the corresponding theoretical curves of the normal distribution obtained using the mean and standard deviation values computed from error values. The figure shows the observed probability distribution of the error for SVR and RF methods is in good agreement with a normal distribution with zero mean.

SVR and RF are used to define the confidence intervals around the predicted natural frequencies for a new observation. It is found that the error in the training data for SVR and RF also have a normal distribution with zero mean. Thus, the confidence interval is defined based on the error variance of the training data. Figure 16 shows the identified and predicted natural frequencies for RF between July 20 and July 30 (2009), together with the 95% confidence interval for the second natural frequency. For the test set, the ratio of the data that falls within the 95% confidence level to the full set of the data is referred to as the success rate. For SVR, the success rates for frequencies 1 to 5 are 98%, 91%, 98%, 94% and 91%, respectively. The corresponding success rates for RF are 98%, 91%, 98%, 94% and 89%. These high success rates indicate that the variations in bridge natural frequencies can be accounted for by measuring temperature, wind and traffic loading. These rates also demonstrate the consistency of continuously monitored data from 2007 to 2010, thereby establishing a baseline data for continuous health monitoring of the bridge. In addition, the success rate can be used as a damage-detection index. If the success rates for future natural frequencies change, it is likely that the bridge has experienced some kind of structural change.

5. Effects of environmental factors and traffic loading on natural frequencies of the bridge

The changes in bridge natural frequencies are adequately accounted for by three factors: temperature, wind and traffic loading. This study identifies the degree to which each factor has an effect on the frequency change. Simultaneous effects of these factors on the first five natural frequency responses are evaluated. This is carried out by using the measure of relative importance of input variables in regression analysis. The measure of relative importance indicates the variables that are highly related to the response for interpretation purposes.

5.1. Evaluation of effects using relative importance metrics of the multiple linear regression method

Multiple linear regression can be used to evaluate the contribution of an individual input variable $x_j\left(j=1,..,p\right)$ to the prediction of a response y. The contribution of each variable is compared with that of other variables using a metric of so-called relative importance. Several relative importance metrics have been proposed to assess the amount of variation in the response that is explained by each individual variable [48-49]. In this study, since the correlation between input variables is negligible, the relative importance of each individual variable is defined as the squared correlation coefficient of an input variable x_j with the response y.

Figure 17 shows the relative importance of temperature, wind and traffic loading using MLR for the first five natural frequencies of the bridge. The effects of temperature, wind and traffic loading on the first frequency are 8%, 18% and 74%. Such effects correspond to 34%, 10% and 56% for the second frequency. They are 28%, 10% and 62% for the third frequency and 22%, 10% and 68% for the fifth frequency. Except for the fourth frequency (i.e. 70%, 21% and 9%), based on relative importance metrics defined using MLR, traffic loading is the main factor that affects the natural frequencies.

5.2. Evaluation of effects using relative importance metrics of the random forest method

The random forest method has improved the prediction accuracy in comparison to other prediction methods. Besides, RF also evaluates the relative importance of variables in a dataset in order to measure the prediction strength of each variable.

As mentioned in Section 2.5, approximately 63% of the observations in the original training set are used for each sub-dataset on which to grow each individual tree. The non-chosen observations (about 37%) are utilized as validation observations for that tree. The computation of the importance of an input variable x_j is carried out one tree at a time. First, when the b^{th} tree T_b is grown, the validation observations are then used to determine the mean squared error from the validation data MSE_b. Next, the values of variable x_j in the validation data are randomly permutated while leaving the values of all other variables unchanged. Then, the permuted observations are used in the tree T_b and the mean squared error from the permuted validation data $MSE_b\left(x_j\right)$ is computed. If x_j is important, permuting its observed values will reduce the prediction accuracy of each observed value in the validation data. Thus, $MSE_b\left(x_j\right)$ from the permuted validation data is larger than MSE_b from the un-permuted data.

Finally, a measure of the importance of the j^{th} variable x_j is obtained by averaging the mean squared errors from the permuted validation data over all of the trees:

$$imp\left(x_j\right) = \frac{1}{B}\sum_{b=1}^{B}\left(MSE_b\left(x_j\right) - MSE_b\right). \tag{13}$$

The relative importance of each variable is computed by normalizing its importance to the summation of the importance of all variables. The relative importance metrics are expressed in percentage. Figure 18 shows the relative importance of temperature, wind and traffic loading on the natural frequency responses of the bridge. There is a significant effect of traffic loading on the frequency. Figure 18 also indicates that while the effect of traffic loading decreases from frequencies 1 to 5, the effect of temperature increases respectively. Both effects are almost similar for frequency 5 and the effect of temperature is more dominant than that of traffic loading for frequency 4.

5.3. *Discussion*

Comparing Figure 17 and Figure 18 shows that although variable importance metrics are defined in two different ways using multiple linear regression and random forest, the importance rankings for temperature, wind and traffic are identical. For the first frequencies, the averaged percentages of the effects taken from both variable importance metrics are about 8%, 17% and 75% respectively. Such percentages correspond to 26%, 15% and 59% for the second frequency, 24%, 9% and 66% for the third frequency, 60%, 22% and 18% for the fourth frequency and 31%, 11% and 58% for the fifth frequency. A possible reason for the effect of traffic loading is that there is a significant contribution of the traffic mass to the total mass of the truss-span suspension bridge. Despite the strong influence on other frequencies, the relative effect of the traffic on the fourth frequency is quite small. This could be due to the fact that the fourth frequency refers to a torsional vibration mode while other frequencies refer to vertical and lateral modes.

As for temperature effects, the influence on the variation of the fourth and fifth frequencies is larger than that of the other frequencies. This could be caused by the non-linear temperature distribution due to solar radiation. In general, for successful data interpretation when monitoring natural frequency responses of suspension bridges, the effects of both traffic loading and temperature need to be taken into account.

6. Conclusions

This paper compares five methodologies to predict the natural frequency responses of a suspension bridge using measurements of temperature, wind and traffic loading. The following conclusions are drawn

- Random forest and support vector regression are the most appropriate methods for predicting the natural frequencies of a suspension bridge using measurement data of temperature, wind and traffic loading. This may be due to non-linear behavior.

- The relative importance of input variables of regression analysis is a useful metric to evaluate the simultaneous effects of environmental factors and traffic loading on the long-term natural frequency responses of a bridge.
- Traffic loading and temperature are the most influential parameters on natural frequencies of the suspension bridge studied. Obtaining these parameters should be a priority when using natural frequency changes to detect damage.

Acknowledgements

The authors are grateful to Dr. Minh-Hai Pham and Dr. Ki-Young Koo for their contributions to this research.

References

[1] Wahab MA, De Roeck G. Effect of temperature on dynamic system parameters of a highway bridge. Structural Engineering International. 1997;7:266-70.

[2] Yuen K-V, Kuok S-C. Ambient interference in long-term monitoring of buildings. Engineering Structures. 2010.

[3] Liu C, DeWolf JT. Effect of temperature on modal variability of a curved concrete bridge under ambient loads. Journal of Structural Engineering. 2007;133:1742-51.

[4] Xu Z-D, Zhishen Wu. Simulation of the effect of temperature variation on damage detection in a long-span cable-stayed bridge. Structural Health Monitoring. 2007;6:177-89.

[5] Brownjohn JM. Thermal effects on performance on tamar bridge. in *4th International Conference on Structural Health Monitoring of Intelligent Infrastructure (SHMII-4)*. Zurich, Swirzerland. 2009.

[6] Laory I, Trinh TN, Smith IFC. Evaluating two model-free data interpretation methods for measurements that are influenced by temperature. Adv Eng Inform. 2011;25:495-506.

[7] Posenato D, Lanata F, Inaudi D, Smith IFC. Model-free data interpretation for continuous monitoring of complex structures. Advanced Engineering Informatics. 2008;22:135-44.

[8] Posenato D, Kripakaran P, Inaudi D, Smith IFC. Methodologies for model-free data interpretation of civil engineering structures. Computers & Structures. 2010;88:467-82.

[9] Sohn H, Dzwonczyk M, Straser EG, Kiremidjian AS, Law K, Meng T. An experimental study of temperature effect on modal parameters of the alamosa canyon bridge. Earthquake Engineering and Structural Dynamics. 1999;28:879-97.

[10] Trinh TN, Koh CG. An improved substructural identification strategy for large structural systems. Structural Control and Health Monitoring. 2011:doi: 10.1002/stc.463.

[11] Moser P, Moaveni B. Environmental effects on the identified natural frequencies of the dowling hall footbridge. Mechanical Systems and Signal Processing. 2011;25:2336-57.

[12] Peeters B, Maeck J, Roeck GD. Vibration-based damage detection in civil engineering: Excitation sources and temperature effects. Smart Materials and Structures. 2001;10:518.

[13] Peeters B, De Roeck G. One-year monitoring of the Z24-bridge: Environmental effects versus damage events. Earthquake Engineering & Structural Dynamics. 2001;30:149-71.

[14] Yang D, Youliang D, Aiqun L. Structural condition assessment of long-span suspension bridges using long-term monitoring data. Earthquake Engineering and Engineering Vibration. 2010;9:123-31.

[15] Bishop C. Neural networks for pattern recognition: Oxford University Press, USA; 1996.

[16] Raphael B, Smith IFC. Fundamentals of computer-aided engineering: West Sussex, England : Wiley; 2003.

[17] Chan THT, Ni YQ, Ko JM. Neural network novelty filtering for anomaly detection of tsing ma bridge cables in *The 2nd International Workshop on Structural Health Monitoring*. Stanford University, Stanford, CA, USA. 1999.

[18] Fernandez B, Parlos AG, Tsai WK. Nonlinear dynamic system identification using artificial neural networks (anns). New York, NY, USA: IEEE. 1990.

[19] Huang C-C, Loh C-H. Nonlinear identification of dynamic systems using neural networks. Computer-Aided Civil and Infrastructure Engineering. 2001;16:28-41.

[20] Adeli H. Neural networks in civil engineering: 1989–2000. Computer-Aided Civil and Infrastructure Engineering. 2001;16:126-42.

[21] Ni YQ, Zhou HF, Ko JM. Generalization capability of neural network models for temperature-frequency correlation using monitoring data. J Structural Engineering. 2009;135:1290.

[22] Zhou HF, Ni YQ, Ko JM. Constructing input to neural networks for modeling temperature-caused modal variability: Mean temperatures, effective temperatures, and principal components of temperatures. Engineering Structures. 2010;32:1747-59.

[23] Zhou HF, Ni YQ, Ko JM. Performance of neural networks for simulation and prediction of temperature-induced modal variability. in *Smart Structures and Materials 2005: Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems, 7 March 2005*. USA: SPIE-Int. Soc. Opt. Eng. 2005.

[24] Freitag S, Graf W, Kaliske M, Sickert JU. Prediction of time-dependent structural behaviour with recurrent neural networks for fuzzy data. Computers & Structures.In Press, Corrected Proof.

[25] Graf W, Freitag S, Sickert J-U, Kaliske M. Prediction of time-dependent structural behavior with recurrent neural networks. in *The Sixth International Structural Engineering and Construction Conference*. Zürich, Switzerland. 2011. 789-94.

[26] Vapnik V, Golowich S, Smola A. Support vector method for function approximation, regression estimation, and signal processing. in *Advances in Neural Information Processing Systems 9*. 1996.

[27] Saitta S, Kripakaran P, Raphael B, Smith IFC. Feature selection using stochastic search: An application to system identification. Journal of Computing in Civil Engineering. 2010;24:3-10.

[28] Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and Computing. 2004;14:199-222.

[29] Basak D, Pal S, Patranabis D. Support vector regression. Neural Information Processing – Letters and Reviews. 2007;11.

[30] Zhang J, Sato T. Experimental verification of the support vector regression based structural identification method by using shaking table test data. Structural Control and Health Monitoring. 2008;15:505-17.

[31] Ni YQ, Hua XG, Fan KQ, Ko JM. Correlating modal properties with temperature using long-term monitoring data and support vector machine technique. Engineering Structures. 2005;27:1762-73.

[32] Hua XG, Ni YQ, Ko JM, Wong KY. Modeling of temperature-frequency correlation using combined principal component analysis and support vector regression technique: ASCE; 2007.

[33] Breiman L, Friedman J, Stone C, Olshen RA. Classification and regression trees: Chapman & Hall/CRC; 1984.

[34] Izenman A. Modern multivariate statistical techniques : Regression, classification, and manifold learning: Springer New York; 2008.

[35] Breiman L. Random forests. Machine Learning. 2001;45:5-32.

[36] Zheng L, Watson DG, Johnston BF, Clark RL, Edrada-Ebel R, Elseheri W. A chemometric study of chromatograms of tea extracts by correlation optimization warping in conjunction with pca, support vector machines and random forest data modeling. Analytica Chimica Acta. 2009;642:257-65.

[37] Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis. 2008;52:2249-60.

[38] Grömping U. Variable importance assessment in regression: Linear regression versus random forest. The American Statistician. 2009;63:308-19.

[39] Suykens JAK, Gestel TV, Brabanter JD, Moor BD, Vandewalle J. Least squares support vector machines. Singapore: World Scientific; 2002.

[40] Hsu C, Chang C, Lin C. A practical guide to support vector classification. 2010.

[41] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Data mining, inference and prediction. 2 ed: Springer; 2009.

[42] Koo KY, Brownjohn JMW, List D, Cole R. Innovative structural health monitoring for suspension bridges by total positioning system. in *IABMAS10*. Philadelphia, USA. 2009.

[43] Westgate R, Brownjohn JMW. Development of a tamar bridge finite element model. in *IMAC XXVIII*. Florida, USA. 2010.

[44] Koo KY, Brownjohn JMW, Cole R, List DI. Structural health monitoring of tamar suspension bridge. Submitted to Journal of Structural Control and Health Monitoring. 2011.

[45] Koo KY, Brownjohn JMW. Robotic total station for long-term structural health monitoring of the tamar bridge. Submitted to Engineering Structures. 2011.

[46] Ni YQ, Zhou HF, Ko JM. Generalization capability of neural network models for temperature-frequency correlation using monitoring data: ASCE; 2009.

[47] Freund RJ, Wilson WJ, Mohr D. Statistical methods. London: Academic Press; 2010.

[48] Chao Y-CE, Zhao Y, Kupper LL, Nylander-French LA. Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. Journal of Occupational and Environmental Hygiene. 2008;5:519 - 29.

[49] Groemping U. Relative importance for linear regression in r: The package relaimpo. Journal of Statistical Software. 2006;17.

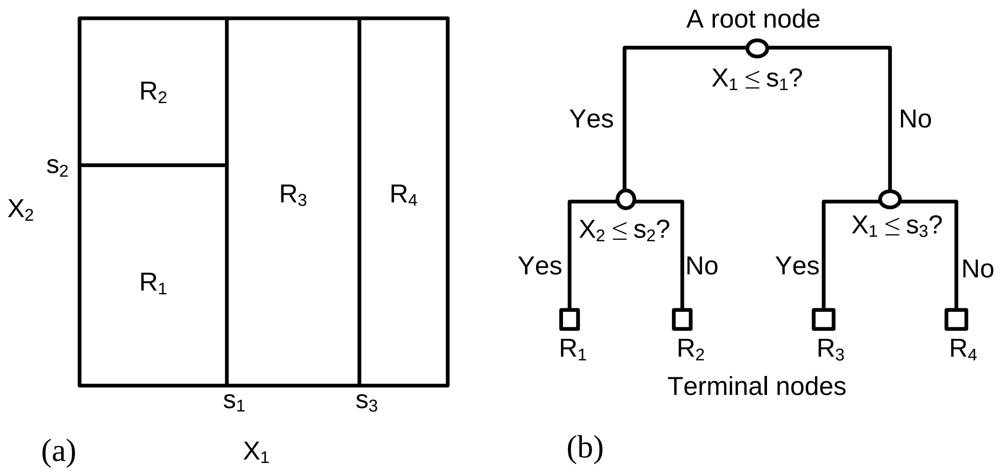


Figure 1. (a) The partitioning of a two-dimensional feature space into four regions, R_1-R_4; (b) a decision tree with three splits and four terminal nodes corresponding to the four partitions.

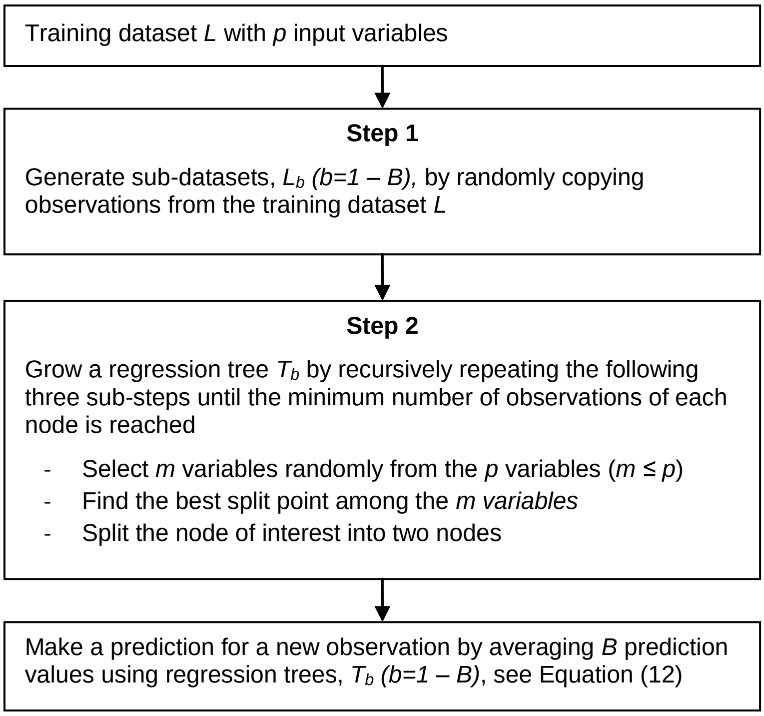


Figure 2. A layout of the random forest analysis method.

Figure 3. The Tamar Suspension Bridge

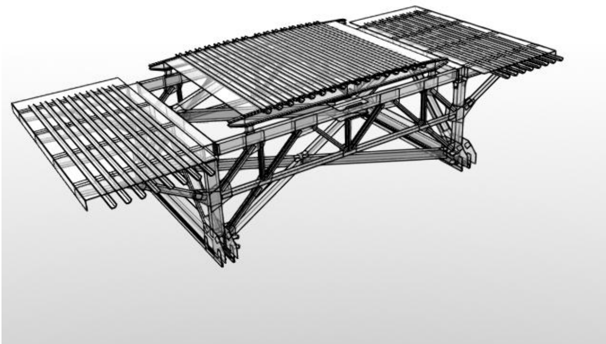


Figure 4 The truss section with the main orthotropic deck and two cantilever lanes

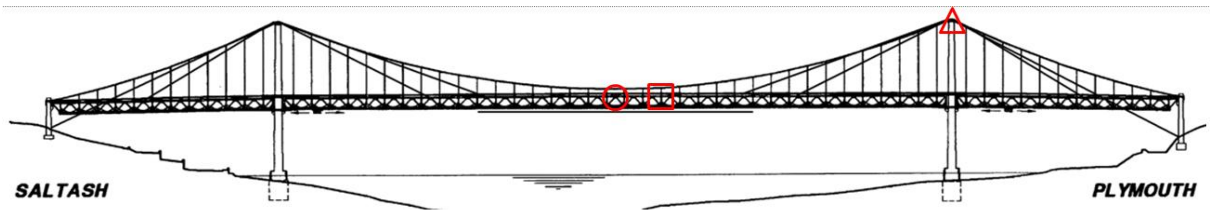


Figure 5 Sensor locations (circle for accelerometers, square for thermistors and triangle for anemometer)

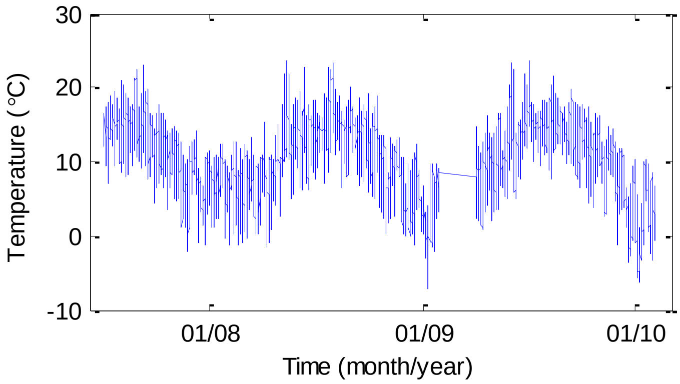


Figure 6. Time history of temperature measured from 2007 to 2010

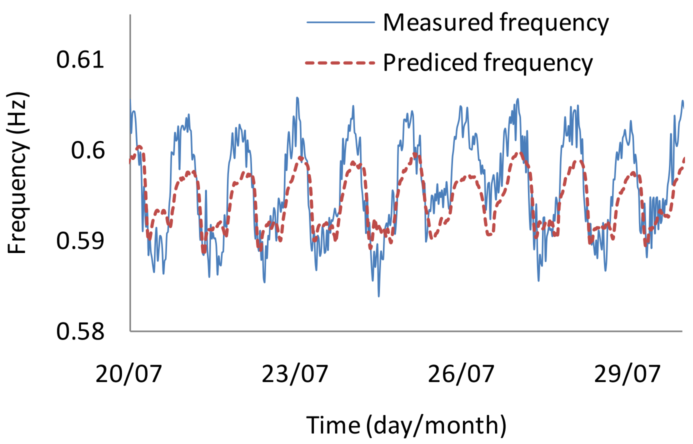


Figure 7. Measured and predicted natural frequencies between July 20 and July 30, 2009 using the multiple linear regression method.

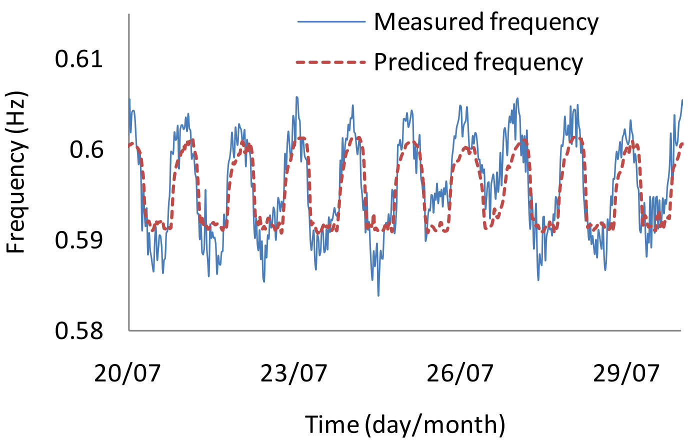


Figure 8. Measured and predicted natural frequencies between July 20 and July 30, 2009 using the artificial neural network method.

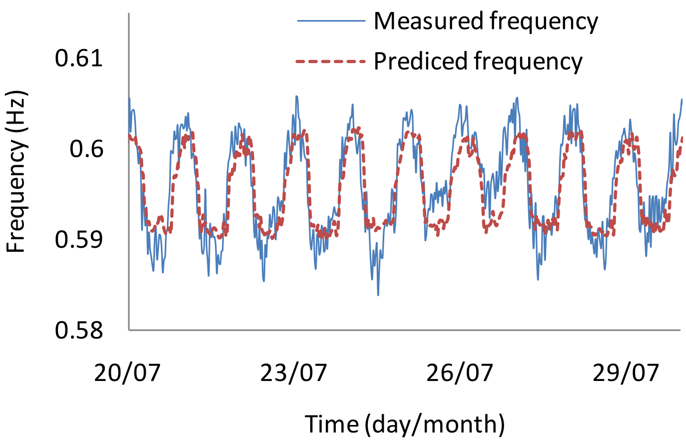


Figure 9. Measured and predicted natural frequencies between July 20 and July 30, 2009 using the support vector regression method.

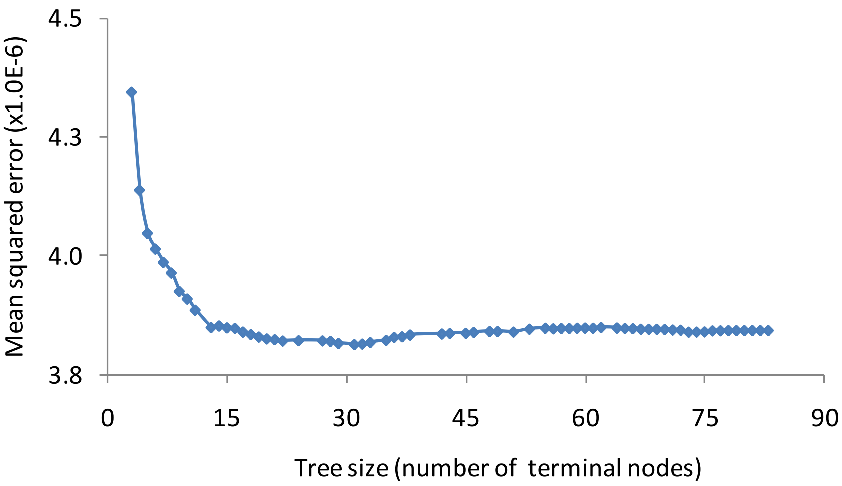


Figure 10. Mean squared errors versus the number of terminal nodes of a tree.

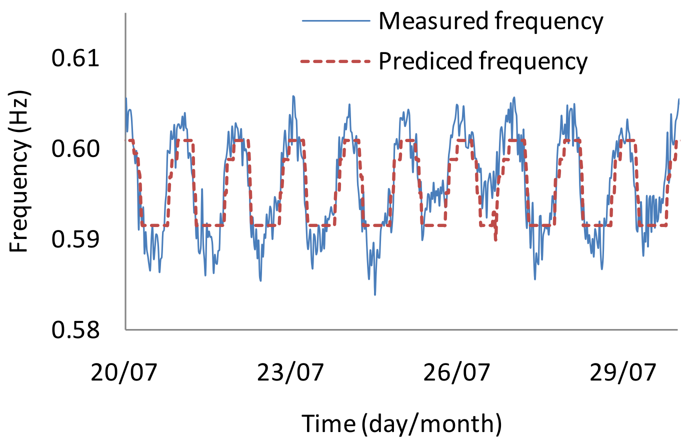


Figure 11. Measured and predicted natural frequencies between July 20 and July 30, 2009 using the regression tree method.

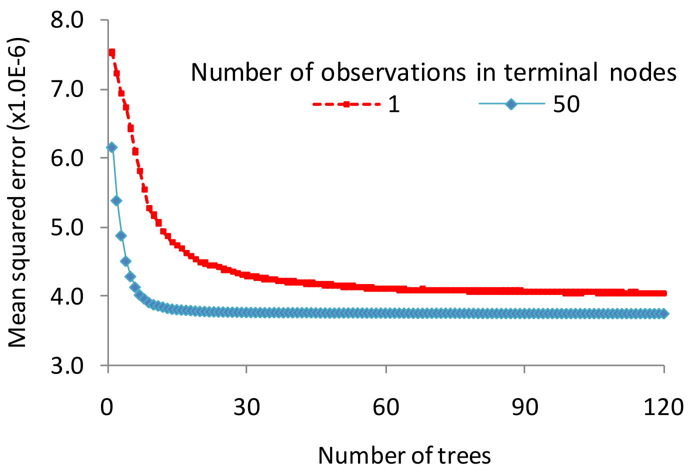


Figure 12. Mean squared errors versus number of trees for two cases of 1 and 50 observations in terminal nodes.

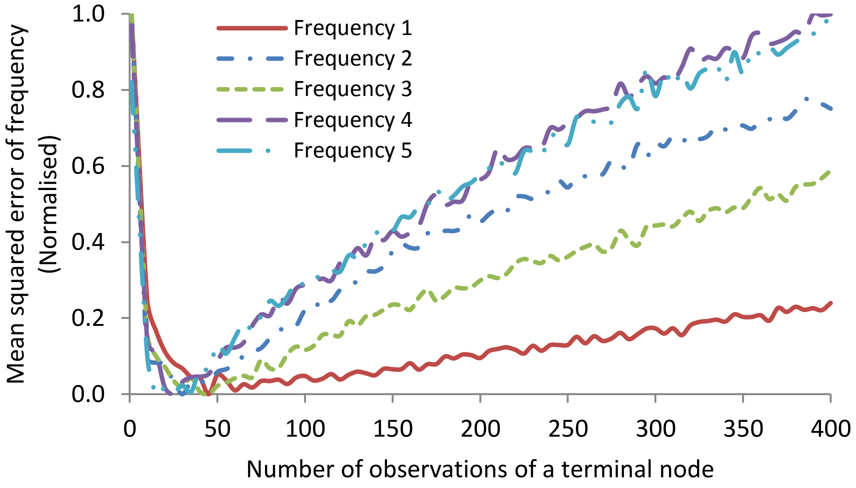


Figure 13. Mean squared errors versus the number of observations in a terminal node

627

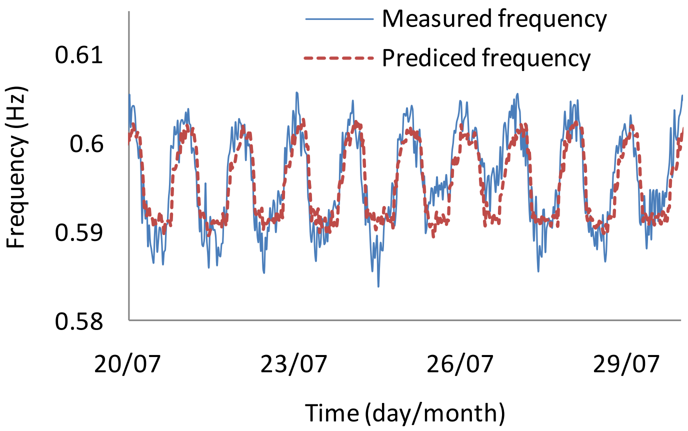


Figure 14. Measured and predicted natural frequencies between July 20 and July 30, 2009 using the random forest method

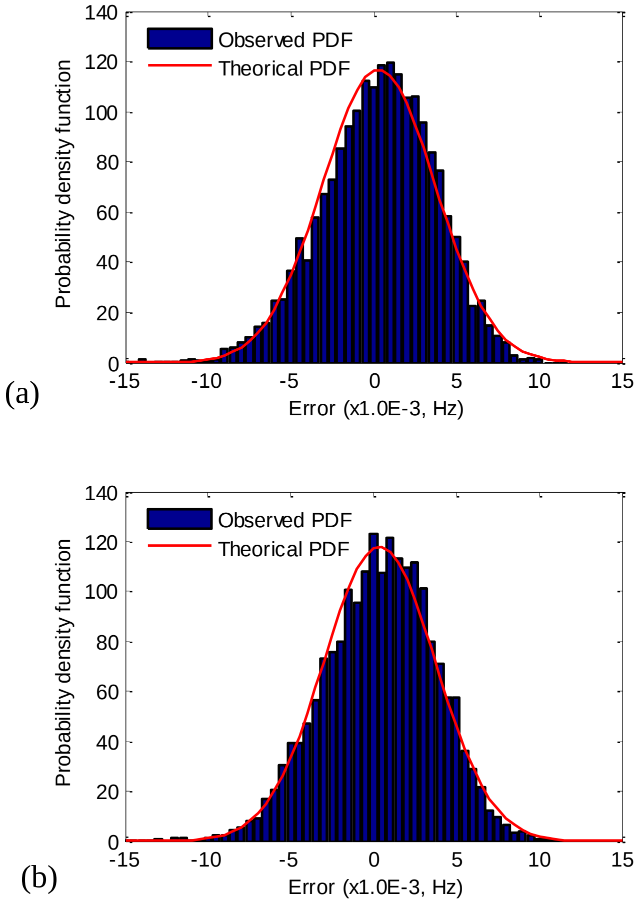
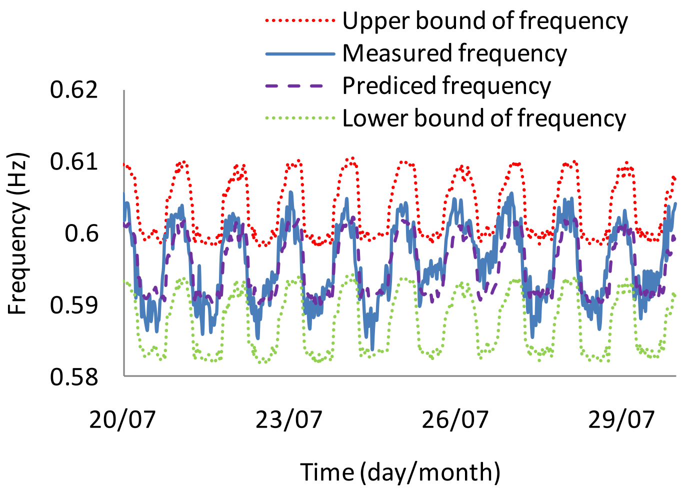
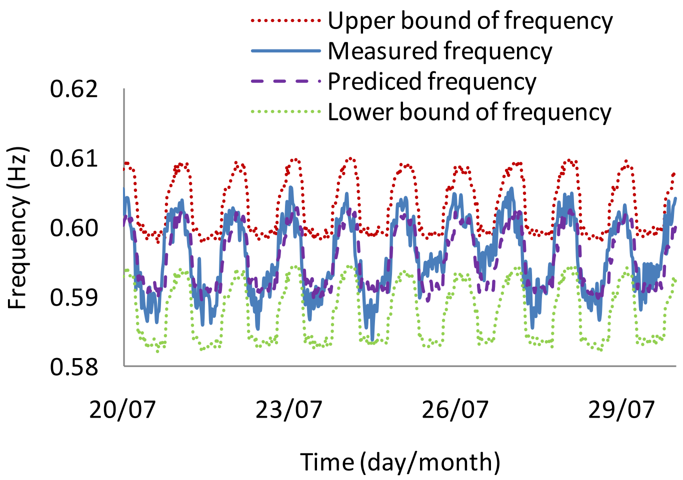


Figure 15. Probability distribution of errors for (a) support vector regression and (b) random forest



(a)



(b)

Figure 16. Measured and predicted natural frequencies (20 – 30 July, 2009) together with the 95% confidence interval using (a) support vector regression and (b) random forest

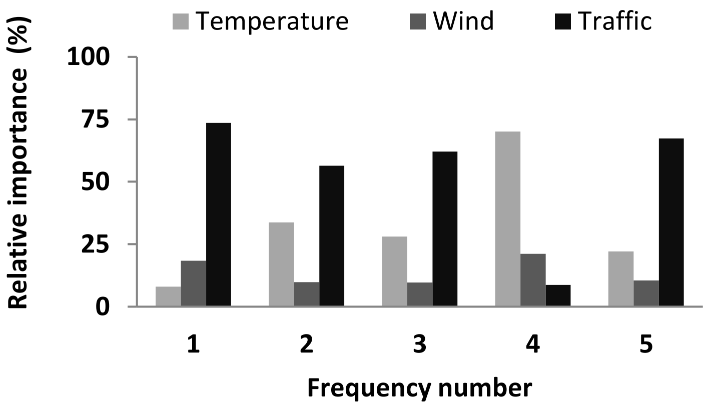


Figure 17. Evaluating simultaneous effects of temperature, wind, and traffic loading on the natural frequency responses through the relative importance of variables using the MLR method.

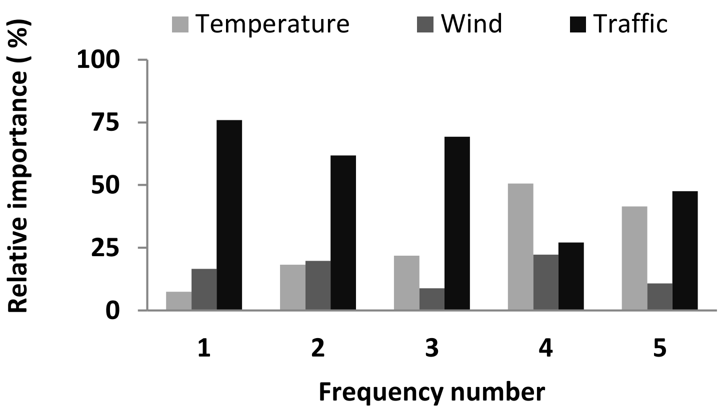


Figure 18. Evaluating simultaneous effects of temperature, wind, and traffic loading on the modal frequency responses through the relative importance of variables using the RF method.

Table 1. Parameters of measured natural frequencies of the bridge.

Mode number	Average frequency (Hz)	Frequency range (Hz)		Maximum difference (%)	Standard deviation (Hz)
		minimum	maximum		
1	0.39	0.38	0.41	8	0.00
2	0.47	0.41	0.57	34	0.01
3	0.60	0.58	0.61	5	0.01
4	0.69	0.67	0.70	4	0.00
5	0.73	0.71	0.75	6	0.01

Table 2. Results of the MLR method for the first five modes of the bridge

Frequency number	Mean squared error ($\times 10^{-6}$)	
	Training set	Test set
1	4.0	2.8
2	84.4	99.1
3	20.5	13.7
4	11.7	11.6
5	18.5	23.1

Table 3. Results of the ANN method for the first five modes of the bridge

Frequency number	Number of hidden nodes	Mean squared error ($\times 10^{-6}$)	
		Training set	Test set
1	11	3.9	2.7
2	10	87.6	95.7
3	21	22.8	12.3
4	33	15.7	18.6
5	21	19.5	20.0

Table 4. Results of the SVR method for the first five modes of the bridge

Frequency number	γ	α	Mean squared error ($\times 10^{-6}$)	
			Training set	Test set
1	20	0.8	3.7	2.5
2	9	0.17	69.5	96.7
3	20	0.56	16.6	11.8
4	12	0.28	9.4	10.6
5	14	0.36	15.7	18.6

Table 5. Results of the R_Tree method for the first five modes of the bridge

Frequency number	Number of terminal nodes	Mean squared error ($\times 10^{-6}$)	
		Training set	Test set
1	31	3.7	2.6
2	44	74.1	98.1
3	35	17.1	11.7
4	54	9.7	10.8
5	78	16.4	19.4

Table 6. Results of the RF method for the first five modes of the bridge

Frequency number	The optimal number of observations in terminal nodes	Mean squared error ($\times 10^{-6}$)	
		Training set	Test set
1	45	3.5	2.5
2	30	68.8	96.7
3	45	15.2	11.4
4	25	9.1	10.1
5	15	13.6	18.4

Table 7. Reduction in prediction errors of the ANN, SVR, R_Tree and RF methods when using the prediction error of the MLR method as a basis.

Frequency number	Error reduction (%)			
	ANN*	SVR	R_Tree	RF
1	6.5	10.8	9.8	12.9
2	3	2.4	1.1	2.4
3	10.4	14.2	14.9	17.0
4	8.6	8.5	6.9	12.9
5	13.5	19.4	15.8	20.1

(*) The error reduction for ANN when compared with the error of MLR, MSE_{MLR}, is equal to $\left(MSE_{ANN} - MSE_{MLR}\right) \times 100 / MSE_{MLR}$; the same formulation is also applied for SVR, R_Tree and RF.