Proverbio, M., Costa, A. and Smith, I.F.C. "Sensor data interpretation with clustering for interactive asset management of urban systems" Journal of Computing in Civil Engineering, Vol. 32, No. 6, 2018.
https://doi.org/10.1061/(ASCE)CP.1943-5487.0000793 http://cedb.asce.org

Sensor data interpretation with clustering for interactive asset-management of urban systems

Marco Proverbio[1]

Alberto Costa[2]

Ian F.C. Smith[3], F. ASCE

ABSTRACT

In responsive cities, user feedback and information provided by sensors are combined to improve urban design and to support asset managers in performing decision making. Optimal management of infrastructure networks requires accurate knowledge of current asset conditions, in order to avoid unnecessary replacement and expensive interventions when cheaper and more sustainable alternatives are available. Structural model updating is a discipline that focuses on improving behaviour-model accuracy by means of measurements taken from the built environment. Error-domain model falsification (EDMF) is a simple and practice-oriented methodology that employs measurements at sensor locations to identify plausible models among an initial population that is generated according to engineering judgment. However, many plausible models are often identified, making result interpretations difficult for practising engineers. In this paper, a clustering methodology based on bipartite-modularity optimisation (BMO) is employed to clarify identification outputs. Compared with classical clustering methods such as K-means, BMO clustering provides more accurate interpretations and better visualization of the results. Moreover, engineers can actively interact with the clustering framework to obtain the knowledge that is needed at several stages of the decision-making process.

[1]Future Cities Laboratory, Singapore-ETH Centre, ETH Zurich, CREATE Tower, Singapore 138602; Applied Computing and Mechanics Laboratory, School of Architecture, Civil and Environmental Engineering, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland (corresponding author). E-mail: marco.proverbio@epfl.ch

[2]Future Cities Laboratory, Singapore-ETH Centre, ETH Zurich, CREATE Tower, Singapore 138602. E-mail: costa@lix.polytechnique.fr

[3]Applied Computing and Mechanics Laboratory, School of Architecture, Civil and Environmental Engineering, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland. E-mail: ian.smith@ epfl.ch

Keywords: clustering, bipartite modularity, structural identification, error-domain model falsification, K-means

INTRODUCTION

The growth of cities means that demand for fast, reliable and safe mobility in urban environments is increasing. Transportation networks provide an essential contribution. The core of these networks is made up of infrastructure such as roads and bridges that often have not been designed to meet current needs. It has been recently estimated that a one-trillion-dollar gap per year exists between infrastructure demand and supply (i.e. existing infrastructure plus new construction) (World Economic Forum 2014). Predictions indicate that this supply shortfall will increase in the future since demand reduction and a surge in new construction are both unlikely.

Responsive cities are intended to improve the decision making and to help asset managers optimise the allocation of resources. However, complications arise when the Internet of Things is scaled up to the level of the city since they have not been designed to be measured and monitored. Therefore, several challenges remain in collecting and interpreting the *response*.

Probably the most outstanding challenge is that effects, rather than causes, are generally measured in the built environment. Data interpretation requires advanced model-based analyses to understand the real behaviour of existing infrastructure. Also, design-like approaches often underestimate reserve-capacity sources, which result from conservative practices carried out during design and construction. Therefore, significant savings of resources, time, energy, materials, and as a consequence money, are provided when real-behaviour models are used to compare scenarios such as replacement, retrofit and improvement of infrastructure.

Structural model updating involves identifying suitable models as well as values for model parameters that determine structure behaviour through comparing measurements with predictions. Although sensing provides additional information of structural behaviour, uncertainties and systematic bias affect structural models (Catbas et al. 2013; Raphael and Smith 2003; Simoen et al. 2015). Also, understanding real behaviour is an iterative task (Pasquier and Smith 2016). Engineers need to make assumptions, generate models, collect measurements, update model parame-

ters and eventually use these models for diagnoses and prognoses. This process is repeated several times throughout service lives to appraise the structure and fix management priorities. Therefore, model-updating methodologies and results need to be understandable to decision makers (Smith 2016).

Among several available model-updating techniques, error-domain model falsification (EDMF) (Goulet and Smith 2013) is an approach that provides parameter identification without having to make assumptions on values of uncertainty correlations between sensor locations. Initially, this methodology requires the generation of a model population, which is based on engineering judgment and available knowledge. Uncertainties associated with modelling and measurements are combined and threshold bounds are evaluated according to a reliability of identification.

Falsification is performed through comparing model predictions with field measurements. Models for which residuals between predictions and measurements exceed threshold bounds, at one or more sensor locations, are falsified. Models for which residuals are within these bounds at each sensor location are included in the candidate model set (CMS). When candidate models are identified, they are then employed to perform predictions, for example, at unmeasured locations (Pasquier and Smith 2015) and those predictions may be used to assess the reserve capacity of the structure (Pasquier et al. 2014; Proverbio et al. 2018b). Furthermore, reserve capacity estimations form the basis for well-engineered interventions, such as retrofitting for capacity improvement.

Compared with other structural-identification methodologies such as Bayesian model updating, EDMF does not require advanced statistics knowledge and, therefore, it is easy to understand for practising engineers. However, result interpretation in population-based methods may be demanding when many equivalently-likely models are identified. Engineers may be overwhelmed with managing results consisting of multiple models for the same structure. Therefore, data-mining techniques are examined in this paper. The need for such support has been previously highlighted in (Smith and Saitta 2008; Saitta et al. 2008a)

Techniques such as decision trees (Saitta et al. 2005a), neural networks (Yun and Bahng 2000), case-based reasoning (Portinale et al. 2004), have already been integrated into diagnostic method-

ologies. Other studies describe methods that are specifically tailored to dynamic systems (Abad et al. 2002), automatic repair and automatic defect classification (McNamara et al. 2004; Saunders et al. 2000), consistency-based diagnosis (Alonso et al. 2004), and hierarchical clustering for bridge performance (Magalhaes et al. 2009). Preliminary data-fusion aspects such as data preparation, combination and data quality for civil infrastructure have been studied in (Soibelman and Kim 2002).

Feature extraction methods such as principal component analysis (PCA) (Smith and Saitta 2008) and self-organizing maps (SOM) (Flexer 2001) have been applied to reduce dimensionality and to visualize clustering results. However, results that are provided in the principal-component space provide weak support for engineers. Other data mining techniques such as K-means clustering have already been employed to extract knowledge from a set of candidate models (Saitta et al. 2005b). Although K-means requires that the number of clusters is given as input, methods are available to determine reasonable values for this parameter (MacQueen 1967; Pelleg and Moore 2000; Saitta et al. 2008b). Previous research involving clustering of candidate models mainly focused on reducing the number of clusters in the CMS by iteratively adding new sensor locations. Previous research mainly focused on allocating candidate models to specific clusters and reducing the number of clusters in the CMS by iteratively adding sensors at new locations.

An alternative way to represent the CMS is to associate each model with a node of a graph, and the features of the models (i.e., the values of the parameters analysed) to another set of nodes. Relationships between models and parameter values are represented by edges connecting the corresponding nodes. This produces a bipartite graph because each edge connects nodes belonging to two separate groups.

Bipartite graphs are employed in many domains, for example (Guimera and Amaral 2005; Garcia et al. 2018; Good et al. 2010). In the context of recommender systems, where relationships between users and purchased items are represented, clustering is used for applications such as targeted marketing. An emerging technique to find clusters for bipartite graphs is the bipartite modularity (Barber 2007). This method is an extension to bipartite graphs introduced in (New-

man 2006) and defined as the fraction of edges within clusters minus the expected fraction of such edges in a random graph with the same degree distribution. Clustering of bipartite graphs – based on bipartite-modularity optimization strategies – can support population-based structural identification frameworks to identify groups of models that share common features. The usefulness of bipartite clustering has not yet been studied for improving knowledge related to solution spaces that are generated by structural identification.

This paper introduces an EDMF-compatible framework, based on BMO clustering, to investigate properties of the CMS and to improve understanding of measurement data. The proposed methodology is not confined to assigning candidate models to clusters. Rather, it originally represents the challenge of interpreting population-based structural-identification results as relationships in a bipartite network. BMO clustering provides clear result visualization regardless of the number of parameters that are involved, resulting in transparent human-computer interaction that supports informed decision making. Finally, the methodology supports the iterative nature of measurement, interpretation and action within an opportunistic sequence–free framework that includes model–based diagnosis and prognosis.

The remainder of the paper is organized as follows. The next section contains a discussion of background information on EDMF and the clustering algorithms. The subsequent section presents the new framework for clustering the CMS. Finally, two case-studies are used to compare the proposed approach with traditional clustering methods and to suggest subsequent action within an iterative identification framework.

BACKGROUND

Population-based structural identification (EDMF)

Error-domain model falsification (EDMF) (Goulet and Smith 2013) is a recently developed methodology for structural identification in which finite-element (FE) model predictions are compared with measurement data in order to identify plausible model instances that are defined by assigning unique combinations of parameter values to a model class. Each model class consists of

a FE parametric model that includes parameters such as material properties, geometry, boundary conditions and actions.

Let R_i be the real response of a structure – unknown in practice – at a sensor location i, and y_i be the measured value at the same location. Model predictions at location i, $g_i(\boldsymbol{\theta})$, can be evaluated through assigning a vector of parameter values $\boldsymbol{\theta}$ to the selected FE model class. Model-prediction uncertainty $U_{i,g}$ and measurement uncertainty $U_{i,y}$ are estimated and linked to the real behavior using the following equation:

$$g_i(\boldsymbol{\theta}) + U_{i,g} = R_i = y_i + U_{i,y} \quad \forall i \in \{1, \ldots, n_y\} \tag{1}$$

The two sources of uncertainty $U_{i,g}$ and $U_{i,y}$ can be combined in a unique term $U_{i,c}$, while the difference between a model prediction and a measurement at location i, is referred to as the residual $r_i = g_i(\boldsymbol{\theta}) - y_i$.

The error in measurements U_y includes sensor accuracy – based on manufacturing specifications and site conditions – and measurement repeatability that is usually estimated by conducting multiple series of tests on site. The error in the model class U_g, which is usually much larger than U_y, is estimated using values taken from the literature, stochastic methods (to estimate uncertainties of parameters that are not included in the model class parametrization), engineering judgment and local knowledge. Plausible behaviour models are identified indirectly by falsifying those for which residuals exceeds thresholds boundaries that are defined in the uncertainty domain (i.e., the error domain). Being a falsification approach, EDMF initially requires that a set of model instances, which is referred to as the initial model set (IMS), is generated by assigning parameter values to the model class. Then, threshold bounds are defined at each sensor location, according to a 95% confidence level. Finally, models for which residuals are within threshold bounds at each sensor location are included in the candidate model set (CMS). In real situations, considering the number of parameters and the computation times to obtain FE model predictions, the IMS can be generated using adaptive sampling approaches such as radial-basis-function sampling, an approach

that exploits derivative-free optimization techniques to help improve the search of plausible models (Proverbio et al. 2018a). Consequently, the CMS may consist of tens or hundreds of models that are all equivalently likely and some confusion may arise in interpreting identification results.

Clustering

Cluster analysis aims at finding subsets of nodes (called clusters or communities) of a graph, where nodes in the same cluster are somehow similar, and those in distinct clusters are different. Indeed, in the literature, many definitions of what is similar and what is different have been proposed.

When nodes are points in the space, the Euclidean distance can be used as a metric for the similarity among nodes. This way, when two nodes are close they are considered similar and, therefore, they are likely to be assigned to the same cluster. However, in some applications, graphs represent relationships between nodes rather than points in the space. These relationships are defined by edges connecting nodes, which can be directed or undirected, and they can have weights.

The task considered in this paper is represented both ways, i.e., as a set of points in the space and as two sets of nodes connected by unweighted undirected edges. The methods used to find clusters in these two settings are presented below.

K-means

Given a set of n points $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ where each point is a d-dimensional vector, clustering can be carried out using K-means. This algorithm, already employed in (Saitta et al. 2008b) to explain structural identification outcomes, aims to find a set of K clusters $C = \{C_1, \ldots, C_K\}$ which minimizes the following quantity:

$$\sum_{i=1}^{K} \sum_{\boldsymbol{x}_j \in C_i} ||\boldsymbol{x}_j - \boldsymbol{\mu}_i||^2 \tag{2}$$

where $\boldsymbol{\mu}_i$ is the d-dimensional vector (called centroid) representing the mean of the points belonging to cluster C_i. The function to minimize is the sum, for each cluster, of the square distance between the points in the cluster and their mean. Finding the optimal solution has been shown to

be a **NP**-hard problem.

K-means is an iterative heuristic algorithm to efficiently find good quality solutions for this problem, and the main steps are briefly summarized next. First, the number of clusters and initial values for the means of the clusters (called centroids) are set. Then, the algorithm iteratively repeats these two steps:

- assign each point $\boldsymbol{x}_i$ to the cluster containing the nearest centroid;
- update each centroid with the mean of the points in its cluster.

The method stops when no new assignment can be performed in Step 1. K-means provides a local optimum in general, and the solution can change according to the initial guess for the centroids and the number of clusters.

Bipartite-modularity optimisation (BMO)

A possible way to model the problem under study is a bipartite graph, where two distinct sets of nodes, called red (R) and blue (B), are connected by edges. No edge exists between two nodes having the same colour.

In this situation, clustering can be performed by solving the BMO problem. Bipartite modularity was introduced in (Barber 2007) as an extension to bipartite graphs of the modularity metric (Newman 2006). The bipartite modularity of each cluster can be expressed as the difference between the fraction of edges in the cluster and the expected fraction of such edges in a random graph whose nodes have the same expected degree. A good partition of a graph into clusters is obtained by maximizing the sum of the bipartite modularities of the cluster. This problem can be mathematically formulated using binary variables indicating whether a node belongs to a specific cluster (Costa and Hansen 2014). However, this would require the knowledge of the optimal number of clusters, which is unknown a priori. An alternative formulation of BMO, where the optimal number of clusters is not required as input, can be derived by exploiting the definition of bipartite modularity presented in (Zhan et al. 2011) and the transitivity conditions of the clique partitioning formulation (Grötschel and Wakabayashi 1989). More precisely, defining V as the union of the

two sets of nodes ($R \cup B$), the BMO can be defined as:

$$
\begin{aligned}
\frac{1}{m} \max \quad & \sum_{i \in R} \sum_{j \in B} \left(a_{ij} - \frac{k_i k_j}{m} \right) x_{ij} \\
\text{s.t.:} \quad & \forall i < j < l \in V \quad -x_{ij} + x_{il} + x_{jl} \leq 1 \\
& \forall i < j < l \in V \quad x_{ij} + x_{il} - x_{jl} \leq 1 \\
& \forall i < j < l \in V \quad x_{ij} - x_{il} + x_{jl} \leq 1 \\
& \forall i < j \in V \quad x_{ij} \in \{0, 1\},
\end{aligned}
\tag{3}
$$

where x_{ij} is a binary variable equal to 1 if nodes i and j belong to the same cluster, 0 otherwise, a_{ij} is a parameter equal to 1 if nodes i and j are connected by an edge, 0 otherwise, k_i is the degree of node i (i.e., the number of nodes connected to i), and m is the total number of edges of the graph. The objective function is the bipartite modularity, and the constraints of the problem are the transitivity conditions imposing that if nodes i and j belong to the same cluster, and nodes j and l belong to the same cluster, then nodes i and l must belong to the same cluster.

There are alternative ways to formalize the bipartite-modularity optimization problem, but the one reported here has the advantage of not requiring the number of clusters as input.

Solution approach for BMO clustering

Bipartite-modularity maximization is a **NP**-hard problem (Costa and Hansen 2011; Miyauchi and Sukegawa 2015) and it can be solved with Mixed-Integer Linear Programming (MILP) only when the size of the instance is not too large. In practice, the model presented by Equation (3) can be given as input to a MILP solver like CPLEX (IBM-ILOG 2014), and the output will be the optimal values of the variables x_{ij} that can be used to derive the optimal partition of the graph into clusters.

Heuristics have been proposed to solve larger instances. Some of them, e.g., (Barber and Clark 2009; Liu and Murata 2010) are extensions of a label propagation method, where each vertex is iteratively assigned to the cluster containing the majority of its neighbours until convergence.

The heuristic employed in this paper, which produces good results with medium-size instances, is the locally optimal divisive heuristic presented in (Costa and Hansen 2014). Starting from an initial partition with one cluster containing all the vertices, the divisive heuristic recursively splits each cluster into two new clusters in an optimal way, i.e., by maximizing the resulting bipartite modularity. The procedure stops when additional splits do not further improve the bipartite modularity.

RESULT-INTERPRETATION METHODOLOGY

Framework

A new framework based on bipartite-modularity optimisation (BMO) is described in this section. This framework – shown in Figure 1 – helps extract knowledge from the CMS while providing engineering-oriented result visualizations.

Initially, parameter identification is performed according to the EDMF methodology and plausible models are included in the CMS. Since engineers may be overwhelmed with managing results from multiple models for the same structure, BMO clustering is applied to the CMS in order to support the decision-making process.

Once the CMS is identified, the bipartite graph, which is a network of nodes divided into two partitions that are connected by edges, is generated. The first partition is the CMS, which includes all models that are identified using EDMF. The second partition consists of ranges for parameters $\boldsymbol{\theta}$ that define each candidate model. This subdivision of the identified intervals of parameter values into ranges is performed by engineers considering: i) the candidate-model parameter distributions, and ii) the current stage of the structural identification process. The global performance of identification – reduction of parameter initial ranges – varies according to the measurement system adopted and the sampling technique used to generate the IMS.

Parameters that are well identified – updated interval smaller than the initial one – may be divided into a few (i.e., two or three) parameter ranges for clustering. Alternatively, they may be omitted from clustering, since all candidate models have similar values for well-identified parameters. Parameters that are poorly identified – with similar updated and initial intervals – may be

divided into several ranges, to represent many behaviours. This situation may happen, for example, when few sensors are employed.

Once the bipartite graph is generated, each node is assigned to a cluster according to the results of the bipartite modularity optimization procedure. More precisely, the optimal values of the variables x_{ij} are used to assign each model to its corresponding cluster, thus allowing the visualization step.

As mentioned in the Background section, when the size of the graph is large solving exactly the bipartite modularity optimization problem can be computationally challenging. Therefore, an approach based on the locally optimal divisive heuristic (Costa and Hansen 2014) is employed when large instances are addressed. This method, other than being computationally more efficient, provides accurate results.

The value of modularity obtained after the optimization is a direct metric of the goodness of classification: the higher the value of modularity, the better the classification. However, high modularity values do not guarantee the visualization to be effective for engineers. Therefore, if the knowledge provided by visualization is not sufficient, engineers may define either alternative initial subdivisions of ranges or modify the selection of parameters for clustering. This results in a new clustering that can be visualised. Since graph representation is not affected by the number of parameters that are involved, feature extraction techniques such as PCA are not required. When engineers are satisfied with the results, they can decide to proceed with the next stage of structural identification.

To help engineers perform decision-making tasks – as explained more into details in the next section – clustering results can be condensed by defining centroid models – one for each cluster – that are able to represent the entire CMS (Saitta et al. 2008a). In BMO clustering, centroid models can be easily computed for each cluster in the optimal solution. Centroid-model predictions can be subsequently evaluated, using a FE solver, through assigning centroid values to the model class that has been employed for falsification. However, care should be taken when the CMS population is replaced by few centroid models. For example, inaccurate clustering may generate centroid

models whose predictions are not compatible with measurements – and, therefore, should not be employed to represent the CMS. To avoid this issue, a centroid-model check is included in the BMO clustering framework. Notice that, since centroid models of each cluster are assigned mean values of parameters, extreme values of CMS predictions are likely to be omitted.

However, the information provided by BMO cluster visualization and centroid-model predictions can effectively support engineers that are performing structural identification, as explained in the next section.

Decision making

Structural identification is an iterative process that includes the following six tasks which may be executed in any order: modelling, monitoring, in-situ inspection, model falsification, diagnosis, and prognosis (Pasquier and Smith 2016). Engineers select the next iteration based on the current stage and the knowledge obtained from previous steps. The BMO clustering is a tool that can assist practising engineers who perform structural identification using population-based methodologies such as EDMF. The contribution of BMO clustering in the structural identification framework, which is depicted in Figure 2, is briefly discussed in the following.

The modelling task consists of building a FE model that describes the structural behaviour and a statistical model of the errors associated with the physics-based model. When time-consuming non-linear FE analyses are performed, BMO clustering helps reduce the number of model instances through providing few centroid models that: i) are compatible with measurements and ii) are able to represent the entire CMS.

At the early stages of structural identification, a subset of measurements is often compared with model predictions, thus limiting the computational demand for preliminary comparisons. As knowledge is acquired, the size of measurement sets usually increases. BMO clustering can be used to improve the sensor configuration by providing information on sensor types and measurement locations that are able to falsify entire clusters.

In-situ inspection comprises visual inspection and other non-destructive testing techniques. It allows engineers to improve their basic knowledge – based on structural drawings – with infor-

mation such as in-situ boundary conditions, as-built geometry and material deterioration. Visualisations of BMO-clustering may provide information that helps adjust the focus during future site inspections. For example, information on material properties can be employed to falsify a cluster that corresponds to a specific behaviour of the structure, thus refining result interpretation.

In the diagnostic phase, engineers interpret identification results of physical properties of the structure and draw conclusions about the structural conditions. BMO clustering helps clarify and organise the information provided by the CMS and convert it into knowledge.

CASE STUDY A – EXETER (UK)

A case study that involves the structural identification of the Exeter Bascule Bridge in the UK is employed to demonstrate the applicability of the proposed framework. The steel bridge, built in 1972, has a single span of 17.3 m and is designed to be lifted in order to allow boat passing along the canal. The light-weight aluminium deck is connected to several secondary beams that are bolted to two longitudinal girders (W36x12 section). The bridge has a total width of about 8.2 m and carries the carriageway and a footway. A static load test has been performed and deflection measurements have been collected by means of a target and a precision camera. Figure 3 shows the side elevation and a view of the bridge during the load test. Additional information about the Exeter Bascule Bridge can be found in (Kwad et al. 2017).

Model falsification

According to a sensitivity analysis, the following three parameters that influence the most the structural behaviour are selected for model updating: Youngs modulus of aluminium deck (θ_1), rotational stiffness of the North-bank hinges (θ_2), and axial stiffness of hydraulic jacks (θ_3). Table 1 contains initial intervals for the adopted parameters. Bounds for Youngs modulus are defined using engineering judgment, while values for the rotational stiffness cover the full range from a constrained to a pinned support in order to include the potential effects of corrosion at the bearings. The axial stiffness of hydraulic jacks is used to simulate their contribution as additional load-carrying supports. An initial model set of 1,000 instances is generated from the uniform distribution of each parameter value using Latin hypercube sampling. Model uncertainties are

defined considering model class simplifications, mesh refinements and finite-element numerical approximations. Measurement uncertainties take into account the sensor accuracy – provided by the manufacturer specification – and measurement repeatability, which is estimated by performing multiple measurements under site conditions.

Uncertainty sources are combined using the Monte Carlo method and threshold bounds are computed for a confidence level that is fixed at 95%. Residuals between deflection predictions and measurements collected using a precision camera are computed. Out of 1,000 initial model instances, a CMS consisting of 103 models is identified.Threshold bounds are computed and the CMS, consisting of 103 models, is identified. Table 1 reports initial intervals (top rows) and updated intervals (italics) for the parameters that have been considered for model updating. The performance of identification depends on several factors, such as the initial sampling and the sensor configuration. Using only one sensor, the longitudinal stiffness of the hydraulic jack has been clearly identified, while for the Youngs modulus of aluminium deck and the rotational stiffness of the bearing devices initial intervals and identified intervals are similar.

Clustering of bipartite graph

To represent the CMS a bipartite graph is generated. The first partition consists of the 103 candidate models that constitute the CMS. The second partition, which consists of ranges for candidate-model parameter values, has to be defined according to engineering judgment. Since θ_3 has been well identified a plausible choice is to cluster the CMS considering only 2 parameters (θ_1, θ_2). The parameter ranges adopted in this study are reported in Table 2.

Once both partitions are defined, the bipartite graph can be generated as shown in Figure 4. The left-hand side partition represents the 103 candidate models and the right-hand side represents the parameter ranges according to Table 2. Each edge joins a candidate model with the corresponding range for each parameter. Figure 5 shows the partition of the graph into clusters obtained by BMO clustering, and clusters are defined using node colours. To improve visualization, clusters are vertically separated. While clusters C-3 and C-5 have at least one range for each parameter that has been considered for clustering, both C-1, C-2 and C-4 do not include ranges for all parameters. This

is a direct consequence of the non-uniform distribution of parameter values in the CMS (see Figure 6). Since only two parameters have been considered for classification, the cluster visualisation is possible in the parameter space (Figure 6). Also, for each cluster, a centroid model is computed. According to BMO clustering, models that show high value of θ_2 are grouped regardless of their values of θ_1. On the contrary, when θ_2 is low, models are clustered according to θ_1-values.

Comparison with K-means

In order to highlight the advantages of the suggested approach, a comparison with a traditional clustering algorithm (i.e., K-means) is presented. K-means is a popular unsupervised learning method, which is already implemented in several engineering analysis tools. Moreover, previous studies that focused on interpreting results of model populations – for example (Saitta et al. 2005b) – employed K-means clustering. Finally, K-means is often applied by practicing engineers since it is easy-to-use and does not require an advanced machine-learning background.

K-means requires that the number of clusters K is given as input. As mentioned in the Background section, some techniques are available to help define K. In this case study, a reasonable value for the number of clusters was found to be between 2 and 5. Figure 7 shows K-means clustering while varying the K-value. Resulting clusters are defined by only θ_1 and all centroid models have almost identical values of θ_2.

The two parameters θ_1 and θ_2 represent respectively the bending stiffness and the rotational retain of a single span bridge. Therefore, in this situation, a negative correlation between these parameters is expected. Looking at the centroids in Figure 6, obtained by BMO, such a relationship – even though weak - can be identified. On the other hand, the centroids obtained by K-means in Figure 7 seem to suggest that the value of θ_2 is around its mid-range and the value of θ_1 is irrelevant.

CASE STUDY B – SINGAPORE

A second case study that involves the structural identification of a reinforced concrete bridge in Singapore is presented. The bridge (Figure 8), which consists of 4 precast prestressed beams that support a concrete deck, has a single span of 32 m.

During a static load test measurements have been collected by means of a laser tracker targeting 4 prisms (P) and 8 strain gauges (S) which have been attached to the bottom face of the main precast beams. Moreover, 2 inclinometers (I) have been positioned on the bridge deck close to the expansion joints. A detailed description of the sensor configuration and the case study is available in (Proverbio et al. 2018a).

The initial model set is generated through sampling the five-dimensional parameter space defined by the Youngs modulus of cast-in-place concrete, the Youngs modulus of precast concrete, the Youngs modulus of barrier concrete, the rotational and the vertical stiffness of the bearing devices. The initial interval for each parameter, defined according to engineering judgment, is reported in Table 3. The initial model set consists of 2,000 instances that are sampled using an adaptive sampling approach, as described in (Proverbio et al. 2017a). Model uncertainties take into account model simplifications, which are estimated using engineering judgment and considering the model class features. Moreover, this source takes into account that only the most sensitive parameter uncertainties are selected for identification. The uncertainty associated with mesh refinement and numerical approximations are also included. Measurement uncertainties are estimated by comparing multiple measurements under site conditions and considering the type of sensors that are employed. Sensor accuracies which represent the lowest source of measurement uncertainty are based on manufacturer specifications. For strain gauges, an uncertainty also arises from the imperfect alignment of gauges with respect to the longitudinal axis of the bridge. Finally, additional noise associated with sensor installation have been considered for inclinometers and strain gauges using engineering judgment.

Model and measurement uncertainties are estimated and, for each measurement location, a combined uncertainty is computed and threshold bounds are determined for a confidence level fixed at 95%. Table 3 reports initial (top parts of rows) and updated (italics – lower parts of rows) intervals for the parameters that have been considered for model updating.

Candidate model set B.1 – 80 models

Out of the 2,000 initial model instances, 80 candidate models are identified using the information provided by all available sensors (i.e. four deflection prisms, eight strain gauges and two inclinometers).

BMO clustering – 4 parameters

To represent the CMS a bipartite graph is generated. The first partition consists of the CMS while the second partition is based on candidate-model parameter ranges that are defined by the engineer. Parameters which are well identified – updated interval smaller than initial interval – should be omitted from the clustering in order to reduce the graph size. Moreover, further subdivisions of a well-identified range of parameter values provide weak support for decision making.

Referring to Table 3, the parameter θ_4 is fairly well identified and should be omitted from clustering. However, the 53% initial-range reduction for θ_5 may suggest that also this parameter should be omitted. This choice is part of the active interaction between the engineer and the framework. To help clarify the outcomes of such a decision, in this section four parameters (i.e. θ_1, θ_2, θ_3, θ_5) are employed, while the case involving the selection of θ_1, θ_2, and θ_3 is presented in the next section.

Table 4 reports the decomposition ranges defined by the engineer while the bipartite network is depicted in Figure 9, where clusters are identified by different colours.

The cluster visualization proposed in Figure 9 can be further enhanced by vertically separating clusters (see Figure 11) and can be performed regardless the number of parameters that have been considered. Moreover, looking at Figure 9 it is possible to notice that the large majority of the 80 candidate models have values of θ_5 in range II, while only few models (5 over 80) have $\theta_5$values in range I. This observation suggests that a better network representation may be achieved by omitting θ_5 from clustering. In this way, engineers interact with this clustering framework and update both the initial subdivision of ranges and parameter selection until the desired level of knowledge is acquired.

BMO clustering – 3 parameters

Given the identification results and according to the visualisation results of the previous iteration, a plausible choice is to cluster the CMS considering only 3 parameters (i.e. θ_1, θ_2, θ_3). The parameter ranges adopted in this study are reported in Table 5.

Figure 10 shows the graph, and clusters are defined using node colours. Clusters C-3 and C-2 have at least one range for each parameter that has been considered for clustering, while both C-1 and C-4 do not include ranges for all parameters. This is a direct consequence of the non-uniform distribution of parameter values in the CMS, as shown in Figure 11. More precisely, cluster C-4 is associated with values of θ_2 in the range (25 to 30 GPa) while values of parameters θ_1 and θ_3 are spread throughout the domain. Similarly, in cluster C-1 values for parameter θ_1 and θ_3 are clustered around one of the ranges defined in Table 2, while values of θ_2 are gathered in two groups that are at opposite bounds of the domain. Hence, there is no range node for θ_2 in cluster C-1. On the other hand, both clusters C-2 and C-3 identify ranges for each parameter. Therefore, models that belong to these clusters are expected to show distinct trends in the CMS.

To support result interpretation, clustering can be used to define few models – called centroid models (CMs) – that are able to represent the entire CMS. First, centroid coordinates are computed for each parameter as the mean of the values of the models belonging to the considered cluster. Then, centroid-model predictions are evaluated, using a FE solver. Finally, CMs are checked using EDMF threshold bounds and only those that are not falsified can be employed to represent the CMS.

Figure 12 shows the parallel-axis plot of parameter values that define centroid-models, along with CM predictions. CM-2 and CM-3 show different trends – in agreement with the observations made in Figure 10. CM-3 corresponds to a model class with high values of θ_2 and relatively low values of θ_1 and θ_3. An opposite trend is observed for CM-2. CM-1 shows parameter values 5% to 10% higher than CM-4 and a similar trend. All centroid models have similar values of θ_4 and θ_5, which, indeed, have not been considered for clustering. Moreover, all CMs provide predictions that are inside threshold bounds for each sensor locations. Therefore, all CMs can be employed for

decision making.

Understandably, when centroid models are used to represent the entire CMS, information provided by extreme parameter values is lost. For example, in Figure 12, no centroid model has values of θ_1 lower than 0.55 or higher than 0.9. Consequently, CM predictions cannot cover the prediction ranges provided by the entire CMS. This drawback may be reduced through employing uniform sampling to generate the initial model set to which EDMF is applied. Thus, parameter values will be uniformly distributed in the CMS. However, uniform sampling is not efficient since good sample density often requires a number of samples that makes the problem computationally challenging.

Decision making

When the information obtained from result visualisation satisfies the engineer, decision making is carried out. As mentioned above, decision making can result in additional testing and in-situ inspection, additional FE analyses, and structural diagnosis.

Considering the visualisation of centroid models in Figure 12, a first option is to employ non-destructive testing (NDT) to evaluate identified clusters. For example, CM-2 shows high values of θ_1 and low values of θ_2, while CM-3 is characterised by very high values of θ_2. Since θ_1 and θ_2 represent the Youngs modulus of precast barriers and cast-in-place concrete respectively, indications of actual values can be obtained by NDT such as ultrasonic pulse velocity. This information can help reduce initial uncertainty estimations and increase the precision of identification. Moreover, predictions provided by CM-2 and CM-3 are dissimilar at many sensor locations. Therefore, insights on the true behaviour of the structure can be acquired through analysing these two centroid models.

Centroid models may also be employed to obtain reserve-capacity preliminary estimates. A unitary reserve capacity means that the structural safety for a given limit state is verified under design load configurations. Table 6 reports reserve-capacity assessments for the serviceability limit states (SLS) – computed for each centroid model. CM-2 and CM-3 provide the minimum and the maximum reserve capacity respectively. Moreover, the four estimations are close (less

than 4% different) from the average reserve capacity computed using the entire CMS. A detailed explanation of reserve-capacity assessments for this case study is available in (Proverbio et al. 2018b).

Since the estimation of reserve-capacity involves checks of structural safety, extreme predictions are often determinant. However, fast preliminary estimations – based on CMS average predictions – help engineers employ population-based approaches for structural identification. The reserve capacity computed using CMs represents a trade-off between the entire CMS complexity and the synthesis provided by a unique average behaviour.

Comparison with K-means

In order to highlight the advantages of the suggested approach, a comparison with a traditional clustering algorithm (i.e., K-means) is presented.

K-means requires that the number of clusters K is given as input. To simplify the comparison with the previous results, the same number of clusters identified by the BMO clustering is selected (i.e., K=4).

Using the Euclidean distance as a metric, K-means provides the classification of candidate models into 4 clusters. Since 3 parameters have been considered for clustering, result visualization is possible by plotting the models as nodes in a 3D space – Figure 13.

In Figure 13, each node corresponds to a candidate model and crosses correspond to centroid positions. Compared with Figure 10, this visualization provides weak support, since it only allows appreciation of cluster distributions in the CMS. Little information on model properties is provided and engineers cannot actively interact with the framework or refine the visualization since the only parameter that is considered is the number of clusters K.

To support downstream processes and the decision making, CM predictions are evaluated and checked using EDMF threshold bounds. Table 7 shows CM check results for multiple K-values since only CMs that are not falsified can be employed to represent the CMS.

Considering the current choice of K-value (i.e. K=4), CM-4 is falsified. Moreover, Table 7 shows that two CMs will be falsified even though 5 or 6 clusters are initially considered.

Figure 14 shows the parallel-axis plot of parameter values that define centroid-models and the corresponding CM predictions. CM-4 is not compatible with measurements; therefore, it is rejected and plotted using a dashed line. Compared with Figure 12, values of θ_2 and θ_3 are gathered for all centroid models. This results in a weaker interpretation of the CMS. Moreover, all the accepted centroid models show similar trends. In this situation, engineers obtain only one possible interpretation of the CMS, since only values for θ_1 have been effectively explored.

Exact solution and divisive heuristic approach

In case study B.1 the size of the network is relatively small. When larger networks, such as those considered here, are evaluated, the optimisation of bipartite modularity can be computationally challenging. Just to give a reference, those instances could not be solved by the exact method in 2 hours on a server with four Intel Xeon E5-4620 CPUs (2.20 GHz, 8 cores, Hyper-Threading and Turbo Boost disabled) and 128 Gb of RAM (32 GB for each processor). Therefore, an approach based on the locally optimal divisive heuristic (Costa and Hansen 2014) is employed when large instances are addressed and solutions could be obtained in a few seconds on a much less powerful laptop.

As shown in Table 8, when both the exact method and the divisive heuristic can be used, the difference of bipartite modularity value is minimal and the number of clusters identified is the same. Therefore, the divisive-heuristic method, other than being computationally efficient, provides accurate results.

Finally, values of bipartite modularity help engineers estimate the quality of the visualisation at the current step. For example, regardless the optimisation approach used, bipartite-modularity values are higher when 3 parameters are included in the clustering. Therefore, omitting θ_5 from clustering has increased the quality of the clustering, which results in a better explanation of identification results.

Candidate model set B.2 – 260 models

The feasibility of the proposed approach to large CMSs is evaluated through assuming that no sensor can be installed below the bridge, thus excluding deflection prisms and strain gauges.

Consequently, falsification is carried out using the information provided by the two inclinometers positioned on the bridge deck and 260 candidate models are identified.

In real situations, engineers may deal with large CMSs at initial stages of structural identification, which are characterised by high uncertainty levels, or during preliminary monitoring, when only a few sensors are employed.

BMO clustering – 5 parameters

Having omitted the information provided by deflection prisms and strain gauges, the CMS B.2 consists of 260 models – 180 more compared with CMS B.1. Although EDMF succeeded in falsifying many out of the initial 2,000 model instances, initial ranges and updated ranges of parameters are coincident. Therefore, EDMF has excluded only some parameter combinations rather than reducing parameter intervals.

In this situation, all parameters participate in the clustering and parameter intervals are divided into a number of ranges that are equally large. Table 9 shows parameter intervals for each range along with cluster assignments obtained through the BMO-clustering method.

Since this case study involves 5 parameters, the direct visualisation of the parameter domain is not possible. Future extraction techniques such as PCA may be employed; however, result visualisation is possible only in the principal component space. Thus, clusters need to be mapped back in the parameter space to allow result interpretation. Although a bipartite network can be represented regardless the parameter domain dimensionality (see, for example, Figure 9), tabular representations of results are more appropriate for graphs defined by a large number of nodes.

A visual representation of the clusters presented in Table 9 is provided in Figure 15. Each vertical axis represents a parameter and candidate models are plotted as coloured lines. The dashed red line indicates the centroid model for each of the four clusters. Although few models show parameter values far from average values – which define centroid models – centroid models exhibit trends that are aligned with the ranges in Table 9. Therefore, the tabular format is able to effectively condense the information provided by the visualisation of parallel-axis plots.

Engineers can consider centroid models as a synthesis of the CMS. Although the range of

behaviour of 260 models cannot be captured by only four centroid models, it is worth noting that common features are found. For example, a negative correlation between parameters θ_4 and θ_5 characterises all centroid models. This observation is confirmed by checking the correlation coefficients reported in Table 10, which shows that parameters θ_4 and θ_5 exhibits a strong negative correlation.

Decision making

Since initial and identified parameter ranges are similar, little knowledge can be extracted from the CMS. Such a situation suggests that additional monitoring – including new sensor configurations and load cases – and further inspection – involving, for example, non-destructive testing – are carried out.

The goal is to increase the information provided by the real structure to reduce the initial uncertainties. However, one of the main advantages of EDMF lies in its emphasis on accuracy rather than precision. Adding new sensors can only result in a reducing the number of candidate models; therefore, incorrect falsification of plausible models is avoided. In other words, the CMS represents the most accurate knowledge given the current level of information available.

Result-interpretation techniques should follow analogous principles. Centroid models that represent the CMS should describe the range of behaviour that is plausible at the current stage while simplifying the inclusion of additional information that may become available.

For example, the CMS B.2 – obtained using 2 sensors – can be seen as an initial stage of structural identification, while CMS B.1 results from an improved sensor configuration, which consists of 12 sensors. Table 11 represents the updated version of Table 9 when all sensors are employed. Clusters indicated in curly brackets are falsified and some parameter ranges are reduced. As a result, cluster C4, which includes extreme values for parameters, θ_1, θ_4 and θ_5, is falsified for these three parameters and C3 is falsified for θ_4 and θ_5. Interestingly, all falsified models are in the same two clusters; therefore, BMO clustering provided an accurate interpretation of the CMS.

SUMMARY AND DISCUSSION

Engineers may be overwhelmed with managing results from multiple models that explain mea-

surements taken from the same structure. Clustering can be effectively employed to group entities that are *similar*. However, several definitions of similarity, which come out of different clustering strategies are possible.

K-means in the solution space – one of the most well known algorithm for clustering – is chosen as a benchmark since it is easy-to-use and already implemented in several data-analysis tools. The traditional implementation of K-means, which employs Euclidean distance to cluster similar nodes, does not facilitate user interaction since input parameters such as the number of clusters or the selection of initial centroids are not specific to the problem at hand. In order to overcome this limitation, bipartite networks can be used to describe relationships between plausible behaviour models. Moreover, the bipartite-network representation enables use of the BMO approach to highlight existing similarities between subsets of models and specific behaviour regardless the number of parameters that are taken into account. Therefore, dimensionality reduction techniques such as PCA are not necessary. Results show that BMO clustering successfully condenses the information provided by the CMS into a few centroid models that are able to represent plausible behaviour. While it is possible to modify K-means clustering in order to leverage similar domain representations, this strategy may not be easily understandable to engineers who are responsible for asset management.

The following limitations of the framework are recognised. The sampling technique adopted to generate the model population and the assessment of uncertainties influence identification results. Accurate parameter identification can be achieved only when reliable model classes are adopted. Model-class features and model uncertainties should always be verified through visual inspection and iterative model-class updating when new information becomes available.

Although this study focuses on the downstream process of identification – after the CMS is defined – new research directions involve applying classification algorithms to perform structural identification. For example, logistic regression or support vector machine may be employed to investigate hidden relationships between parameter values and model predictions to guide the search for additional candidate models when performing falsification.

CONCLUSIONS

BMO clustering effectively helps interpret structural-identification results when population-based approaches – such as EDMF – are employed. Specific conclusions are as follows:

- BMO clustering helps clarify and interpret the candidate model set.
- The proposed methodology identifies feasible centroid models more successfully than traditional applications of K-means.
- Result visualization is possible regardless of the number of parameters.
- Large CMSs containing many instances (more than 100) are successfully clustered using the divisive heuristic approach.
- Finally, active interaction with the clustering framework is possible to leverage new knowledge during several stages of the asset-management decision-making process.

ACKNOWLEDGMENTS

This paper is a significantly expanded version of a conference paper presented at the IEEE Symposium Series on Computational Intelligence (Costa et al. 2017), from which 3 pictures have been reused. The research was conducted at the Future Cities Laboratory at the Singapore-ETH Centre, which was established collaboratively between ETH Zurich and Singapore's National Research Foundation (FI 370074011) under its Campus for Research Excellence and Technological Enterprise programme. The authors gratefully acknowledge the support of the Land Transport Authority (LTA) of Singapore, the University of Exeter and the Full Scale Dynamics Ltd for support during load tests in the scope of the two case studies.

REFERENCES

Abad, P. J., Surez, A. J., Gasca, R. M., and Ortega, J. A. (2002). "Using supervised learning techniques for diagnosis of dynamic systems." *Report No. ADP012709*, Universidad de Huelva Cantero (Spain).

Alonso, C. J., Rodrguez, J. J., and Pulido, B. (2004). "Enhancing consistency based diagnosis with machine learning techniques." *Lecture Notes in Computer Science*, 312–321.

649 Barber, M. J. (2007). "Modularity and community detection in bipartite networks." *Physical Re-*
650 *view E*, 76(6), 066102.

651 Barber, M. J. and Clark, J. W. (2009). "Detecting network communities by propagating labels
652 under constraints." *Physical Review E*, 80(2), 026129.

653 Catbas, F., Kijewski-Correa, T., Lynn, T., Aktan, A. E., and others (2013). "Structural identification
654 of constructed systems." American Society of Civil Engineers.

655 Costa, A. and Hansen, P. (2011). "Comment on Evolutionary method for finding communities in
656 bipartite networks." *Physical Review E*, 84(5), 058101.

657 Costa, A. and Hansen, P. (2014). "A locally optimal hierarchical divisive heuristic for bipartite
658 modularity maximization." *Optimization Letters*, 8(3), 903–917.

659 Costa, A., Proverbio, M., and Smith, I. F. C. (2017). "Cyber civil infrastructure and IoT for
660 cities." *Proceeding of IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE,
661 2140–2147.

662 Flexer, A. (2001). "On the use of self-organizing maps for clustering and visualization." *Intelligent*
663 *data analysis*, 5(5), 373–384.

664 Garcia, J. O., Ashourvan, A., Muldoon, S. F., Vettel, J. M., and Bassett, D. S. (2018). "Applications
665 of community detection techniques to brain graphs: Algorithmic considerations and implications
666 for neural function." *Proceedings of the IEEE*.

667 Good, B. H., de Montjoye, Y.-A., and Clauset, A. (2010). "Performance of modularity maximiza-
668 tion in practical contexts." *Physical Review E*, 81(4), 046106.

669 Goulet, J.-A. and Smith, I. F. C. (2013). "Structural identification with systematic errors and un-
670 known uncertainty dependencies." *Computers & structures*, 128, 251–258.

671 Grötschel, M. and Wakabayashi, Y. (1989). "A cutting plane algorithm for a clustering problem."
672 *Mathematical Programming*, 45(1-3), 59–96.

673 Guimera, R. and Amaral, L. A. N. (2005). "Functional cartography of complex metabolic net-
674 works." *nature*, 433(7028), 895.

675 IBM-ILOG, C. (2014). *12.6 User's Manual*. IBM.

Kwad, J., Alencar, G., Correia, J., Jesus, A., Calada, R., and Kripakaran, P. (2017). "Fatigue assessment of an existing steel bridge by finite element modelling and field measurements." *Journal of Physics: Conference Series*, Vol. 843, IOP Publishing, 012038.

Liu, X. and Murata, T. (2010). "An efficient algorithm for optimizing bipartite modularity in bipartite networks." *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 408–415.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations." *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, Oakland, CA, USA., 281–297.

Magalhaes, F., Cunha, A., and Caetano, E. (2009). "Online automatic identification of the modal parameters of a long span arch bridge." *Mechanical Systems and Signal Processing*, 23(2), 316–329.

McNamara, J. D., di Scalea, F. L., and Fateh, M. (2004). "Automatic defect classification in long-range ultrasonic rail inspection using a support vector machine-based smart system." *Insight-Non-Destructive Testing and Condition Monitoring*, 46(6), 331–337.

Miyauchi, A. and Sukegawa, N. (2015). "Maximizing Barbers bipartite modularity is also hard." *Optimization Letters*, 9(5), 897–913.

Newman, M. E. (2006). "Modularity and community structure in networks." *Proceedings of the national academy of sciences*, 103(23), 8577–8582.

Pasquier, R., Goulet, J.-A., Acevedo, C., and Smith, I. F. C. (2014). "Improving fatigue evaluations of structures using in-service behavior measurement data." *Journal of Bridge Engineering*, 19(11), 04014045.

Pasquier, R. and Smith, I. F. C. (2015). "Robust system identification and model predictions in the presence of systematic uncertainty." *Advanced Engineering Informatics*, 29(4), 1096–1109.

Pasquier, R. and Smith, I. F. C. (2016). "Iterative structural identification framework for evaluation of existing structures." *Engineering Structures*, 106, 179–194.

Pelleg, D. and Moore, A. W. (2000). "X-means: Extending K-means with Efficient Estimation of

the Number of Clusters..” *ICML*, Vol. 1, 727–734.

Portinale, L., Magro, D., and Torasso, P. (2004). “Multi-modal diagnosis combining case-based and model-based reasoning: a formal and experimental analysis.” *Artificial Intelligence*, 158(2), 109–153.

Proverbio, M., Costa, A., and Smith, I. F. C. (2018a). “Adaptive sampling methodology for structural identification using radial-basis functions.” *Journal of Computing in Civil Engineering*, 32(3), 04018008.

Proverbio, M., Vernay, D. G., and Smith, I. F. C. (2018b). “Population-based structural identification for reserve-capacity assessment of existing bridges.” - *Under review*.

Raphael, B. and Smith, I. F. C. (2003). *Fundamentals of computer-aided engineering*. John Wiley & Sons.

Saitta, S., Kripakaran, P., Raphael, B., and Smith, I. F. C. (2008a). “Improving system identification using clustering.” *Journal of Computing in Civil Engineering*, 22(5), 292–302.

Saitta, S., Raphael, B., and Smith, I. F. C. (2005a). “Data mining techniques for improving the reliability of system identification.” *Advanced Engineering Informatics*, 19(4), 289–298.

Saitta, S., Raphael, B., and Smith, I. F. C. (2005b). “Supporting engineers during system identification.” *Computing in Civil Engineering (2005)*, 1–12.

Saitta, S., Raphael, B., and Smith, I. F. C. (2008b). “A comprehensive validity index for clustering.” *Intelligent Data Analysis*, 12(6), 529–548.

Saunders, C., Gammerman, A., Brown, H., and Donald, G. (2000). “Application of support vector machines to fault diagnosis and automated repair.” *Eleventh International Workshop on Principles of Diagnosis (DX ’00)*.

Simoen, E., De Roeck, G., and Lombaert, G. (2015). “Dealing with uncertainty in model updating for damage assessment: A review.” *Mechanical Systems and Signal Processing*, 56, 123–149.

Smith, I. F. C. (2016). “Studies of Sensor Data interpretation for Asset Management of the Built environment.” *Frontiers in Built Environment*, 2, 8.

Smith, I. F. C. and Saitta, S. (2008). “Improving knowledge of structural system behavior through

multiple models." *Journal of structural engineering*, 134(4), 553–561.

Soibelman, L. and Kim, H. (2002). "Data preparation process for construction knowledge generation through knowledge discovery in databases." *Journal of Computing in Civil Engineering*, 16(1), 39–48.

World Economic Forum (2014). "Strategic infrastructure, steps to operate and maintain infrastructure efficiently and effectively." *Report No. 180314*, World Economic Forum, Davos.

Yun, C.-B. and Bahng, E. Y. (2000). "Substructural identification using neural networks." *Computers & Structures*, 77(1), 41–52.

Zhan, W., Zhang, Z., Guan, J., and Zhou, S. (2011). "Evolutionary method for finding communities in bipartite networks." *Physical Review E*, 83(6), 066120.

List of Tables

TABLE 1. Initial intervals (top parts of rows) and updated intervals (italics – lower parts of rows) for parameters that have been considered for falsification

Parameters	Lower bound	Upper bound	Initial-range reduction
θ_1 – Youngs modulus of aluminium deck [GPa]	60.00 *60.53*	80.00 *79.53*	5%
θ_2 – Rotational stiffness of bearing devices [log(Nmm/rad)]	8.00 *8.01*	12.00 *11.98*	0.7%
θ_3 – Longitudinal stiffness of hydraulic jack [log(Nmm)]	2.00 *4.02*	8.00 *4.39*	94%

TABLE 2. Parameter ranges for clustering defined by the engineer

Parameters	Parameter ranges			
	I	II	III	IV
θ_1 [GPa]	60-65	65-70	70-75	75-80
θ_2 [log(Nmm/rad)]	8-9	9-10	10-11	11-12

TABLE 3. Initial intervals (top rows) and updated intervals (italics) for parameters that have been considered for falsification

Parameters	Lower bound	Upper bound	Initial-range reduction
θ_1 – Young's modulus of barrier concrete [GPa]	3.00 *8.48*	40.00 *39.99*	14.8%
θ_2 – Youngs modulus of cast-in-place concrete [GPa]	20.00 *20.01*	35.00 *34.98*	0.8%
θ_3 – Rotational stiffness of bearing devices [log(Nmm/rad)]	9.00 *9.01*	13.00 *12.87*	3.5%
θ_4 – Vertical stiffness of bearing devices [log(N/mm)]	8.00 *8.34*	11.00 *8.67*	89%
θ_5 – Youngs modulus of precast concrete [GPa]	25.00 *38.31*	50.00 *49.97*	53.4%

TABLE 4. Parameter ranges for clustering defined by the engineer

Parameters	Parameter ranges			
	I	II	III	IV
θ_1 [GPa]	$<$20	20-30	$>$30	–
θ_2 [GPa]	$<$25	25-30	$>$30	–
θ_3 [log(Nmm/rad)]	$<$10	10-11	11-12	$>$12
θ_5 [GPa]	$<$45	$>$45	–	–

TABLE 5. Parameter ranges for clustering defined by the engineer

Parameters	Parameter ranges			
	I	II	III	IV
θ_1 [GPa]	$<$20	20-30	$>$30	–
θ_2 [GPa]	$<$25	25-30	$>$30	–
θ_3 [log(Nmm/rad)]	$<$10	10-11	11-12	$>$12

TABLE 6. Preliminary reserve-capacity assessments computed using each centroid model

Centroid model	Reserve capacity (SLS)
CM-1	1.36
CM-2	1.32
CM-3	1.38
CM-4	1.33
CMS average	1.37

TABLE 7. Centroid-model check using EDMF threshold bounds. K-means clustering is employed with values of K from 1 to 6 (✓: the CM is a candidate model; ✘: the CM is falsified)

K	Centroid-model check					
	CM-1	CM-2	CM-3	CM-4	CM-5	CM-6
1	✓	-	-	-	-	-
2	✓	✓	-	-	-	-
3	✓	✓	✓	-	-	-
4	✓	✓	✓	✘	-	-
5	✘	✘	✓	✓	✓	-
6	✘	✓	✓	✓	✘	✓

TABLE 8. Comparison of the exact solution and the divisive heuristic approach

Number of parameters for clustering	Approach	Number of clusters	Bipartite modularity	Δ bipartite modularity
4 Parameters	Exact	4	0.271	1.1%
	Divisive heuristic	4	0.268	
3 Parameters	Exact	4	0.353	0.3%
	Divisive heuristic	4	0.352	

TABLE 9. Parameter ranges used for BMO clustering (CMS B.2). Clustering results are indicated by different colours and the cluster identification is reported for each range

Ranges	θ_1 [GPa]	θ_2 [GPa]	θ_3 [log(Nmm/rad)]	θ_4 [log(Nmm)]	θ_5 [GPa]
IV	[30.75;40.00] **C1**	[31.25;35.00] **C2**	[12.00;13.00] **C3**	[10.25;11.00] **C4**	[43.75;50.00] **C1**
III	[21.50;30.75] **C2**	[27.50;31.25] **C4**	[11.00;12.00] **C4**	[9.50;10.25] **C4**	[37.50;43.75] **C2**
II	[12.25;21.50] **C3**	[23.75;27.50] **C1**	[10.00;11.00] **C2**	[8.75;9.50] **C3**	[31.25;37.50] **C3**
I	[3.00;12.25] **C4**	[20.00;23.75] **C3**	[9.00;10.00] **C1**	[8.00;8.75] **C1**	[25.00;31.25] **C4**

TABLE 10. Parameter correlation matrix (CMS B.2)

	θ_1	θ_2	θ_3	θ_4	θ_5
θ_1	1	-	-	-	-
θ_2	-0.01	1	-	-	-
θ_3	-0.19	-0.10	1	-	-
θ_4	-0.13	-0.07	-0.26	1	-
θ_5	-0.03	-0.08	-0.20	**-0.69**	1

TABLE 11. BMO clustering results considering the CMS B.1. Clusters are indicated by different colours and bold letters, while falsified clusters are indicated in curly brackets

Ranges	θ_1	θ_2	θ_3	θ_4	θ_5
IV	**C1**	**C2**	**C3**	$\{C4\}$	**C1**
III	**C2**	**C4**	**C4**	$\{C4\}$	**C2**
II	**C3**	**C1**	**C2**	$\{C3\}$	$\{C3\}$
I	$\{C4\}$	**C3**	**C1**	**C1**	$\{C4\}$

List of Figures

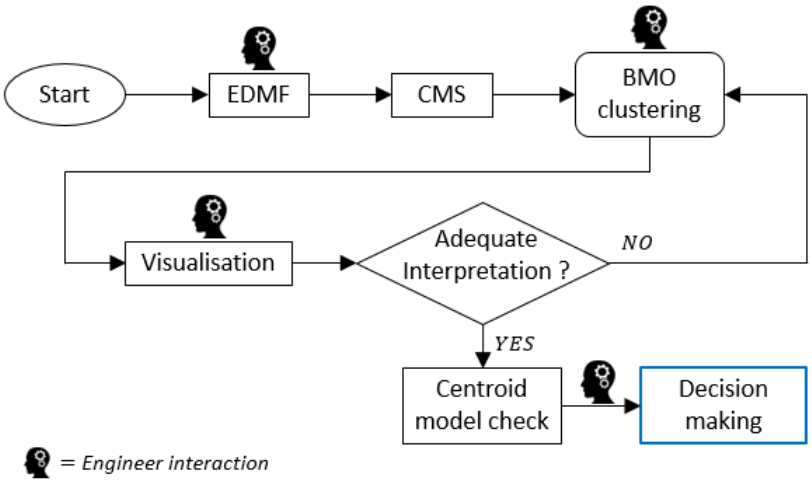

FIG. 1. Bipartite-modularity optimization (BMO) clustering framework

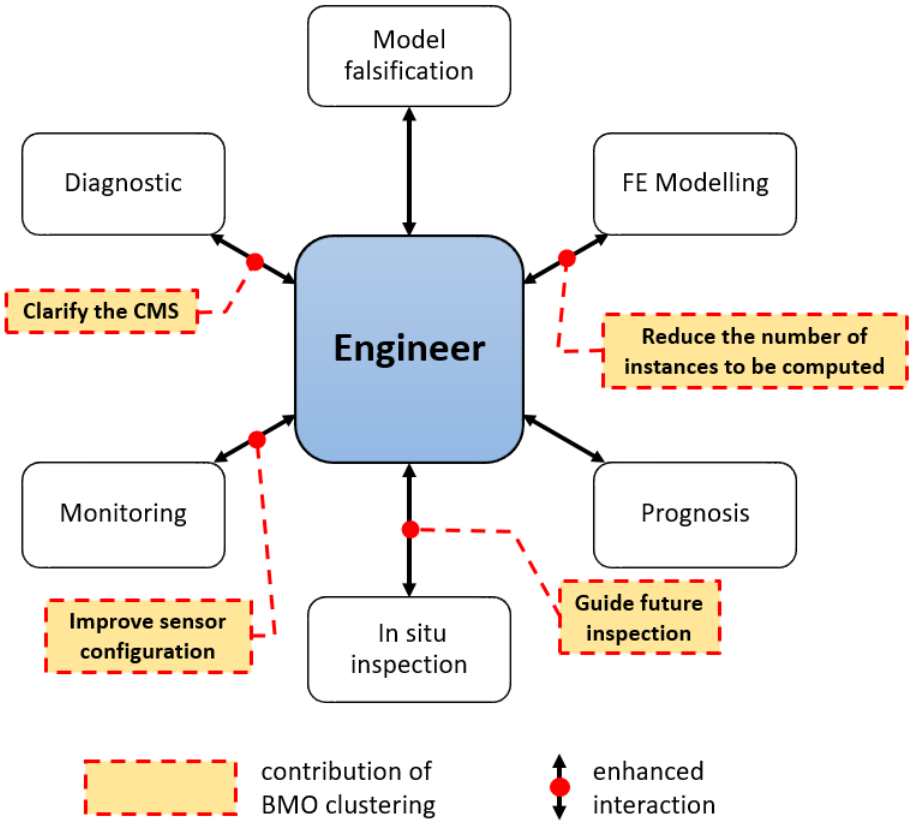

FIG. 2. The contribution of BMO clustering in the iterative sequence–free structural–identification framework. Adapted and enhanced from (Pasquier and Smith 2016)

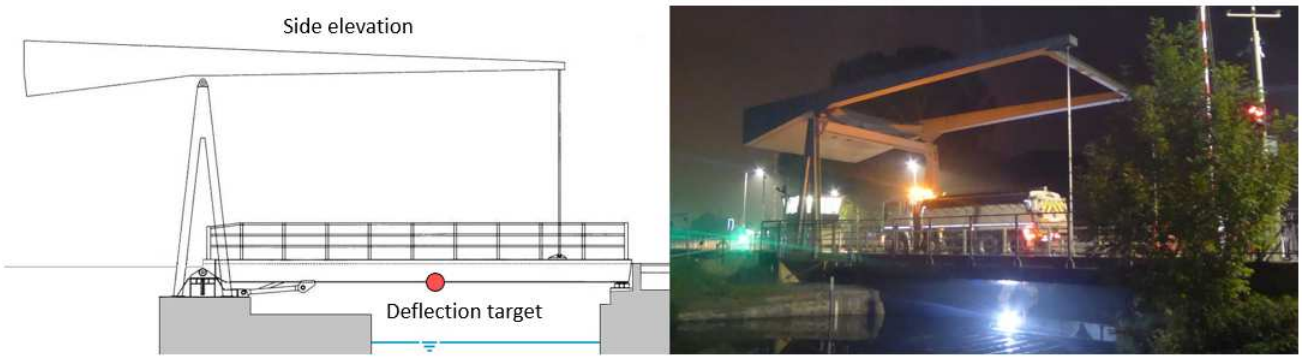

FIG. 3. Side elevation and view of the Exeter Bascule Bridge during the load test

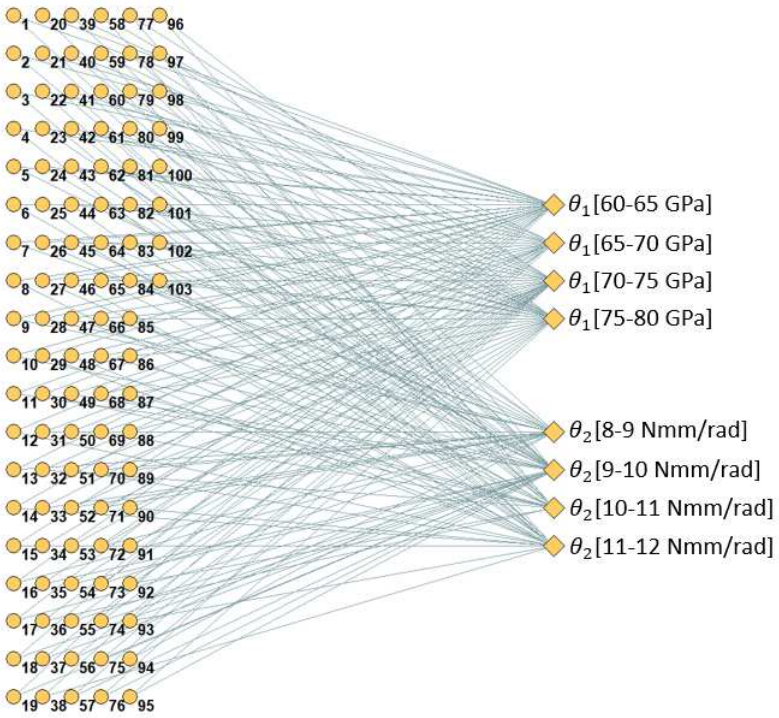

FIG. 4. Bipartite graph representation. The two partitions consist of candidate models (left) and parameter ranges (right)

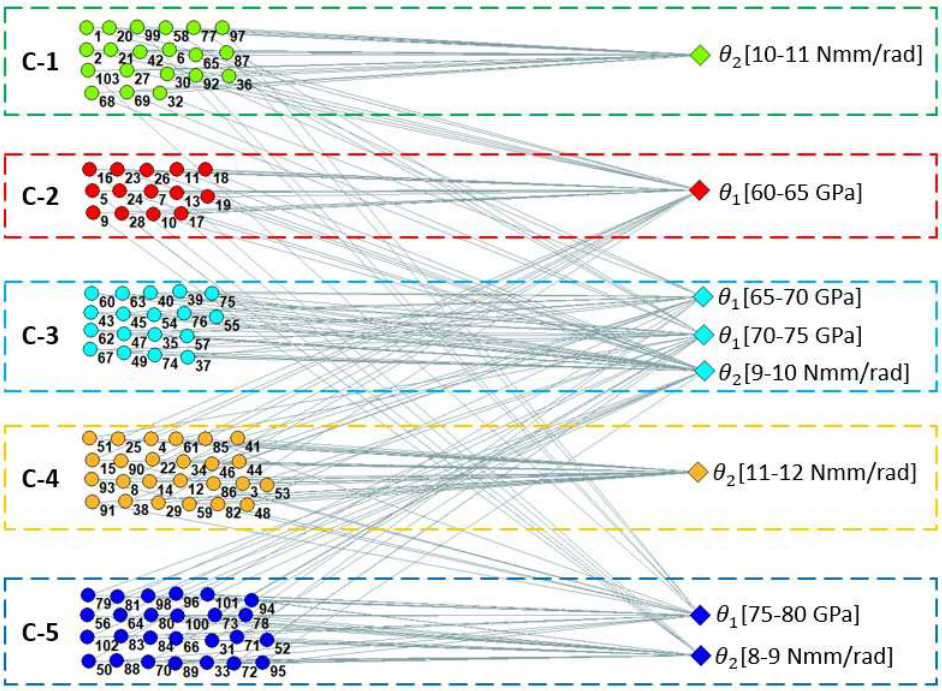

FIG. 5. BMO clustering visualization (CMS A – 2 parameters)

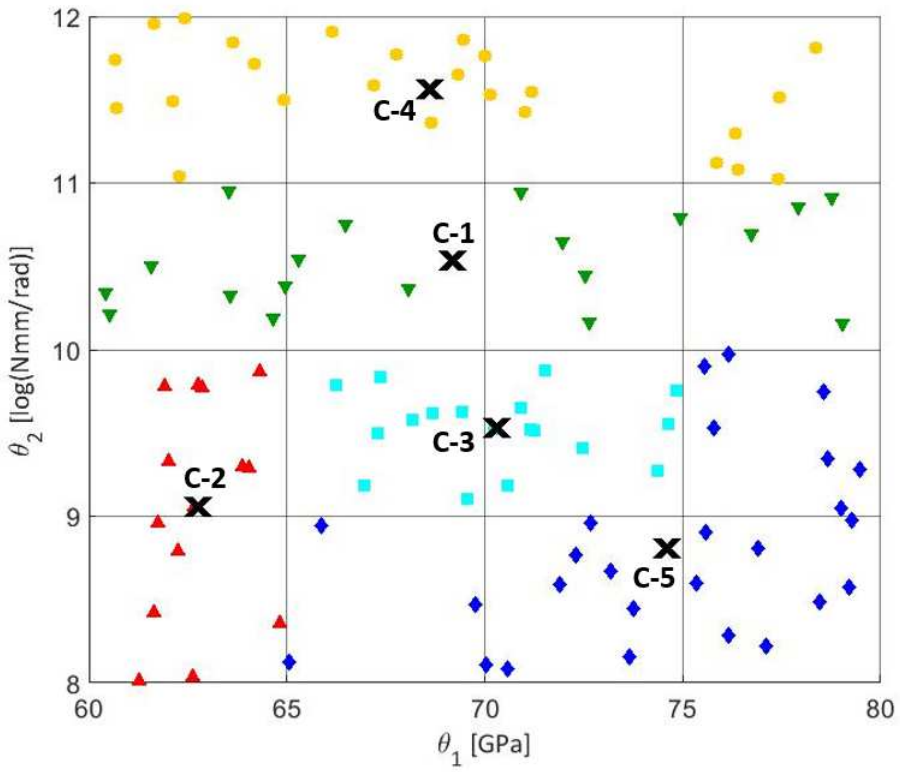

FIG. 6. BMO clustering visualization in the parameter space. Five clusters are identified using different symbols and centroid positions (X)

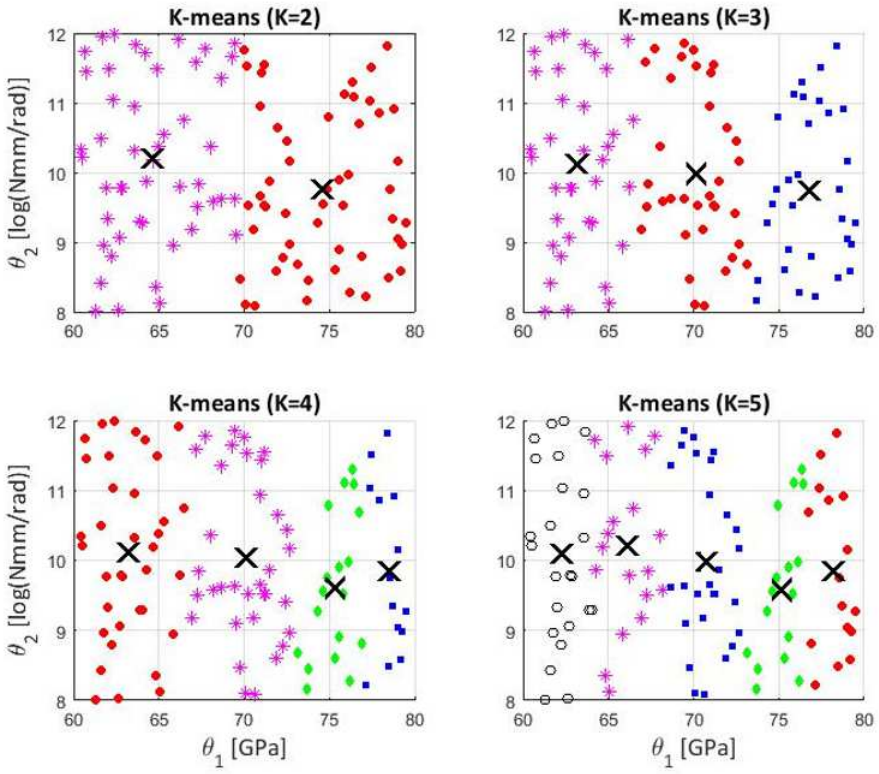

FIG. 7. K-means clustering visualization in the parameter space. Clusters are depicted using different symbols and centroid positions (X)

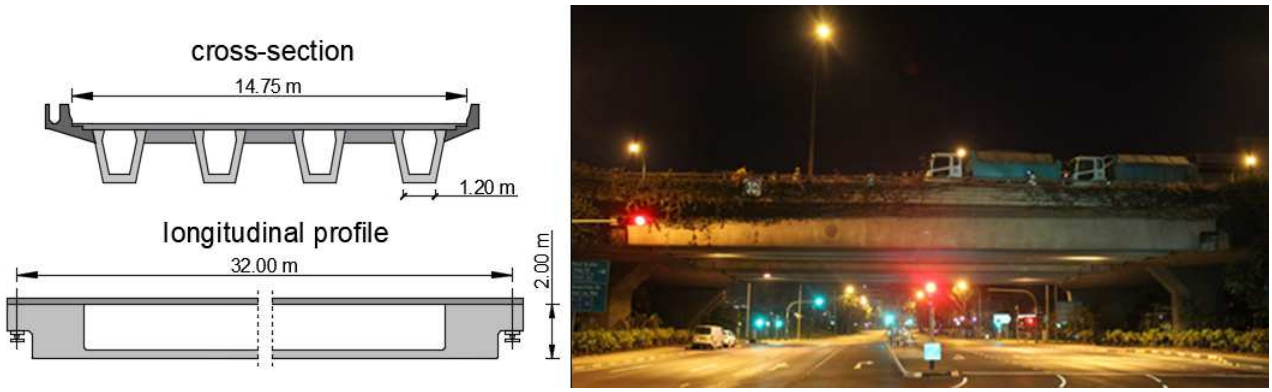

FIG. 8. Cross-section, longitudinal profile and view of the bridge in Singapore during the load test

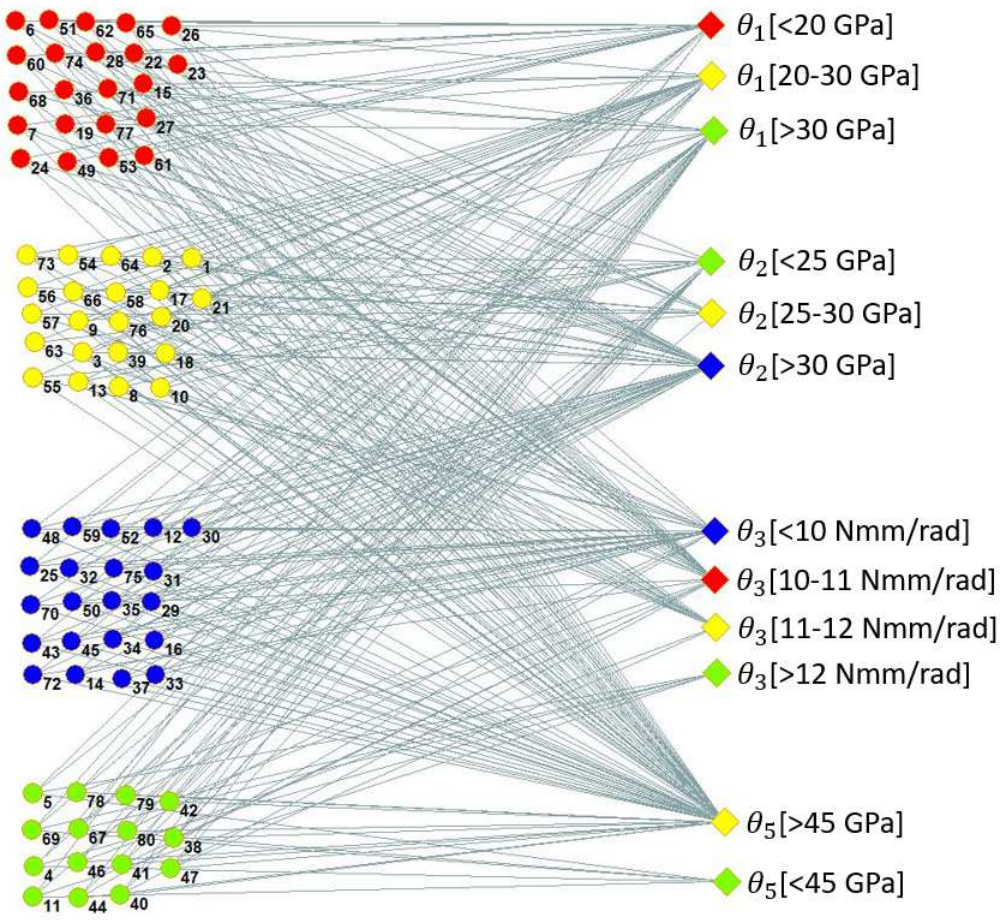

FIG. 9. Bipartite network and BMO clustering visualization (CMS B.1 4 parameters)

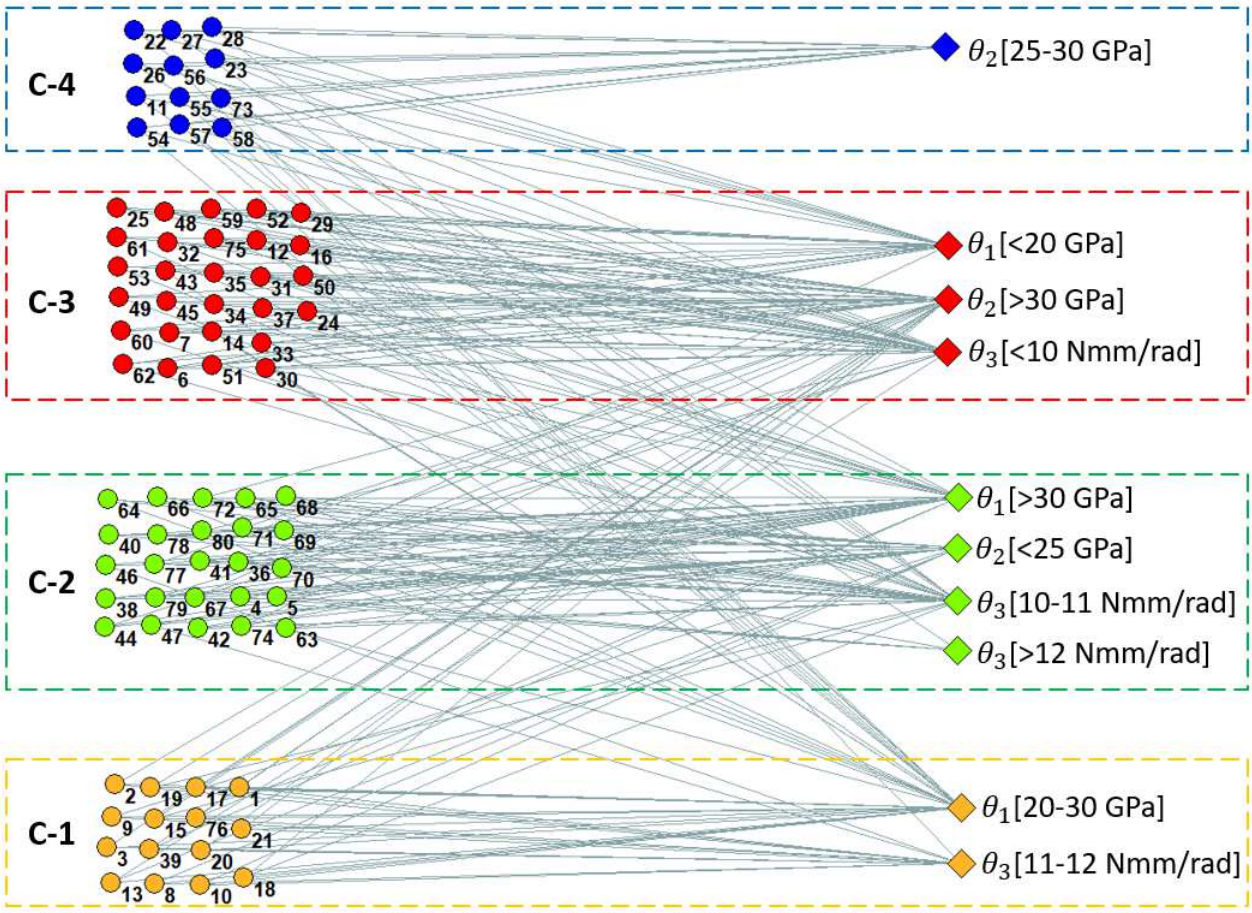

FIG. 10. BMO clustering visualization (CMS B.2 3 parameters) ©2017 IEEE

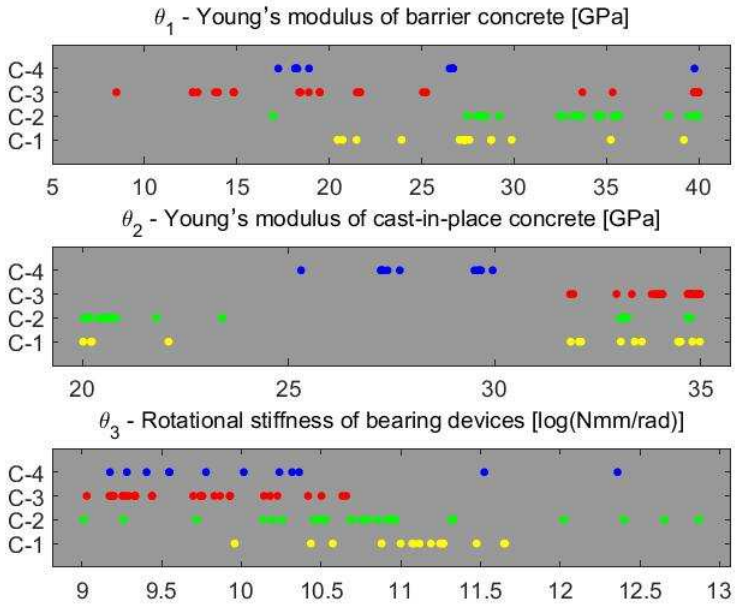

FIG. 11. Clustering of candidate models for each parameter ©2017 IEEE

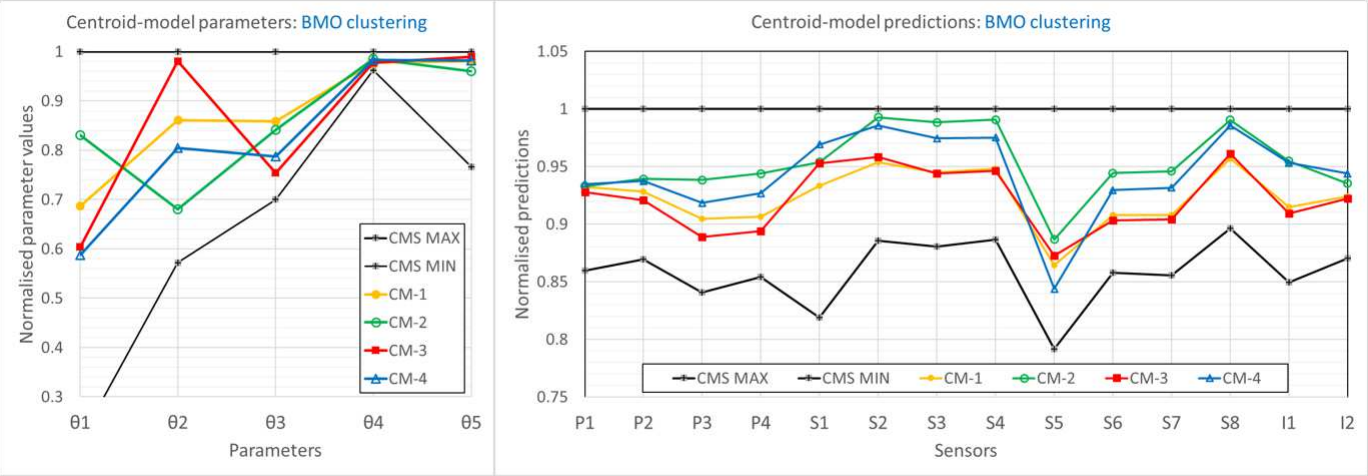

FIG. 12. Parallel-axis plot of parameter values that define centroid models (left) obtained using BMO clustering. CM predictions are within EDMF threshold bounds for each sensor location (right)

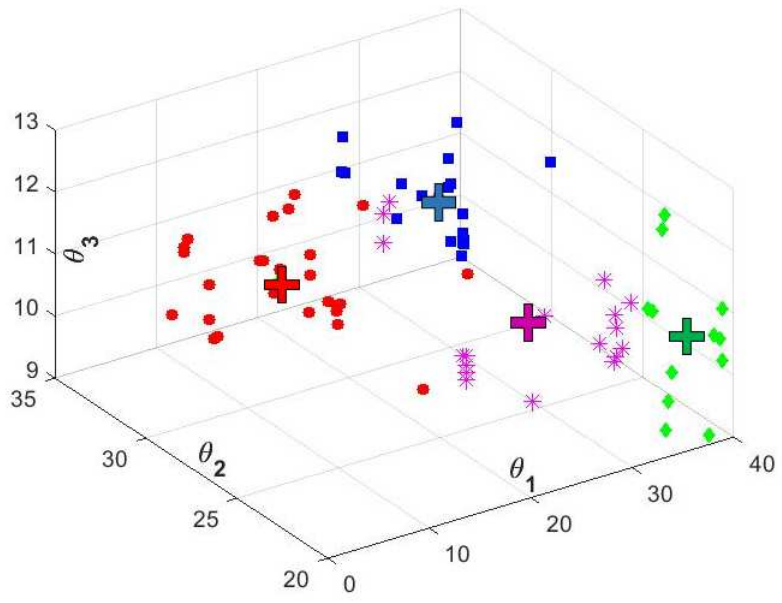

FIG. 13. K-means clustering visualization. Each node represents candidate model and clusters are indicated by different markers. Crosses represent cluster centroids ©2017 IEEE

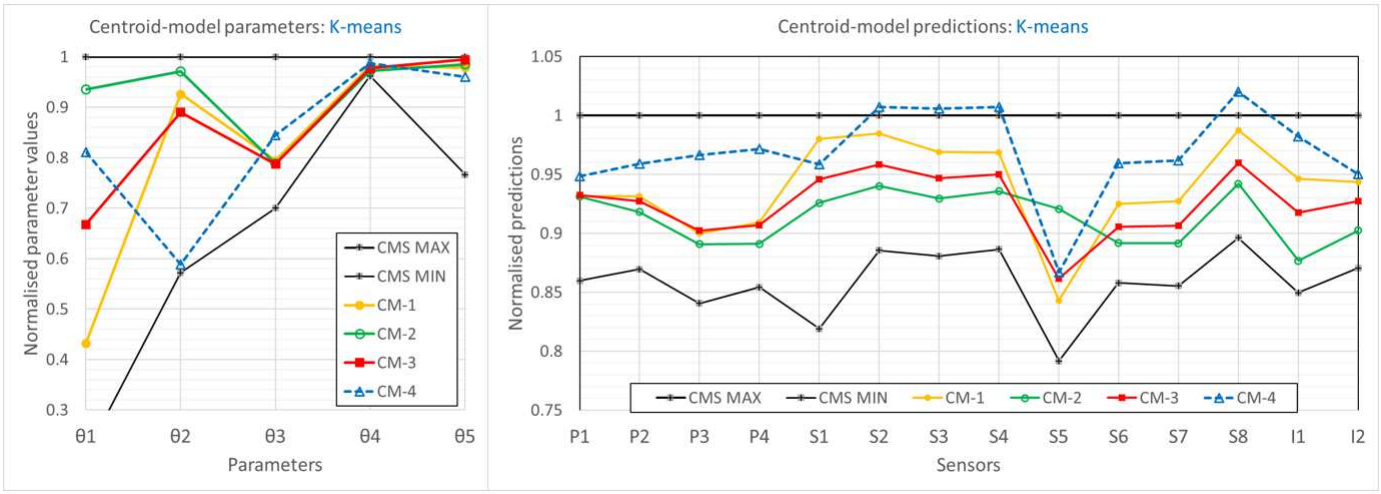

FIG. 14. Parallel-axis plot of parameter values that define centroid models (left) obtained using K-means (K=4). CM-4 is plotted with a dashed line since it is not compatible with measurements

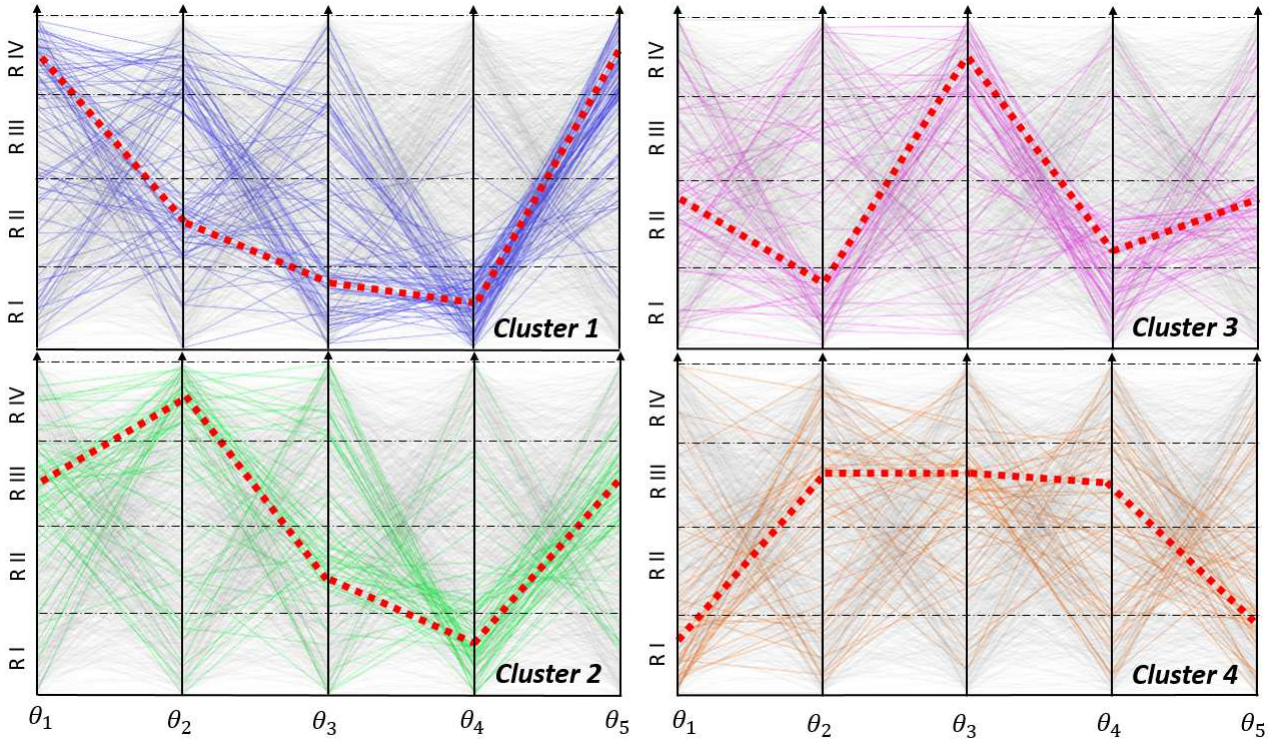

FIG. 15. Parallel axis plot of model parameters for BMO clustering (CMS B.2). Each vertical axis represents a parameter (divided into 4 ranges) and candidate models that belong to each cluster are plotted as coloured lines. To improve visualisation clusters are plotted separately and red dashed lines indicate cluster centroid models