The 3rd International Conference on Structural Health Monitoring of Intelligent Infrastructure
Vancouver, British Columbia, Canada
November 13-16, 2007

SYSTEM IDENTIFICATION: DATA MINING TO EXPLORE MULTIPLE MODELS

Prakash Kripakaran
Structural Engineering Institute, Switzerland

Sandro Saitta
Structural Engineering Institute, Switzerland

Suraj Ravindran
Structural Engineering Institute, Switzerland

Ian F. C. Smith
Structural Engineering Institute , Switzerland

Abstract

System identification involves identification of a behavioral model that best explains the measured behavior of a structure. Unlike traditional studies that focus on identifying parameters in a single model for system identification, this research uses a strategy of generation and iterative filtering of multiple candidate models. Robert Nicoud et al. [1] used stochastic search for generation of a set of candidate models that could represent the behavior of a structure. Exploring this often large set of candidate models without suitable computing tools is difficult. In this paper, a data mining based methodology is proposed to cluster models thereby providing information related to model distribution. An application of the methodology to a full-scale bridge is shown. From the distribution of models and their predictions, it is possible to identify the best location for the next measurement. Measuring at this location reduces the number of candidate models and this leads to more rapid identification of the correct model. The bridge example demonstrates how this is done.

INTRODUCTION

Sensor-based monitoring and diagnosis of structures is a rapidly evolving field in the domain of structural engineering. Advances in sensor technology have provided engineers with a wide variety of sensors that are capable of measuring various types of structural response. The task of interpreting these measurements, however, remains problematic for engineers. System identification [2] is the task of determining the state of the system from measurements. Research in system identification has focused on model updating or model calibration techniques [3, 4] that estimate values of unknown parameters of a mathematical model of the structure. These techniques are based on the assumption that the model that best fits measurements is the correct model. However, Robert-Nicoud et al. [1, 5] showed that this assumption may be false due to the presence of compensating errors in modeling and measurement. They proposed a strategy involving generation and iterative filtering of multiple candidate models for system identification.

Robert-Nicoud et al. [1, 5] used stochastic search to generate sets of candidate models. The objective function of the stochastic search is defined as the root mean square error between measurements and corresponding model predictions. It was proposed that models which have a root mean square error less than a certain threshold value, are equally capable of representing the measured structure. For a complex structure, large numbers of candidate models are generated by stochastic search. Engineers need computing tools to infer meaningful information from model sets. For instance, engineers may want to know where to take the next measurement to filter models most efficiently. This paper explores the support that data mining tools can provide to engineers during the system identification process.

Data mining methods have been successfully applied to tasks such as image recognition [6], speech processing [7] and web mining [8]. Many applications of data mining techniques to structural health monitoring also exist in literature. Posenato et al. [9] have used principal component analyses for detecting structural damage by analyzing time series data from sensors. Bulut et al [10] have employed statistical pattern recognition tools for vibration-based monitoring of structures. A key aspect of these data mining applications is that they detect damage by operating directly on measured data. However, multiple model system identification requires data mining techniques that can extract knowledge from candidate model sets generated by stochastic search.

In this paper, clustering techniques [11, 12] are proposed to extract useful knowledge from candidate model sets and assist with identification of subsequent sensor locations. The objective of clustering is to group together models according to a distance metric. The clustering algorithm identifies numbers of clusters in model sets and when possible generates clusters that are compact and well-separated. A new measurement at an appropriate location should eliminate the maximum number of models. An algorithm that finds this best location for the next measurement based on cluster information is presented. The methodology is illustrated for the Schwandbach bridge in Switzerland. Models generated using stochastic search are grouped using the clustering algorithm and the next location for measurement is identified. The proposed methodology is also compared with one that does not involve clustering.

MULTIPLE MODEL SYSTEM IDENTIFICATION

In conventional system identification [3], a suitable model is identified by matching measurement data with model predictions. This involves identifying values of model parameters that minimize the difference between predictions and measurements. These methods are based on the assumption that the model that best fits observations is the most reliable model. This assumption is flawed due to the following reasons: (1) system identification is an inverse problem and thus, multiple models can predict the same measurements, and (2) errors in modeling and measurement [1, 13, 14] may compensate such that the model that best predicts the measurements is not the correct model. Therefore, a strategy of generation and iterative filtering of multiple models is necessary for identification.

Figure 1 represents the framework for multiple model system identification [1, 5, 15]. The framework supports an iterative process that employs measurements for identification and then information from identification to improve the measurement system. The framework involves four modules: (1) model generation module, (2) data mining module, (3) measurement system design module and (4) engineer-computer interaction module.

Modeling assumptions and measurements from existing measurement system are provided by engineers. Given these, the model generation module generates candidate model sets for the data mining module. Modeling assumptions define the parameters for the identification problem. The set of model parameters may consist of quantities such as elastic constant, connection stiffness and moment of inertia. Each set of values for the model parameters corresponds to a model of the structure. The model generation module uses an objective function to evaluate the quality of candidate models. The objective function E is defined as follows.

$$E = \begin{cases} \varepsilon, \text{ if } \varepsilon > \tau \\ 0, \text{ if } \varepsilon \le \tau \end{cases} \quad \text{and } \varepsilon = \sqrt{\sum (m_i - p_i)^2} \qquad (1)$$

ε is the error which is calculated as the difference between predictions p_i and measurements m_i. τ is a threshold value evaluated from measurement and modeling errors in the identification process. The set of models that have $E = 0$ form the set of candidate models for the structure. Stochastic search [16] is used to generate candidate model sets.

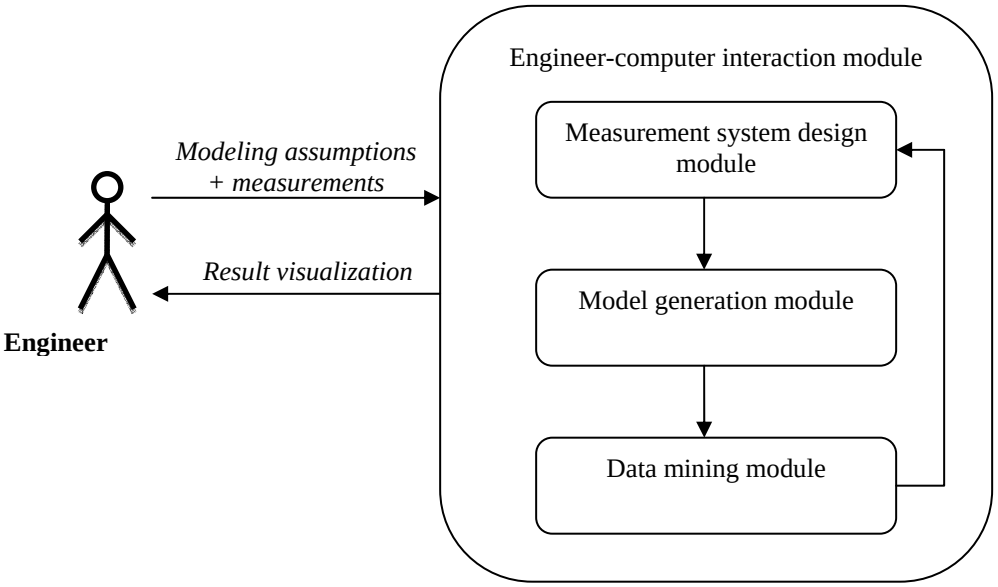

Figure 1. A framework of system identification tasks.

Generated model sets are examined by the data mining module. Data mining techniques are used to extract relationships between models and group similar models. This information is used by the measurement system design module to determine locations for subsequent measurements. The model generation module, data mining module and measurement system design module involve various degrees of engineer-computer interaction and this interaction is handled by the engineer-computer interaction module. For instance, this module has visualization tools for displaying results from the data mining module.

This paper focuses on the role of the data mining module in multiple model system identification. Specifically, the support that clustering techniques can provide for identifying the correct model for the structure is discussed.

DATA MINING MODULE

The correct model for the structure should be contained in the model sets given by model generation module. Clustering techniques aid in eliminating incorrect models from these model sets and thus rapidly converge to the correct model. To cluster models, a methodology that combines principal component analysis [17] and K-means clustering [12] is developed.

Principal Component Analysis (PCA)

PCA is a method for linearly transforming data in parameter space to a new and uncorrelated feature space [17]. In the machine learning community, PCA is usually used as a preprocessing technique, for example before a supervised learning algorithm. In this research, PCA is also used for visualization purposes. It is difficult to visualize clusters when clustering techniques such as K-means are directly applied to model sets of dimensionality greater than three. PCA finds a set of principal components (PC) that are sorted such that the first few components explain most of the variability in the model sets. By plotting the two first PC instead of two randomly chosen parameters, the clusters are easier to see.

The first step in evaluating the principal components of a data set is the construction of the covariance matrix *S*. The formula for evaluating *S* is given below.

$$S_{ij} = \text{cov}(x,y) = \sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y}) \quad (2)$$

S_{ij} represents element at the i^{th} row and j^{th} column of S. x and y are the i^{th} and j^{th} parameter and $\bar{x}$ and $\bar{y}$ are their respective means. N is the number of samples. The particular case of cov*(x, x)* corresponds to the variance of parameter x. After constructing S, its eigenvectors and eigenvalues are found (more details can be found in [17]). The principal components are obtained by sorting eigenvectors in decreasing order of their eigenvalues.

K-means Clustering

K-means [12] is a widely used clustering algorithm that is simple to understand and implement. However, it is effective only when applied and interpreted correctly. The K-means algorithm divides the data into K clusters according to a given distance measure. Although the Euclidean distance is often chosen as the distance measure, other metrics may be more appropriate in certain cases. The algorithm iterates over K clusters in order to minimize their intra-cluster distances, shown as the measure J in Equation 2:

$$J = \sum_{j=1}^{K} \sum_{x_i \in D_j} \left\| x_i - c_j \right\|^2 \quad (3)$$

K is the number of clusters, x_i is the i^{th} data point and c_j is the centroid of j^{th} cluster D_j. The K starting centroids are chosen randomly among all data points. The data set is then partitioned according to the minimum squared distance J. The cluster centers are updated by computing the mean of the points belonging to the clusters. The process of partitioning and updating is repeated until a stopping criterion is reached. The stopping criterion is attained if there is no significant change in the values of c_j or Equation 2 over two consecutive iterations of the algorithm.

The methodology for grouping models into clusters combines PCA and K-means. Model sets in parameter space are transformed using PCA into an uncorrelated feature space. Next the best number of clusters is estimated using a score function [18]. Once the number of clusters is known, K-means algorithm is applied to the data in feature space.

CLUSTERING FOR SYSTEM IDENTIFICATION

The purpose of system identification is to eliminate incorrect candidate models and converge to the correct model. Subsequent measurement locations that are found using cluster information help in achieving this goal. The proposed clustering procedure is completely integrated into the overall methodology for system identification. The methodology can be divided in two phases: i) original measurement system design and ii) periodic monitoring.

During the original measurement system design phase, engineers provide modeling assumptions that define parameters of the structure. Stochastic sampling is used to generate model sets. A global search [19] uses these model sets to determine the optimal number and position of sensors for the starting measurement system.

Figure 2 shows a flowchart with the methodology for system identification during the periodic monitoring phase. As in the initial measurement system design phase, engineers provide structural assumptions that define the parameter set for the problem.

The next step, *model generation*, creates a set of candidate models that may represent the real state of the structure using stochastic search. Measurements, a set of model parameters and an objective function (Equation 1) that defines candidate models are needed to generate the set of candidate models.

Once the models have been generated, the described *clustering* algorithm is used to group models. Models are grouped into clusters to i) facilitate visualization of the model space and ii) reduce the number of models given to the engineer (the centroid of the cluster is a possible representative model for the entire cluster). Visualization of

clusters is improved through the use of principal components. As described earlier, PCA is first applied to models before the K-means algorithm is used.

In the *representative model selection* step, a few models representing each cluster are selected. Only models which are close to the center of the cluster are selected. In this study, 5% of the total number of models in each cluster are taken to be representative models.

After clustering, entropy [20, 21] is used as a measure of model separability to identify the next measurement location. Entropy H_s of model sets at measurement location s is given by the following equation.

$$H_s = -\sum_i P_i . \log_2 P_i \qquad (4)$$

P_i is the probability of the i^{th} interval in the prediction distribution at location s. P_i is calculated as the ratio between the number of models that have predictions within the i^{th} interval and the total number of models. If model sets have high values of entropy, more candidate models can be filtered. Then the next step is sensor addition and further measurements. If the entropy of predictions is not significant (close to zero), then it is checked if there are multiple clusters. If true, it means that the current set of measurement locations is incapable of further filtering models. The engineer has to provide other measurement locations to the algorithm in order to find the correct model. If there is only one cluster and the entropy is close to zero, center of all remaining models is given to the engineer as the correct model for the structure (*model identification* step).

During the *sensor addition and further measurements* step, the entropies of selected representative models are used to find the position of the next sensor. The location with the highest entropy is chosen as the best position for the next measurement. Then, the measurement is taken on the structure.

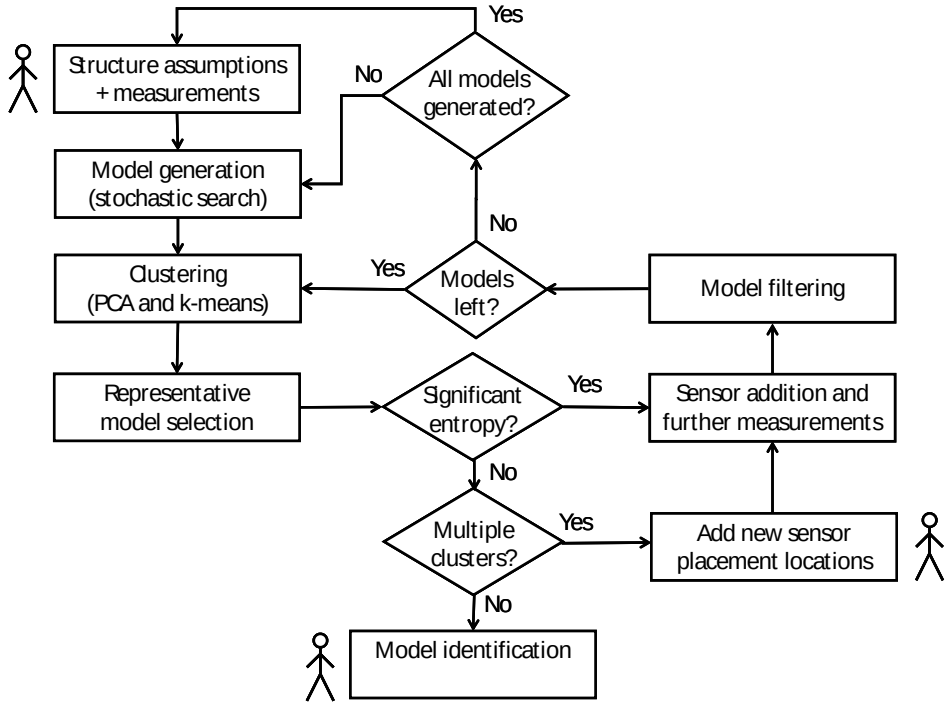

Figure 2. Flowchart showing the methodology for iterative sensor placement using multiple models. The human icon means that user interaction is required.

In the *model filtering step*, sensor measurement at the new location is compared for every candidate model. Candidate models that do not predict the measurement are eliminated from the current set of models. If there are models left, then the next step is *clustering*. However, if no models are left, then it is likely that all models were not generated by the model generation module. While it may be possible to generate all models for a simple problem, it is practically impossible to generate all possible models in a complex structure. In that case, the *model generation* phase is revisited. On the other hand, if all models have been generated, then some assumptions related to modeling the structure are incorrect. Therefore, *structure assumptions* have to be checked and modified by the engineer. This methodology is illustrated for the Schwandbach bridge in Switzerland.

SCHWANDBACH BRIDGE

The Schwandbach Bridge [22] (see Figure 3) designed by Maillart in 1933 is an early example of a deck stiffened open-spandrel arch. The elliptic horizontal ground-plan curve that is supported by a vertical curved thin-walled arch is also an example of daring structural engineering that has inspired engineers for over seventy years. The proposed methodology is demonstrated for identifying connection behavior of the Schwandbach bridge.

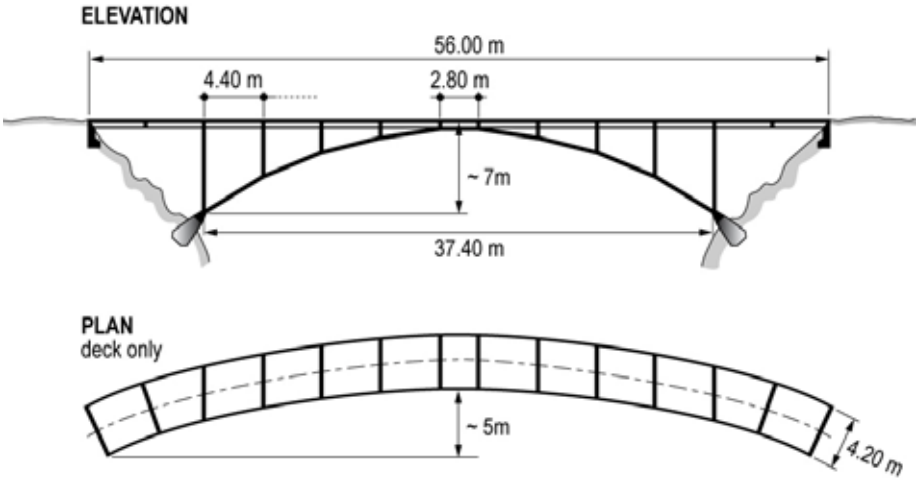

Figure 3. A schematic diagram of the Schwandbach Bridge (1933).

Connection behavior is difficult to understand in most structures. While they are modeled as fully hinged or fully rigid during design, it is well-known that connections exhibit semi-rigid behavior in practice. There are 20 connections in the Schwandbach bridge as shown in Figure 4. A finite element model of the bridge is created. Connections are modeled as rotational springs. To simulate sensor measurements, stiffness values are specified for all connections. This set of stiffness values is the correct model for the structure that should be found using system identification. A truck load test is simulated by modeling equivalent loads on the bridge model. Three load cases that represent two trucks at three different positions on the bridge are simulated (see Figure 5). Each truck has a front axle and rear axle loads of 17 kN and 44 kN respectively.

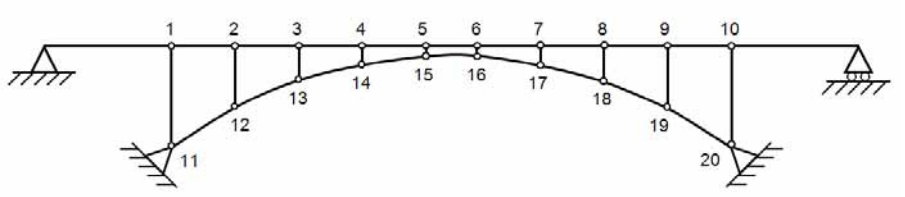

Figure 4. Modeling assumptions for the Schwandbach Bridge (1933).

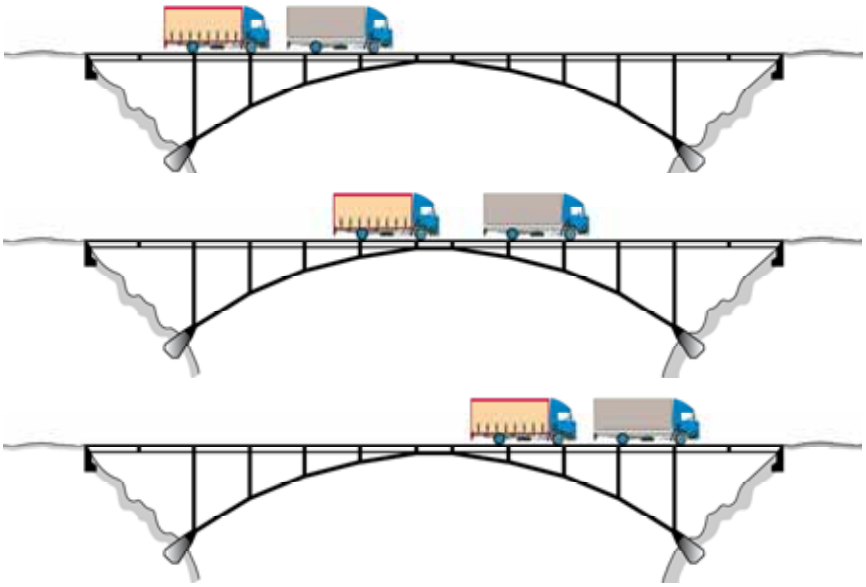

Figure 5. Truck positions on Schwanbach bridge for the three load cases.

The starting measurement system consists of inclinometers at the following locations – 1, 3, 6, 8, 10, 14 and 17. The inclinometers have a precision of 1 microradian. The rotations at the following locations for the three load cases are taken as the measurements from the inclinometers. These measurements are given as input to the model generation module. The spring stiffness values k_i of connection i are permitted to vary between 0 (hinged) and 10^8 (fully rigid). The model parameters are log k_i - the logarithms of the spring stiffness k_i. 1000 candidate models are generated using the model generation module for the data mining module.

RESULTS

This section focuses on the use of clustering results to iteratively add sensors on the structure. As stated previously, a score function is used to evaluate the number of clusters among models. The present model set contains 1000 rows (the models) and 20 columns (the parameter values for each model). Using the score function [18], the model sets are found to contain three clusters. This number can be visually estimated as well. For that, the models are plotted using the two first principal components as axes in Figure 6.

The three clusters present in Figure 6 already give the engineer an idea about the candidate model space. Since models in a cluster have similar values for parameters, they must also represent similar states of the structure. Instead of having to examine 1000 models, the engineer can examine the three groups of models, each represented by its center. The center of each cluster represents a bridge with a particular set of stiffness values for the connections. One cluster has a low value of stiffness for three connections. This may indicate cracking at these locations.

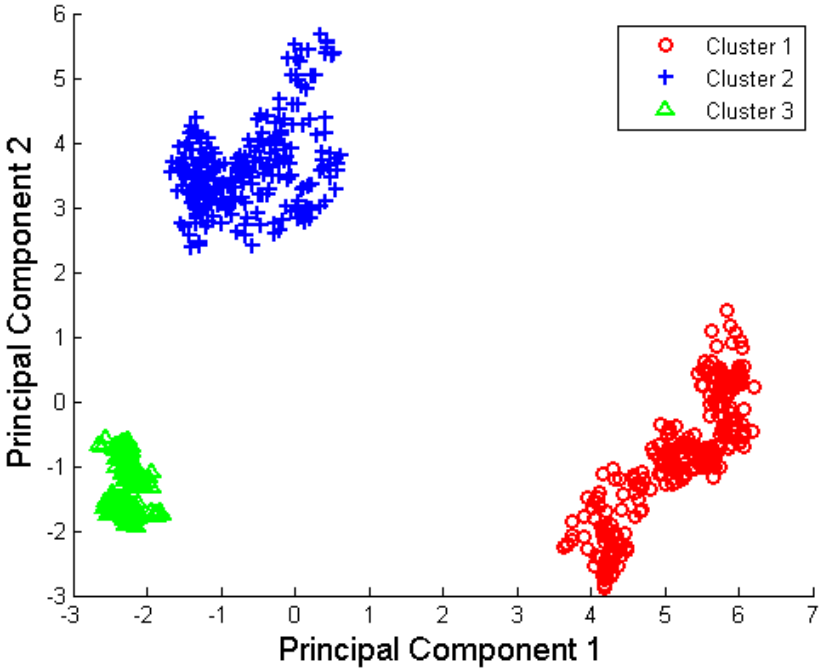

Figure 6. Model space after placing the seven original sensors. Three clusters are present. Axes are the two first principal components.

The next step is to iteratively add sensors to reduce the total number of models. The methodology is described by the flowchart in Figure 2. Representative models are selected in each cluster. Entropy is measured on these representative models and a new sensor is chosen where entropy is greatest. Once this sensor is known, a new measurement is taken. All models whose predictions do not match the new measurement are eliminated. This is repeated until the entropy of predictions is less than 1. At each iteration, the number of models is either reduced or the same.

The proposed strategy is compared with a strategy without clustering. Results using the proposed strategy are found to be better. Figure 7 shows a comparison of the two strategies on the basis of number of models remaining after each iterative measurement. The remaining number of candidate models is plotted versus the iteration number (i.e. the number of sensors placed). The starting conditions include seven sensors. Using the clustering strategy, the number of models decreases more rapidly. The reason for this behavior is that the strategy involving clustering finds better measurement locations. As seen from Figure 7, measurement from the location identified using clustering in the first cycle is able to filter as many models as the measurements from locations identified over two cycles using the strategy without clustering.

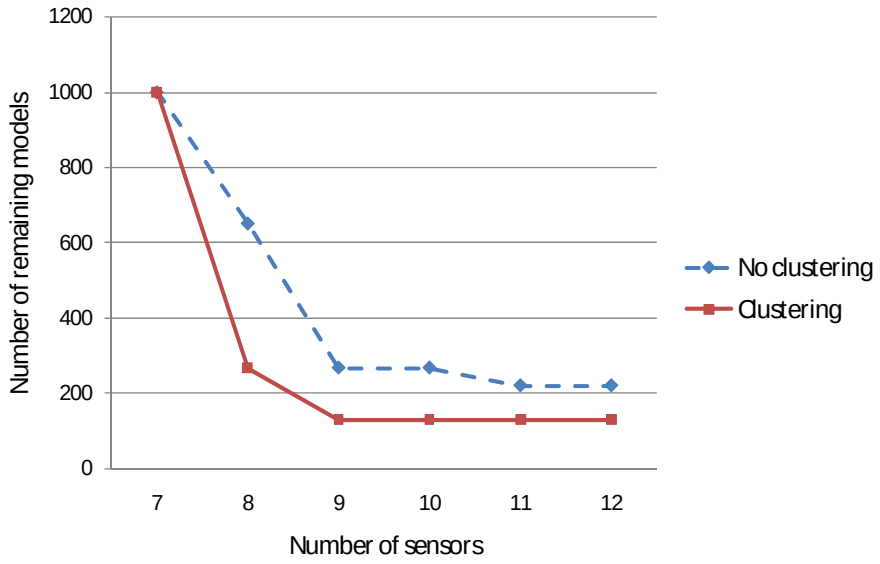

Figure 7. Comparison of the number of models left after sensor addition for methodologies with and without clustering.

After two sensor additions, the clustering strategy stabilizes. Addition of further sensors does not reduce the number of remaining models. The reason for such behavior is that entropy values of predictions at all measurement locations are low. Placing sensors at the given measurement locations cannot filter more models. However, there are multiple clusters present among the remaining models. This indicates that the correct model has still not been identified. The engineer has to choose new measurement locations and possibly, new sensor types to identify the correct model.

CONCLUSIONS

Conclusions from this study are the following.

1. Clustering is capable of organizing large numbers of candidate models generated by stochastic search into small numbers of clusters and thus, engineers may concentrate only on a smaller set of representative models from each cluster.
2. The information from clustering helps identify subsequent sensor placement locations that eliminate the highest number of models from the current set.

Several extensions to this work are in progress. Application of other clustering algorithms and especially fuzzy clustering is under study. Use of a different stopping criterion for the methodology is also being explored. Lastly, work is in progress for devising a standard way of estimating the number of representative models required from each cluster to identify subsequent measurement locations.

ACKNOWLEDGMENTS

The work described in the paper is funded by the Swiss National Science Foundation under contract no. 200020-109257. The authors would like to thank B. Raphael for discussions.

REFERENCES

1. Robert-Nicoud, Y., Raphael, B., and Smith, I. F. C., 'System Identification through Model Composition and Stochastic Search,' Journal of Computing in Civil Engineering (19)(3)(2005) 239-247.
2. Ljung, L. 'System Identification - Theory for the User,' Prentice Hall, (1999).
3. Friswell, M., and Motterhead, J. 'Finite Element Model Updating in Structural Dynamics,' Kluwer Academic Publishers, (1995).
4. Doebling, S. W., Farrar, C. R., and Prime, M. B., 'A Summary Review of Vibration-Based Damage Identification Methods,' The Shock and Vibration Digest (30)(2)(1998) 91-105.
5. Robert-Nicoud, Y., Raphael, B., Burdet, O., and Smith, I. F. C., 'Model Identification of Bridges Using Measurement Data,' Computer-Aided Civil and Infrastructure Engineering (20)(2)(2005) 118-131.
6. Dirk, W., and Thomas, R. 'Realtime Object Recognition Using Decision Tree Learning,' (2005).
7. Zhou, L., Shi, Y., Feng, J., and Sears, A., 'Data Mining for Detecting Errors in Dictation Speech Recognition,' Speech and Audio Processing, IEEE Transactions on (13)(5)(2005) 681-688.
8. Nasraoui, O., Krishnapuram, R., and Joshi, A. 'Relational clustering based on a new robust estimator with application to web mining,' In Proceedings of NAFIPS 99 New York, (1999) 705-709.
9. Posenato, D., Lanata, F., Inaudi, D., and Smith, I. F. C., 'Model free interpretation of monitoring data,' In Intelligent Computing in Engineering and Architecture, Lecture Notes in Artificial Intelligence, (2006) 529-533.
10. Bulut, A., Singh, A. K., Shin, P., Fountain, T., Jasso, H., Yan, L., and Elgamal, A. 'Real-time nondestructive structural health monitoring using support vector machines and wavelets,' In Advanced Sensor Technologies for Nondestructive Evaluation and Structural Health Monitoring SPIE, San Diego, CA, USA, (2005) 180-189.
11. Tan, P.-N., Steinbach, M., and Kumar, V. 'Introduction to Data Mining,' Addison Wesley, (2006).
12. Webb, A. 'Statistical Pattern Recognition,' Wiley, (2002).
13. Banan, M. R., Banan, M. R., and Hjelmstad, K. D., 'Parameter Estimation of Structures from Static Response. II: Numerical Simulation Studies,' Journal of Structural Engineering (120)(11)(1994) 3259-3283.
14. Sanayei, M., Imbaro, G. R., McClain, J. A. S., and Brown, L. C., 'Structural Model Updating Using Experimental Static Measurements,' Journal of Structural Engineering (123)(6)(1997) 792-798.
15. Saitta, S., Raphael, B., and Smith, I. F. C., 'Data mining techniques for improving the reliability of system identification,' Advanced Engineering Informatics (19)(4)(2005) 289-298.
16. Raphael, B., and Smith, I. F. C., 'A direct stochastic algorithm for global search,' Applied Mathematics and Computation (146)(2003) 729-758.
17. Jolliffe, I. 'Principal Component Analysis,' Springer, (2002).
18. Saitta, S., Raphael, B., and Smith, I. F. C. 'A Bounded Index for Cluster Validity,' In submitted to International Conference on Machine Learning and Data Mining Leipzig, Germany, (2007)
19. Saitta, S., Raphael, B., and Smith, I. F. C. 'Rational design of measurement systems using information science,' In IABSE Conference Budapest, (2006) 118-119.
20. Shannon, C., and Weaver, W. 'The Mathematical Theory of Communication,' University of Illinois Press, (1949).
21. Robert-Nicoud, Y., Raphael, B., and Smith, I. F. C., 'Configuration of measurement systems using Shannon's entropy function,' Computers & Structures (83)(8-9)(2005) 599-612.
22. Billington, D. 'Robert Maillart's Bridges,' Princeton University Press, (1979).