Saitta, S., Raphael, B. and Smith, I.F.C. "Data mining techniques for improving the reliability of system identification" Advanced Engineering Informatics, Vol 19, No 4, 2005, pp 289-298. doi:10.1016/j.aei.2005.07.005

DATA MINING TECHNIQUES FOR IMPROVING THE RELIABILITY OF SYSTEM IDENTIFICATION

S. Saitta, B. Raphael, and I.F.C. Smith

Ecole Polytechnique Fédérale de Lausanne (EPFL),
Applied computing and mechanics laboratory (IMAC),
CH-1015 Lausanne, Switzerland

{Sandro.Saitta, Benny.Raphael,Ian.Smith}@epfl.ch

Abstract: A system identification methodology that makes use of data mining techniques to improve the reliability of identification is presented in this paper. An important aspect of the methodology is the generation of a population of candidate models. Indications of the reliability of system identification are obtained through an examination of the characteristics of the population. Data mining techniques bring out model characteristics that are important.

1. INTRODUCTION

With the development of accurate and inexpensive sensors, interest in system (model) identification has grown. System identification [1] involves determining the state of a system and values of system parameters through comparisons of predicted and observed responses. Appropriate optimisation problems are formulated for the minimization of the difference between analytical predictions of models and measurements. In structural engineering, such procedures are generally known as model updating or model calibration. Friswell and Mottershead [2] provide a survey of model updating procedures using vibration measurements The growing interest in this area is demonstrated by the large number of papers that have been published recently ([3]-[17]). Most work involves computing sets of stiffness coefficients that help predict observed responses of structures.

Reliability of system identification has rarely been studied in previous research. Many thousands of models may predict responses that reasonably match observations. Therefore procedures that match measured and predicted responses might identify wrong models. Robert-Nicoud et al. [18],[19] developed a system identification methodology that takes these

factors into account. A key aspect of this methodology is the generation of a population of candidate solutions in the feasible domain whose objective function values lie below a threshold. An indication of the reliability of identification is obtained through an examination of the characteristics of the population. If all candidate models lie in a narrow well-defined region of the search space, the solution is likely to be unique and identification is reliable. On the other hand, if candidates are distributed among multiple clusters or if the parameter values of candidates vary widely, reliability of identification is poor. A simple semi-automatic feature extraction procedure was used to extract characteristics of the set of candidate models. Features such as the ranges of parameter values and the number of distinct classes of models are extracted with this procedure. The present work extends this methodology to include more sophisticated data mining techniques.

Data mining is an active research area. Even though several textbooks ([20],[21],[22]) and research publications ([23],[24]) on data mining have come out recently, mining model data is a novel concept. Data mining techniques have never been used for identifying characteristics of good models that explain observations. Features extracted using data mining provide indications related to the reliability of system identification and actions to be taken for improving the quality of system identification. This has never been attempted before.

The outline of the paper is as follows: Factors that affect the reliability of system identification are discussed in Section 2. Our methodology for system identification is presented in Section 3. An example of application of the methodology is provided in Section 4. Section 5 contains the conclusions.

2. Reliability of system identification

In system identification, models are selected through computing a distance function that evaluates the difference between predictions and observations. Model selection and calibration involves minimisation of the distance function through searching for appropriate values of model parameters. However, different types of errors could influence the results and accurate estimates of model parameters are difficult to obtain. In this study, **reliability of identification** is defined as the probability that the candidate model(s) obtained through system identification corresponds to reality. Reliability of system identification is poor when many models produce the same response at measured locations.

The degree of match between model predictions and measurements is evaluated using a distance function such as the one given below:

$$\text{Euclidean distance} = \sqrt{\sum_i (x_{i,c} - x_{i,m})^2} \quad (1)$$

where $x_{i,m}$ is the value measured at the i-th measurement point and $x_{i,c}$ is the corresponding value computed using the model. Model identification procedures minimize the distance function through searching for appropriate values of model parameters. If $x_{i,a}$ is the real value of the response at the i-th measurement point and $e_{i,meas}$ is the corresponding measurement error,

$$x_{i,a} = x_{i,m} + e_{i,meas} \quad (2)$$

Modeling error ([25]) consists of 4 components, these are due to mathematical modeling (differential equations) (e_1), numerical computation of solutions, for example, discretisation errors (e_2), wrong assumptions for example, related to boundary conditions (e_{3a}) and wrong values of parameters for example, values of moment of inertia and Young's modulus (e_{3b}). Therefore,

$$x_{i,a} = x_{i,c} + e_{i,1} + e_{i,2} + e_{i,3a} + e_{i,3b} \quad (3)$$

Substituting equations (2) and (3) in (1),

$$\text{Euclidean distance} = \sqrt{\sum_i (e_{i,meas} - e_{i,1} - e_{i,2} - e_{i,3a} - e_{i,3b})^2} \quad (4)$$

When there are modeling and measurement errors, correct values of model parameters result in a non-zero value of the distance function. It is often possible to reduce this value by using incorrect values of model parameters for example if ($e_{i,3b}$) is opposite in sign to the resultant of other terms in Equation (4). Thus, the location of the global minimum of the distance function is not close to the correct values. Therefore, models that are selected through the minimization of the distance function may not correspond to reality, thereby lowering the reliability of system identification.

The measurement system accuracy and configuration is often the factor that has the maximum influence on the reliability of system identification. System identification using more measurements is obviously more reliable than identification with a few measurements. Incorrect models might predict responses that match measurements exactly when only a few measurements are available. Nevertheless, when more measurements are added, it becomes increasingly difficult to obtain zero value of the cost function since model predictions may not match all measurement points.

3. System identification methodology

This methodology accounts for factors that influence the reliability of identification. A schematic of the process is shown in Fig. 1. Users input measurement data and specify a set of modelling assumptions. The model selection process identifies a set of candidate models whose predictions are close to measurements. A feature extraction module extracts characteristics of these models.

Four key modules in the methodology are global search, model composition, measurement system configuration and extraction of model characteristics. Details of these modules are described next.

3.1. Global search

A stochastic global search algorithm called PGSL ([26]) is used to minimise the cost function that evaluates the difference between measurements and model predictions. Search variables are assumptions that are needed to create complete models. These assumptions are related to the condition of the structure such as the presence of cracks and support settlements, as well as values of parameters such as Young's modulus and rigidity of joints. PGSL proposes values of variables which are used by the model composition module to create complete models. These models are analysed by the finite element method in order to compare the predicted responses with measured values. The cost function evaluates the degree of match between predicted and measured responses. The value of the cost function is used by PGSL to identify regions containing good solutions where more intensive search is carried out. At the end of each search, a model whose predictions have a good match with measurements is obtained.

3.2. Model composition

Compositional modelling is a framework for constructing adequate device models by composing model fragments selected from a model fragment library ([27]). Model fragments partially describe components and physical phenomena. A complete model is created by combining a set of fragments that are compatible. For modelling the behaviour of structures, fragments represent support conditions, material properties, geometric properties, nodes, elements and loading. Assumptions are explicitly represented in model fragments so that the model composition module

generates only valid models that are compatible with the assumptions chosen by users.

Model composition makes it possible to search for models containing varying number of degrees of freedom. There is no need to formulate an optimisation problem in which the number of variables is fixed a-priori. Models are automatically generated by combining model fragments and are analysed by the finite element method in order to compare their predictions with measurements.

3.3. Measurement system configuration

The methodology for measurement system configuration is described in detail in Robert-Nicoud et al. ([28]); key features are summarised here. The principal goal of measurement system configuration is to improve the reliability of identification. Reliability of identification is poor when many models predict the same responses at the locations of sensors. Therefore, locations and types of measurement devices are chosen such that there is maximum separation between candidate models. In this work, the degree of separation between models is measured using the entropy function defined by Shannon and Weaver ([29]). The entropy concept has been developed in the field of information theory and is a measure of "disorder" within a set. There is maximum disorder when models and parameter values show wide dispersion. Sensors are ideally placed where there are large variations in predicted responses of candidate models. Therefore, the location and type of measurement devices are chosen such that the entropy of the set of model predictions is the maximum.

3.4. Extracting model characteristics

When many different models are output by the system identification procedure, it is difficult to make conclusive diagnostic assessments. Two situations are encountered.

1. Candidate models belong to different classes having different sets of parameters (heterogeneous model set)
2. All candidate models have the same structure with the same set of variables and differ only in the values of continuous variables

The first situation is difficult to automate. Current data mining techniques are unable to accommodate data containing different sets of parameters. Development of techniques for mining heterogeneous data is a topic of

ongoing research. In the present work, a semi-automatic procedure is employed in the case of heterogeneous models. Users manually separate models into classes using their knowledge of important parameters. Ranges of values of model parameters within each class are then computed in order to assess the reliability of identification.

Data mining techniques are more effective in the second situation. Different types of patterns in model data can be discovered by appropriate data mining techniques. A few examples of situations where data mining helps to discover patterns in data are given in Fig. 2. Each dot corresponds to a model. The values of two model parameters x1 and x2 are plotted for each model. Four situations are shown. In the first case, all candidate models have nearly the same values for one of the parameters. In the second case, there is a non-linear relationship between x1 and x2. In the third case, models appear in three clusters separated from each other. In the final case, models are uniformly distributed throughout the space.

Three data mining techniques were evaluated for bringing out their potential for identifying common characteristics of good models. First, correlation measurement was used to examine if there are relations between model variables. Secondly, Principal Components Analysis (PCA) was used to check if some variables have a greater importance than others. Finally, decision trees were used for the same purpose. These techniques are described below.

Correlation is a measure of the association degree between two random variables. It is derived from the covariance measure and is given by:

$$\mathrm{corr}(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x) \cdot \mathrm{var}(y)}} \tag{5}$$

where *cov* is the covariance and *var* the variances of the specified variables. The correlation between two variables *x* and *y* thus corresponds to the link between them and is written as:

$$\mathrm{corr}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 \right]^{1/2}} \tag{6}$$

where n is the number of samples. The correlation varies between -1 and 1. These bounds are reached when the association between x and y is perfectly linear. If the correlation is zero, it means that the covariance is zero. When the correlation is zero the two variables are independent.

The idea of **PCA** is to generate a new set of variables called principal components that are linear combinations of the original variables. The goal of PCA is to find a system of principal components that are sorted in a manner that the first components can explain most of the data. To obtain the principal components, the covariance matrix S is constructed as follows:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \tag{7}$$

where s_{ij} is the covariance between the parameter p_i and p_j. The formula of the covariance corresponds to the numerator of equation 2. Note that the special cases s_{kk} are equal to the variance of k. The covariance matrix S can be written as:

$$S = VLV^T \tag{8}$$

where L is a diagonal matrix containing the eigenvalues of S and the column of V contains the eigenvectors of S (for more details see [30]). The principal components, which are linear combinations of the original variables, correspond to the eigenvectors of S and can be represented as an orthogonal basis for the new space of the data. The principal components are sorted in decreasing order according to how well they represent the variability of the data. Each sample is transformed to a new dimensional space defined by selected principal components. The goal is to reduce the number of dimensions by choosing only the first two or three principal components. Thus the original data are represented by a linear combination of the original parameters in a new and lower dimensional space.

Decision Tree is an inductive method ([25]) that can be used for regression or classification; here it is used only for classification. The classification task is defined as follows: Given a training set in which each sample contains attributes and a class value, find a hierarchy of rules (tree) that correctly classifies training data samples into their respective classes.

Classification of a data point is done by asking a sequence of questions based on the attributes. Each node of the tree is a question, each branch is an answer and each leaf corresponds to a classification.

In our case, attributes refer to parameters and the class to the boolean value whether the model is good or bad. As parameters have continuous values, questions in the tree are represented by inequalities. Therefore, nodes are questions such as "is the parameter p_i less than a certain value a". One of the most important steps in constructing a decision tree is selecting which attribute to test at each node in the tree. The methodology divides the current sample set into two child sets that are purer than the data in the parent set. The goal is to arrive at a set of leafs that contains either good or bad models. It is not possible to construct all possible trees and to select the one that best classifies the data since the number of different possible trees is exponential with respect to the mean number of possible values of attributes. Many different criteria can be chosen for selecting the best split at each node. The measure of node impurity chosen in this study is the Gini index of diversity ([31]).

3.5. Flowchart

The flowchart for the methodology is shown in Fig. 3. Users specify an initial set of possible hypotheses and potential sensor locations. The measurement system configuration module selects the best combination of sensors. Measurements from these sensors are used to identify a population of models by repeating search several times. The population size is specified by users. Characteristics of candidate models are studied in order to determine whether identification is reliable and whether models are physically possible. If the evaluation is not satisfactory, either assumptions are modified or the measurement system is improved by adding new sensors.

4. Evaluation of methodology

The methodology has been applied to case studies such as a timber beam supported on springs, a beam made of high strength concrete, the Lutrive bridge in Switzerland ([32]) and leak detection in the fresh water supply network of town Martigny in Switzerland. The state of the system has been correctly identified in all cases provided that enough number of sensors is used. Only one case study (timber beam) is described here. Two

experiments are presented. The first illustrates semi-automatic extraction of features. The second experiment illustrates advantages of data mining techniques for extracting features.

4.1. Experiment 1: Simple feature selection

A timber beam supported on springs has been constructed in the laboratory. Eight inductive sensors were used to measure vertical displacements at different locations. These were uniformly distributed over the length of the beam. Positions and magnitudes of applied loads along with characteristics of the structure, such as the material properties and support conditions, were treated as unknown variables (Fig. 4). The order of sensors suggested by the measurement system configuration methodology is given in Table 1.

Rank	1	2	3	4	5	6	7	8
Sensor	5	2	7	3	6	4	1	8

Table 1 The order of sensors for maximum entropy (from [18])

Loads were applied on the structure and measurements were taken from all the sensors in order to test the system identification methodology. Three measurement system configurations were considered. In the first configuration, measurement from a single sensor (sensor 5) was used to identify models. The set of models whose predicted responses were close to the measured value included the correct model as well as those that involved wrong support conditions and loading. The deflections measured by sensors and the corresponding ranges of values predicted by candidate models are shown in Table 2. All models match the deflection at the location of sensor 5, but differ significantly at sensor locations that were not used in model identification. Values predicted by candidate models show a wide dispersion at the location of sensors that were not used in identification.

Sensor	Measured value (m)	Predicted range (m) min	max
1	-0.00017	-0.00587	-0.00005
2	-0.00103	-0.00462	-0.00038
3	-0.00165	-0.00359	-0.00077
4	-0.00211	-0.00299	-0.00125
5	-0.00219	-0.00219	-0.00219
6	-0.00178	-0.00334	-0.00145
7	-0.00103	-0.00464	-0.00075
8	-0.00011	-0.00601	-0.00008

Table 2 When only sensor 5 was used to identify models

Sensor	Measured value (m)	Predicted range (m) min	max
1	-0.00017	-0.00016	-0.00013
2	-0.00103	-0.00114	-0.00091
3	-0.00165	-0.00188	-0.00160
4	-0.00211	-0.00227	-0.00203
5	-0.00219	-0.00230	-0.00210
6	-0.00178	-0.00186	-0.00170
7	-0.00103	-0.00108	-0.00099
8	-0.00011	-0.00012	-0.00011

Table 3 When three sensors (2,5,7) were used to identify models

In the second measurement system configuration, three sensors were used. The candidate models reasonably matched measurements at all sensor locations including those that were not used in system identification (Table 3). Two classes of models were obtained (Table 4). In the first model class, there is a single point-load and a settlement at the second support. In the second model class, there are two point-loads without any support settlement. This class corresponds to the real situation as tested in the laboratory. The results were not significantly different using a third measurement system configuration using 8 sensors.

	Parameter						
Type	X_1	F_1	X_2	F_2	Δ_1	Δ_2	Δ_3
Real structure	✓	✓	✓	✓			
Identification with measurement system consisting of 3 sensors							
Model Class 1	✓	✓				✓	
Model Class 2	✓	✓	✓	✓			

Table 4 Classes of models obtained through the second measurement system configuration. The symbol ✓ indicates that the parameter belongs to the model.

The application of the methodology resulted in a set of candidate models which included those that correctly modelled the state of the system. Three sensors are sufficient to reduce the number of model classes to two. These sensor locations have been obtained using the methodology for measurement system configuration that is described in Section 3.3. The envelope of predicted responses of candidate models contains independent measurements that were not used for system identification. This is an indication of the validity of candidate models.

4.2. Experiment 2: Feature selection using data mining techniques

The timber beam supported on springs described in Experiment 1 is also used in this study. Instead of a two-point load, a single point load is applied on Node 10 which is located near the middle of the left span. Only models containing single point loads were searched for candidate models. Measurements taken from three sensors are used and the mean square error between model predictions and measurements was minimised by PGSL. All the models generated by PGSL were saved for data mining. Most good models (candidates) that are identified within a PGSL run are located near the best solution found in that run. Therefore multiple PGSL runs are carried out in order to obtain good models in different parts of the search space. By changing the parameters of the PGSL algorithm, it is also possible to perform a uniform and pure random search over the entire search space. The term, random run, is used in this paper to denote this procedure. Random run does not always result in the identification of good models. The term focused run will be used to denote a normal PGSL search that converges to a good solution.

The input attributes of the data set are the following.

- p_1 Load position (Node number)
- p_2 Load magnitude (kN)
- p_3 Axial stiffness of the spring at the mid-span (kN/m)
- p_4 Rotational stiffness of the spring at the mid-span (kNm)

Data set used for the decision tree algorithm consists of an additional attribute. That is, the output Boolean attribute that indicates whether the model is good or bad according to the mean square error function. Results of applying the three data mining techniques are discussed below.

4.2.1. Correlation

With correlation, we examine how two parameters p_i and p_j are related in good models. For achieving this, only good models are selected from the case study and a correlation matrix is constructed in which each element is computed as explained in Section 3.4. Each row and each column of the matrix correspond to one of the four parameters. For example, the element (2, 3) represents the correlation between p_2 and p_3. The correlation matrix is symmetric about its diagonal since correlation is a commutative operation.

The correlation matrix was computed for several PGSL runs, both random and focused. It was found that the correlations between parameters in a random run are different with the ones in focused runs. In the first case, there is no significant correlation between parameters. In the second case, correlations between certain parameters are observed, See Table 5.

	p_1	p_2	p_3	p_4
p_1	1.0000	0.0000	-0.0000	0.0000
p_2	0.0000	1.0000	0.4682	-0.5541
p_3	-0.0000	0.4682	1.0000	-0.3109
p_4	0.0000	-0.5541	-0.3109	1.0000

Table 5 Correlation matrix for four parameters in a focused run.

If two parameters have a high degree of correlation, for example, greater than 0.5, it is assumed that there is a relationship between them. It was found that correlation values change from one PGSL run to the next. This result is due to the manner in which PGSL generates models. Nevertheless, two results are important. First, the first parameter (i.e. the

load position) is not correlated with any other parameter since the first column of the correlation matrix is zero (except for the first element). This means that the load position is an independent parameter for system identification. This is because the position of the load is always the same in all good models. A good match between predictions and measurements is not obtained if the load is not in the correct position. Therefore, the load position is a parameter that can be estimated reliably using this system identification methodology.

The second important result concerns the second parameter, the load magnitude. The load magnitude varies significantly among good models in multiple PGSL runs. It is also found that the correlations between this parameter and others vary between different runs. Nevertheless, there are always high values in the second column of the correlation matrix. This implies that the load magnitude has strong correlations with other parameters. Therefore, the load magnitude cannot be estimated independently of other parameters. Different combinations of the load magnitude and other parameters could result in the same degree of match with measurements.

4.2.2. Principal Components Analysis (PCA)

In this study, PCA is used as a "weighting method" that gives an indication of the relative importance of different parameters for determining the characteristics of the set of good models. By examining the principal components, a measure of the importance of each parameter for explaining variations in the data is obtained. Similarly to the correlation measurement, the PCA is an unsupervised data mining method. Only good models are used in the analysis.

After launching PCA on this data the first three principal components are examined. The first two components explain more than 98% of the variations in the model data as shown in Fig. 5. Approximately 78% of the candidate models differ in the value of the first principal component. 20% of the models differ in the value of the second principal component. There is no significant variation in the values of remaining components.

Instead of using the original model parameters, new variables c1, c2, c3 and c4 which contain the values of the principal components in each model are introduced. Using these new variables, a common characteristic of all good models is that the values of variables, c3 and c4 are nearly constant.

Models differ mostly in the values of c1 and c2. It can be seen that the standard goal of PCA has been achieved, since a few components explain most variations in data. For example, a two-dimensional plot (i.e. using only two variables instead of four), show almost all the variability in the data.

When c1 and c2 are plotted in a two dimensional graph, a number of small clusters are observed. It is speculated that this is due to the manner in which PGSL generates models. Since PGSL performs focused search in regions where good solutions are found, when PGSL finds a good model it generates many models in the neighbourhood, and this gives rise to these small clusters. This observation is valid only for focused runs. In the case of a random run, we see only a cloud of models. In this case, PCA cannot group the models since there is no correlation between them as indicated by the correlation measurement. The first three principal components of a sample focused run are given in Table 6.

	c_1	c_2	c_3	c_4
p_1	-0.0000	-0.0000	-0.0000	1.0000
p_2	0.0007	-0.0402	-0.9992	-0.0000
p_3	0.5855	0.8100	-0.0322	-0.0000
p_4	0.8107	-0.5850	0.0241	0.0000

Table 6 The first 3 principal components of a focused run ordered according to their ability in explaining the variability of the data.

Each column contains the coefficients of the original parameters in the linear equation used to compute a principal component. For example, coefficients in the first column should be multiplied by the respective values of original parameters and summed up in order to calculate the value of the new variable c1. In other words, each coefficient is a weight factor that represents the importance of the original parameter in the new dimensional space. The first coefficient is zero for all principal components. This means that the first parameter, the load position, has no influence on the variability of the data. This is because the load position is always the same for good models. This is not the case with other parameters, they vary among good models. However, their coefficients are quite different for two different focused runs. This is because although two focused runs find the same value for the load position, different values are found for other parameters. It can

be shown that the load position can be estimated reliability by plotting the Mean Square Error (MSE) between the model and the measurements against the load position for all the models (Fig. 6).

Each point in the plot corresponds to a model. All the models that have a low MSE have the load on node 10. However, all models that have a load on node 10 need not have a low MSE because the values of other parameters might be wrong. Therefore, a necessary but not sufficient condition for a good model is that the load is on node 10. Therefore, this plot shows that identification is good with respect to the load position. This is not the case with other parameters except the load magnitude as explained in next section. This is evident from the plot of p_3 shown in Fig. 7. For parameter p_3 good models are widely spread out. This parameter cannot be estimated reliably by system identification.

4.2.3. Decision Tree

The primary objective of the decision tree in this study is to assess the importance of parameters in separating good and bad models. A classification tree is created as described in section 3.4. The input is a matrix containing values of parameters for each model with the last column containing a value that indicates whether it is a good or bad model. This is the only method of the three where all the models are used (both bad and good). As with other studies (correlation and PCA), different focused runs gave different results. Fig. 8 shows a tree based on one focused run.

Even though results depend on the run, two conclusions can be made. First, (as confirmed by correlation and PCA), the load position is clearly identifiable. For example, the position must be between 9.5 and 10.5 for the model to be good. The second important result is the importance of the load magnitude. Here, the tree indicates that for a good model the load magnitude must be less than 0.05. This result can be verified by plotting the MSE with respect to the load magnitude (Fig. 9).

4.2.4. Discussion

Correlation brings out information related to the reliability of identification of parameters. For example, the load position is clearly

detected as a reliable parameter for identification. Correlations are different from one run to another. This means that we cannot bring out all relationships between two parameters with correlation. The correlation measure can only accommodate linear relationships between two parameters and there are certainly non-linear relationships between parameters. Another limitation is that we cannot obtain relationships between more than two parameters at a time. This last limitation can be overcome by the principal components analysis.

Two results are obtained using PCA. First, some parameters have more importance then others. This is seen in the coefficients of the principal components (PC). More important parameters have higher values of coefficients than the others. Secondly, the coefficient for the load position is always zero. This does not mean that the parameter p_1 has no influence on separating good and bad models. On the contrary, this means that every good model has the same value for the location of the load. In other words, p_1 is a reliable parameter in identification. This result is confirmed by the two-dimensional plot of the MSE. PCA brings out the fact that there are relationships between parameters of good models. However, it is difficult to determine the exact relationship. Since PCA is a linear data mining method, it is not able to bring out non-linear relationships in the data.

Decision trees provide new information related to the data and confirms others. The new information that was not provided by the correlation measurement or PCA, is the importance of the load magnitude. The first parameter that best separates good and bad models is the load magnitude p_2. For different focused runs, the trees are different. However, two aspects always remain the same. First p_1 and p_2 are always in the tree and second, the load magnitude is always at the root node. The fact that p_2 is an important parameter for system identification - and is therefore reliable - has been brought out only with the decision tree technique. Therefore, decision tree has definite advantages over other techniques. One of the strengths of decision tree is that they generate easily understandable rules. The tree thus brings out meaning of data. The limitation of decision tree is that the method does not perform well when combinations of variables (in the form of linear or non-linear relationships) determine the classes of data points.

5. Concluding remarks

Conclusions from this study are the following

- The system identification methodology that has been developed is able to identify candidate models as well as provide indications related to the reliability of identification.
- Data mining techniques are useful for bringing out common characteristics of the set of good models.
- If there are independent variables whose values can be uniquely identified, they are spotted by the correlation measurement. Non-diagonal terms are nearly zero for these variables.
- The first principal components in PCA consist of independent variables whose values are identified. A few principal components are sufficient for explaining the variation in data.
- Decision trees bring out variables that separate good and bad models.
- All the three techniques namely, correlation, PCA and decision trees, are unable to bring out non-linear relationships between model variables.

The methodology that is described in this paper has the potential to be applied to domains outside structural engineering. The methodology is already proving to be a valuable tool for engineers who are involved in the task of monitoring and maintenance of engineering systems.

ACKNOWLEDGEMENTS

This research is funded by the Swiss National Science Foundation (NSF).

References

[1] Ljung L., System Identification - Theory For the User, Prentice Hall, 1999.

[2] Friswell M.I. and Mottershead J.E., Finite Element Model Updating in Structural Dynamics, Kluwer, New York, 1995.

[3] Brownjohn JMW, Xia P.Q., Hao H., Xia Y., Civil structure condition assessment by FE model updating: methodology and case studies. *Finite elements in analysis and design, 37*, pp. 761-775, 2001.

[4] Koh C.G., Chen Y.F., Liaw C.-Y., A hybrid computational strategy for identification of structural parameters, *Computers and Structures*, 81, pp. 107–117, 2003.

[5] Castello D.A., Stutz L.T., Rochinha F.A., A structural defect identification approach based on a continuum damage model, *Computers and structures*, 80, pp. 417-436, 2002.

[6] Teughels A., Maeck J., Roeck G., Damage assessment by FE model updating using damage functions, Computers and structures, 80, pp. 1869-1879, 2002.

[7] Modak S.V., Kundra T.K., Nakra B.C., Comparative study of model updating studies using simulated experimental data, *Computers and Structures*, 80, pp. 437-447, 2002.

[8] Hemez F.M., Doebling S.W., Review and assessment of model updating for non-linear, tansient dynamics, *Mechanical Systems and Signal Processing*, Vol 15(1), pp.45-74, 2001.

[9] Sohn H., and Law K.H., "Damage Diagnosis using Experimental Ritz Vectors," *Journal of Engineering Mechanics*, ASCE, Vol. 127, No. 11, pp. 1184-1193, 2001.

[10] Hu N., Wang X., Fukunaga H., Yao Z.H., Zhang H.X. and Wu Z.S., Damage Assessment of Structures Using Modal Test Data, *International Journal of Solids and Structures*, 38, pp. 3111-3126, 2001.

[11] Beck J.L. and Katafygiotis L.S., Updating Models and Their Uncertainties. I: Bayesian statistical framework, Journal of Engineering. Mechanics. 124, pp. 455-461, 1998.

[12] Katafygiotis L.S. and. Beck J.L., Updating Models and Their Uncertainties. II: Model Identifiability, J. Engrg. Mech. 124, 463, 1998.

[13] Chaudhary M.T.A., Abe M, Fujino Y., and Yoshida J., Performance Evaluation of two Base-Isolated Bridges using Seismic Data, *Journal of Structural Engineering*, ASCE, Vol.116, No.10, pp.1187-1195, 2000.

[14] Abe M., Vibration control of structures with closely spaced frequencies by a single actuator, *Journal of Vibration and Acoustics*, Transactions of the American Society of Mechanical Engineers, Vol.120, pp.117-124, 1998.

[15] Reich, G. W. and Park, K. C., A Theory for Strain-Based Structural System Identification, *Journal of Applied Mechanics*, 68(4), 521-527, 2001.

[16] Park K.C. and Felippa, C.A., A flexibility-based inverse algorithm for identification of structural joint properties, In Proceedings of ASME symposium on computational methods on inverse problems, 15-20 November 1998, Anaheim, CA, 2001.

[17] Yu L., Law S.S., Link M., Zhang L.M., Damage Detection in Bolted Joint Structures using Element Contribution to Modal Strain Energy, In Proceedings of the Second International Conference on *Identification in Engineering Systems*, Swansea, M.I. Friswell, J.E. Mottershead and A.W. Lees (eds.), pp. 516-526, 1999.

[18] Robert-Nicoud Y., Raphael B., Smith IFC., (2005). System Identification through model composition and stochastic search, In print, ASCE, Journal of computing in civil engineering.

[19] Robert-Nicoud Y., (2003). Une méthodologie mesures-modèles pour l'identification de systèmes de génie civil, PhD. thesis, EPFL, Lausanne.

[20] Webb A.R., Statistical pattern recognition, John Wiley, 2002.

[21] Larose D.T., Discovering knowledge in data: an introduction to data mining, Wiley-Interscience, 2005.

[22] Hand D. Manila H. and Smyth P., Principles of data mining, MIT Press, 2001.

[23] Banks D., Classification, clustering, and data mining applications : proceedings of the meeting of the International Federation of Classification

Societies (IFCS), Illinois Institute of Technology, Chicago, 15 - 18 July 2004 / David Banks editor.

[24] Ghosh A., Evolutionary computation in data mining, Studies in fuzziness and soft computing, vol. 163. 2005.

[25] Raphael B. and Smith I.F.C, Fundamentals of computer aided engineering, John Wiley, UK,. July, 2003a.

[26] Raphael B. and Smith I.F.C. (2003), "A direct stochastic algorithm for global search", J of Applied Mathematics and Computation, Vol 146, No 2-3, pp 729-758.

[27] Falkenhainer B., Forbus K.D. (1991). Compositional modeling: Finding the right model for the job, Artificial Intelligence, vol. 51, pp. 95-143.

[28] Robert-Nicoud Y., Raphael B., Smith IFC., (2004). Configuration of measurement systems using Shannon's entropy function, Computers and structures, 83, pp. 599–612, 2005.

[29] Shannon C. and Weaver W., The Mathematical Theory of Communication, University of Illinois Press, 1949.

[30] Jackson J.E., A user's guide to principal components. *Wiley series in probability and mathematical statistics*, 1991.

[31] Breiman L., Friedman J.H., Olshen R.A. and Stone C.J., Classification and regression trees. Wadsworth International Group, Belmont, California, 1984.

[32] Robert-Nicoud Y, Raphael B, Burdet O, Smith IFC. Model identification of bridges using measurement data. J Comput Aided Civil Infrastruct Eng, Vol 20 (2005) pp. 118–131.

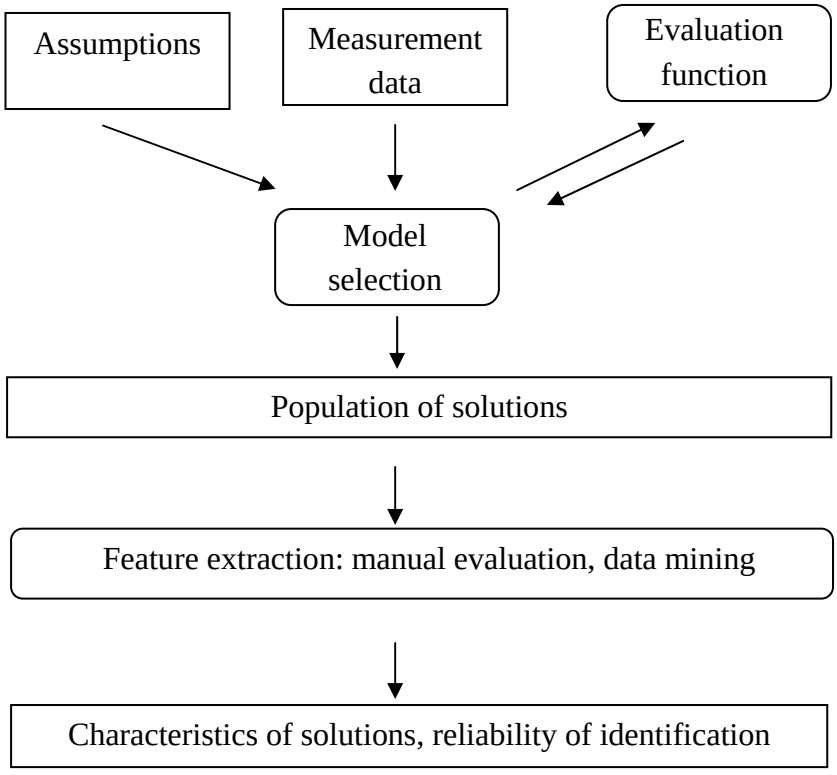

Fig. 1 A system identification methodology using a population of candidate models (adapted from [18]). Rounded rectangles indicate procedures; ordinary rectangles indicate data.

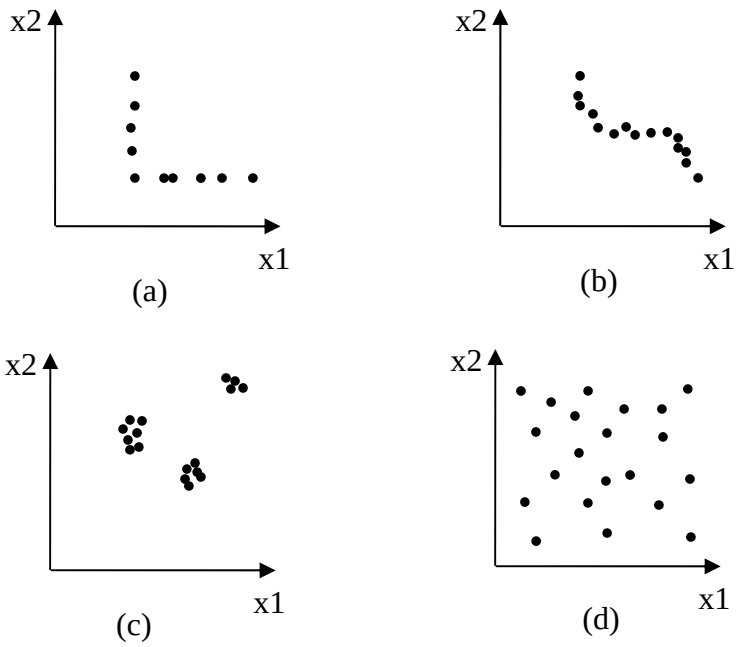

Fig. 2 Examples of situations where data mining helps to discover patterns in data.

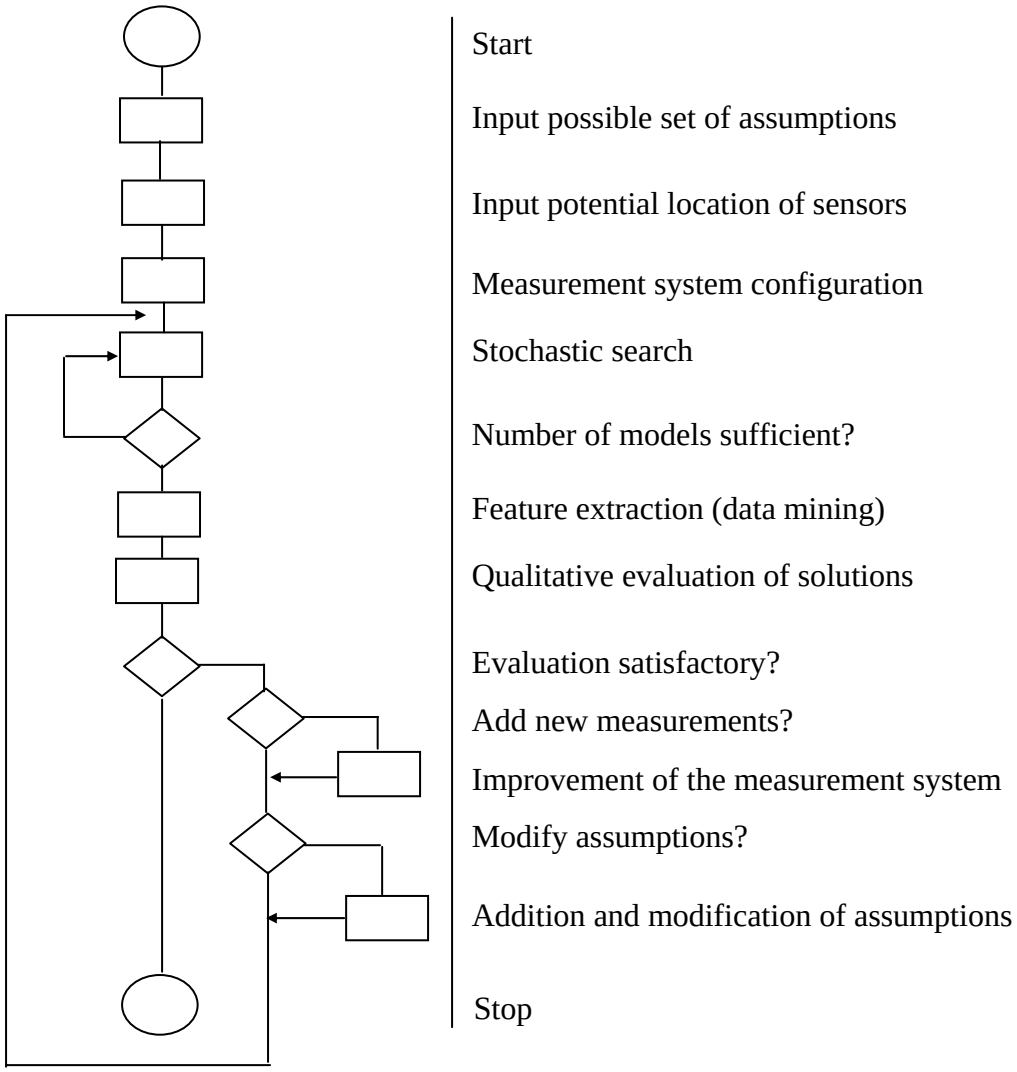

Fig. 3 Flowchart of the methodology (adapted from [18])

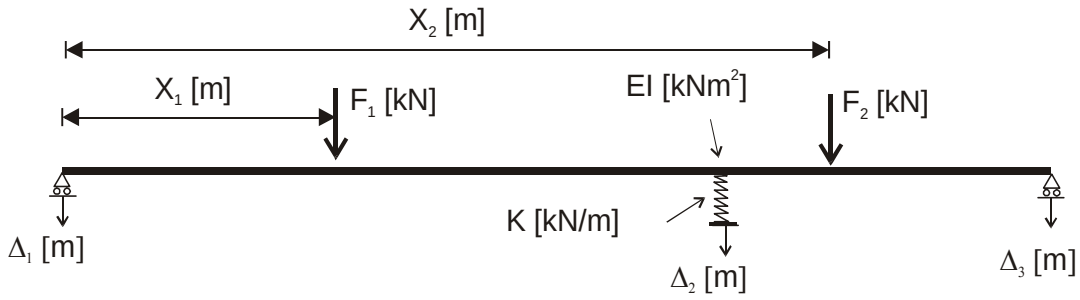

Fig. 4 Variables in the system identification

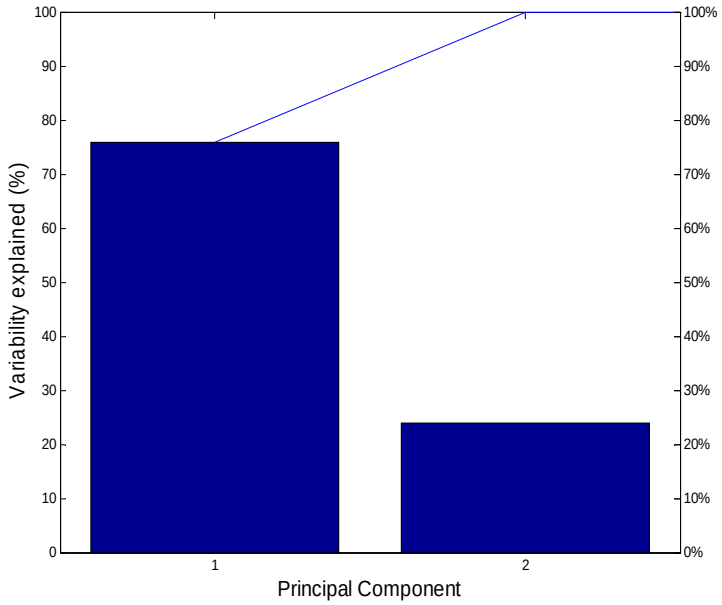

Fig. 5 The first two components explain more than 80% of the variations in the model data

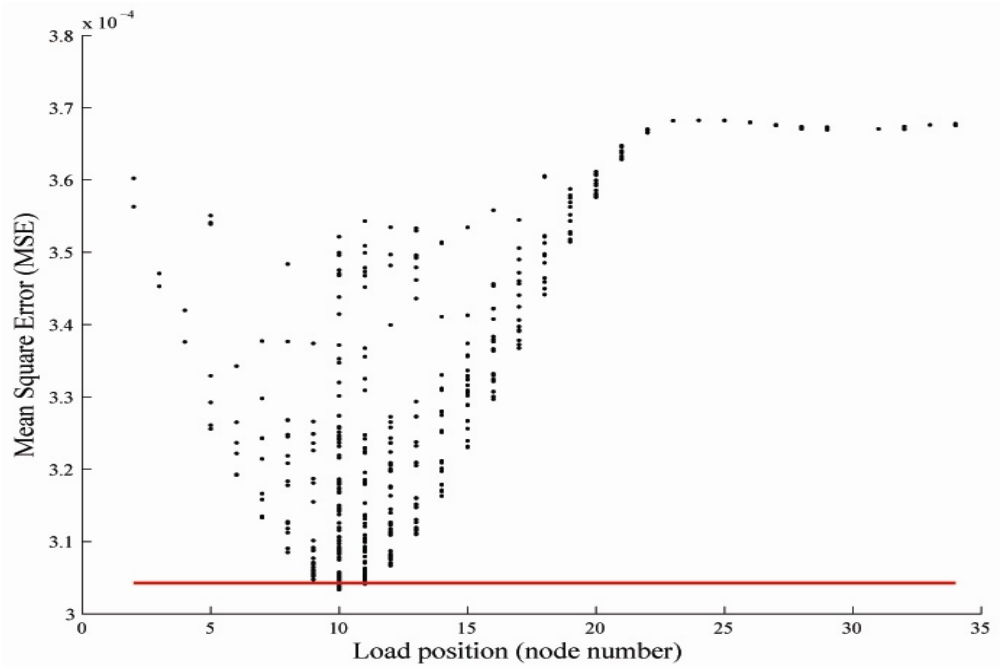

Fig. 6 Plot of mean square error versus load position for the case of timber beam supported on springs. All good models contain the load at node 10.

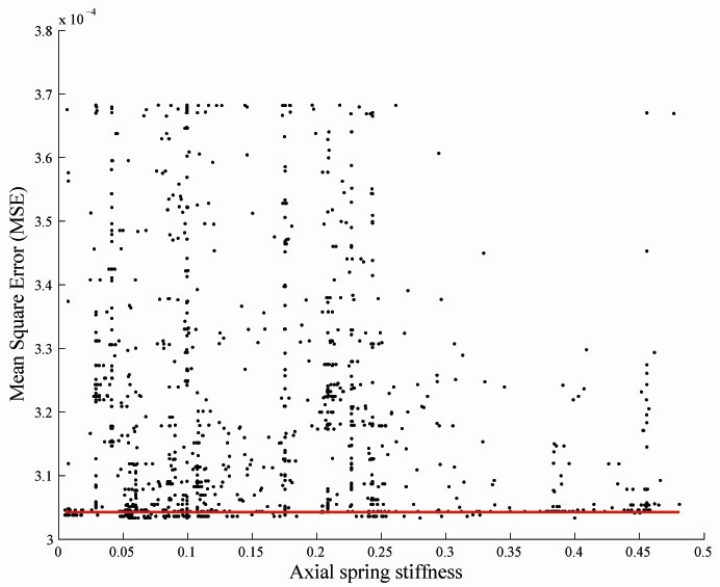

Fig. 7 Plot of mean square error versus spring stiffness for the case of timber beam supported on springs.

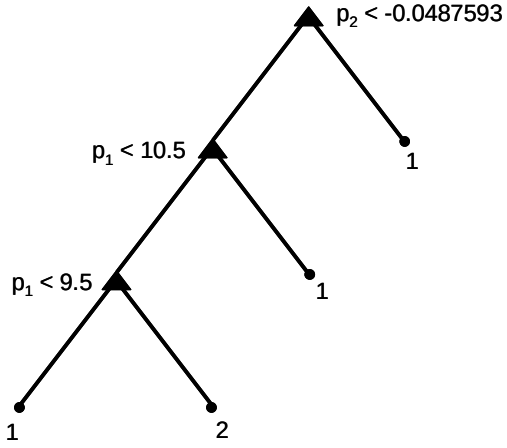

Fig. 8 A decision tree constructed on data from a focus run. The number 1 means a set of bad models and 2 a good set.

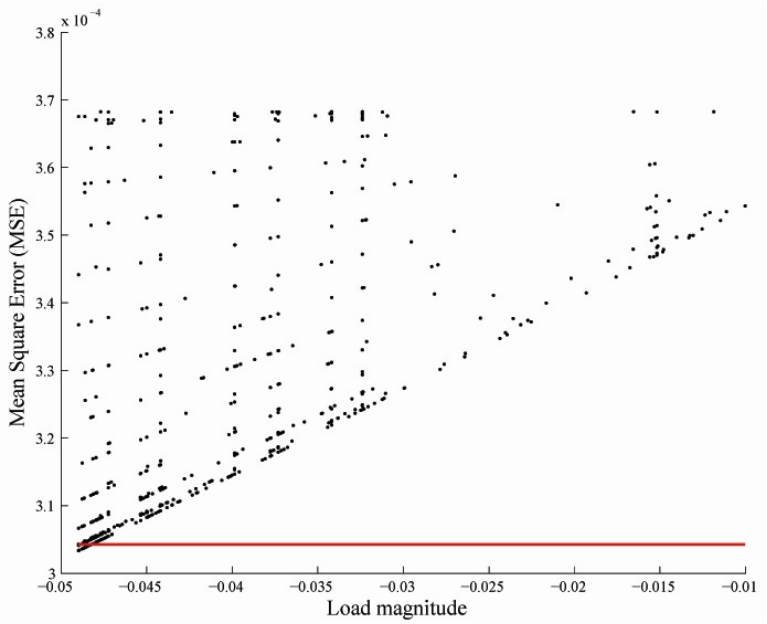

Fig. 9 Mean Square Error (MSE) of a focused run in comparison with p_2, i.e. the load magnitude. Each point on the plot is a model. The line shows the threshold for good models.