ORIGINAL PAPER

Methodology for selecting measurement points that optimize information gain for model updating

Numa Bertola[1] · Ze Zhou Wang[2] · Wen-jun Cao[3] · Ian F. C. Smith[4]

© The Author(s) 2023

Abstract

Information collected through sensor measurements has the potential to improve knowledge of complex-system behavior, leading to better decisions related to system management. In this situation, and particularly when using digital twins, the quality of sensor data determines the improvement that sensors have on decision-making. The choice of the monitoring system, including sensor types and their configuration, is typically made using engineering judgement alone. As the price of sensor devices is usually low, large sensor networks have been implemented. As sensors are often used to monitor at high frequencies over long periods, very large data sets are collected. However, model predictions of system behavior are often influenced by only a few parameters. Informative data sets are thus difficult to extract as they are often hidden amid redundant and other types of irrelevant data when updating key parameter values. This study presents a methodology for selecting informative measurements within large data sets for a given model-updating task. By selecting the smallest set that maximizes the information gain, data sets can be significantly refined, leading to increased data-interpretation efficiency. Results of an excavation case study show that the information gains with refined measurement sets that are much smaller than the entire data set are better than using the data set prior to refinement for the same probability of identification, while the computational time of model updating is significantly reduced. This methodology thus supports engineers for significant data filtering to improve model-updating performance.

Keywords Structural identification · Digital twins · Measurement point selection · Error domain model falsification · Joint entropy

1 Introduction

The Internet of Things (IoT) is a system that consists of connecting devices, objects, and people through the exchange of data and information [1]. It is one of the key aspects guiding the development of responsive cities around the world.

However, there are challenges. When scaling up the IoT concept to a city level, "Things" (physical objects) become much more complicated than simple household items [2]. For example, when performance degrades, replacement is usually neither an easy option nor desirable for city elements.

Monitoring complex infrastructure systems, such as bridges, requires the recognition of several special characteristics. Effects, rather than causes, are generally measured. Data-only methods cannot support decisions related to activities such as retrofitting and replacement avoidance. Most measurement interpretation tasks require model-based diagnosis methodologies to infer causes from effects and then predict future behavior (prognosis), particularly when predictions involve extrapolation [3]. Additionally, uncertainty magnitudes and bias are much larger as city systems are strongly affected by boundary conditions such as the environment. At best, behavior models produce approximate predictions [4].

✉ Numa Bertola
  numa.bertola@epfl.ch

1   Laboratory for Maintenance and Safety of Structures, Ecole Polytechnique Fédérale de Lausanne (EPFL), Station 18, 1015 Lausanne, Switzerland

2   Department of Civil & Environmental Engineering, National University of Singapore, Block E1A, #07-03, No.1 Engineering Drive 2, Singapore 117576, Singapore

3   Department of Civil Engineering, The University of Hong Kong, Hong Kong, China

4   Georg Nemetschek Institute AI for the Built World, Technische Universität München, 8748 Garching, Germany

🖄 Springer

Eventually, when infrastructure "Things" are upgraded due to, for example, a change in demand, replacement should be the last option [5]. More sustainable options usually involve retrofitting and improvement rather than replacement [6, 7]. These actions need to be well-designed to control costs and minimize the impact on users.

The availability of inexpensive sensing devices [8, 9] and computing tools [10] has made it feasible to monitor civil engineering systems. Many bridges are monitored under a structural health monitoring framework [11–13]. A particular characteristic of civil engineering systems, compared with other engineering fields, is that civil structures are often designed using conservative, rather than accurate, behavior models. Also, construction practices often reflect the assumption that design specifications are minimum requirements. Most civil infrastructure thus has reserve capacity that is greater than the reserve which was intended by safety factors [14]. Decisions may also be conservative when asset management is made without quantifying this reserve capacity, leading to unsustainable and uneconomical maintenance activities.

Monitoring has the potential to improve understanding of structural behavior if it is followed by appropriate data interpretation methodologies. This process is called structural identification. Much research has been carried out to develop model-based data interpretation methodologies, see for example [15–17] among many others. Common data interpretation methodologies are residual minimization (RM) [18, 19], Bayesian model updating (BMU) [20–22], and more recently, a special implementation of Bayesian model updating called error domain model falsification (EDMF) [23, 24]. EDMF is an easy-to-use methodology that has been developed to be compatible with typically available engineering knowledge as well as standard civil-structure assessment concepts [25, 26]. These methodologies mostly differ in the criteria used for model updating and in the assumptions related to the quantification of uncertainties. Since each complex system has a unique combination of geometry and material properties as well as specific monitoring goals, the most appropriate data interpretation methodology is case-dependent. A methodology map to provide guidance for the appropriate use of these methodologies has been recently proposed by [27].

The performance of structural identification directly depends on the choice of sensor configuration [28, 29]. To maximize the information gain with a minimum number of sensors, researchers have developed sensor placement algorithms such as those proposed by [30–32], among others. Since the general task has exponential computational complexity with respect to the number of sensor types and locations, most researchers have used greedy algorithms to reduce the computational effort [33–35].

Several studies have focused on determining the best objective function for optimal sensor placement. Multiple approaches have been proposed that involve either minimizing the information entropy in either posterior model parameter distributions [36–38] or maximizing information joint entropy in multiple-model predictions have been proposed [39, 40]. Most approaches provide ranking of potential sensor locations based on their ability to identify either bridge parameter values or structural damage [41, 42]. In addition to structural monitoring, optimal sensor placement studies have involved wind predictions [43], wind turbine condition monitoring [44], and steel pipeline damage detection [45], among others.

Recently, a novel strategy called the hierarchical algorithm has been introduced that explicitly accounts for the shared information within sensor data sets through a joint entropy metric [39, 40]. This methodology has then been adapted to structural identification contexts such as multiple static tests and dynamic excitations [46] and multiple objective criteria [47], and has been validated using field measurements [48].

Evaluation of these sensor placement methodologies has shown that most sensor locations provide redundant information [49]. Measurement system design (also called optimal sensor placement when only one type of sensor is used) is usually carried out by engineers using only qualitative rules of thumb and estimations of signal-to-noise ratios, leading to sensor networks involving unnecessarily large numbers of sensors. As sensor placement algorithms are used for model updating, generation of model classes prior to monitoring as well as defining monitoring goals are necessary. In practice, such activities often take place only after monitoring is carried out. This is unfortunate since these methodologies can be used to filter data sets, thus improving data interpretation. In such cases, algorithms for optimal sensor placement should be adapted for the task of measurement point selection, which refers to the selection of the most informative data among previously collected data sets from multiple sensors. A measurement point is the data collected by a sensor at a given timestep. Each sensor thus collects multiple measurements throughout the duration of the measurement event.

The hierarchical algorithm has already been used to reduce sets of measurements in two applications: an excavation case study [49] and a wheel-flat identification [50]. The aim was to enhance data interpretation by either reducing the computation time or improving the precision of parameter value identification. Efficient methodologies for measurement point selection should involve four stages to avoid removing informative data points. Stages include model generation, data cleansing, data-point selection, and validation. Previous studies [49, 50] proposed stepwise processes for measurement point selection and did not account for the

influence of the number of measurements on the selection process, leading to non-optimal data sets.

This paper contains a proposal for a comprehensive methodology for the last three stages mentioned above to improve structural identification. In this methodology, a strategy is proposed for data cleansing that involves both an outlier detection methodology and filtering of measurement points that provide redundant information. A result validation scheme is also included involving an iterative process to define the optimal number of measurements.

The paper is organized as follows. The error domain model falsification is presented in the background section. The following section presents the methodology for measurement point selection. Then, the methodology is applied to a full-scale case study.

2 Background—error domain model falsification

Error domain model falsification (EDMF) is a recently developed probabilistic methodology for data interpretation that has been introduced by [23]. The particularity of this methodology is its robustness of the model-updating accuracy to variations in correlations between measurement points for most forms of uncertainty.

The methodology involves several steps. First, the model class is chosen. A model class involves the creation of a parametric model through selection of its most critical behavior-sensitive characteristics, such as material properties, geometry, boundary conditions, and excitations, as well as the quantification of nonparametric model uncertainties ($U_{i,g}$) and measurement errors ($U_{i,y}$). Additionally, possible ranges of model parameters that can be identified during monitoring are defined. Then, a set of model instances is generated, where each instance represents a unique combination of model parameter values. The output of EDMF is a set of plausible model instances (i.e., plausible combinations of parameter values) among the initial population through comparisons of model instance predictions with field measurements.

For a measurement point i, the predictions of a model instance $g_i(\Theta)$ are generated by assigning a unique set of parameter values Θ. The true structural response R_i (unknown in practice) is linked to the field measurements y_i and model prediction $g_i(\Theta)$ using Eq. (1), where n_y is the number of measurement points.

$$g_i(\Theta) + U_{i,g} = R_i = y_i + U_{i,y} \forall i \in \left\{ 1, \dots, n_y \right\} \tag{1}$$

Distributions of model $U_{i,g}$ and measurement $U_{i,y}$ uncertainties are combined in a unique source of uncertainty $U_{i,c}$

using Monte Carlo sampling. Equation (1) is then transformed in Eq. (2) as recommended by [51].

$$g_i(\Theta) - y_i = U_{i,c} \tag{2}$$

Plausible instances are selected through the falsification of instances that have residuals larger than defined threshold bounds. These thresholds are calculated using a probability target of identification ϕ on the combined uncertainty distribution $U_{i,c}$. This target ϕ is usually set to 95% [23]. When multiple measurements are considered, the Šidák correction [52] is used to maintain a constant level of confidence. Then, threshold bounds, $t_{i,low}$ and $t_{i,high}$ are evaluated (Eq. 3) for measurement point i. These bounds are calculated to be the shortest intervals in the probability density function of combined uncertainties $f_{U_i}(u_i)$.

$$\forall i = 1, \dots, n_y : \phi^{1/n_y} = \int_{u_{i,low}}^{u_{i,high}} f_{U_i}(u_i) du_i \tag{3}$$

The candidate model set (CMS),Ω'', contains unfalsified model instances defined using Eq. (4) from the initial model set Ω. These instances are set to be equally likely since little reliable information is usually available to describe the combined uncertainty distribution [51]. They are assigned an equal probability, while falsified model instances are assigned a null probability.

$$\Omega'' = \left\{ \Theta \in \Omega | \forall i \in \left\{ 1, \dots, n_y \right\} t_{i,low} \le g_i(\Theta) - y_i \le t_{i,high} \right\} \tag{4}$$

It may happen that all model instances are falsified. In such situations, model predictions are incompatible with measurements given the estimations of nonparametric uncertainty sources. This result means that the current model does not adequately reflect the true structural behavior. Using EDMF thus leads to the re-evaluation of assumptions related to the choice of model class and uncertainty quantification [53]. This situation highlights an important advantage of EDMF compared with other structural identification approaches, such as traditional implementations of BMU, particularly when there are few measurement points.

3 Methodology for measurement point selection

3.1 Overview

In this section, the methodology for selecting measurement points for data interpretation is presented. This methodology aims to improve the performance of data interpretation by removing measurements that do not provide information. By including only informative (i.e., useful) data, model updating requires less computational time than required for the

entire set. A potential increase in precision is linked to the target probability that is corrected (such as using the Šidák correction [52]) to remain constant, independent of the number of measurements.

Figure 1 presents the framework of the methodology for measurement point selection. This methodology is subdivided into two sides, the model side and the measurement side. Inputs on the model side generate a numerical model (such as a finite-element model) and define a model class.

A model class includes the most relevant parameters (i.e., parameters that have the largest influence on model predictions for a given test condition based on a sensitivity analysis). On the measurement side, the inputs are the selection of sensor types and locations.

The first stage involves the collection of data under test conditions, constituting the initial data set. On the model side, the model class is used to generate model instances with a unique combination of model parameter values that
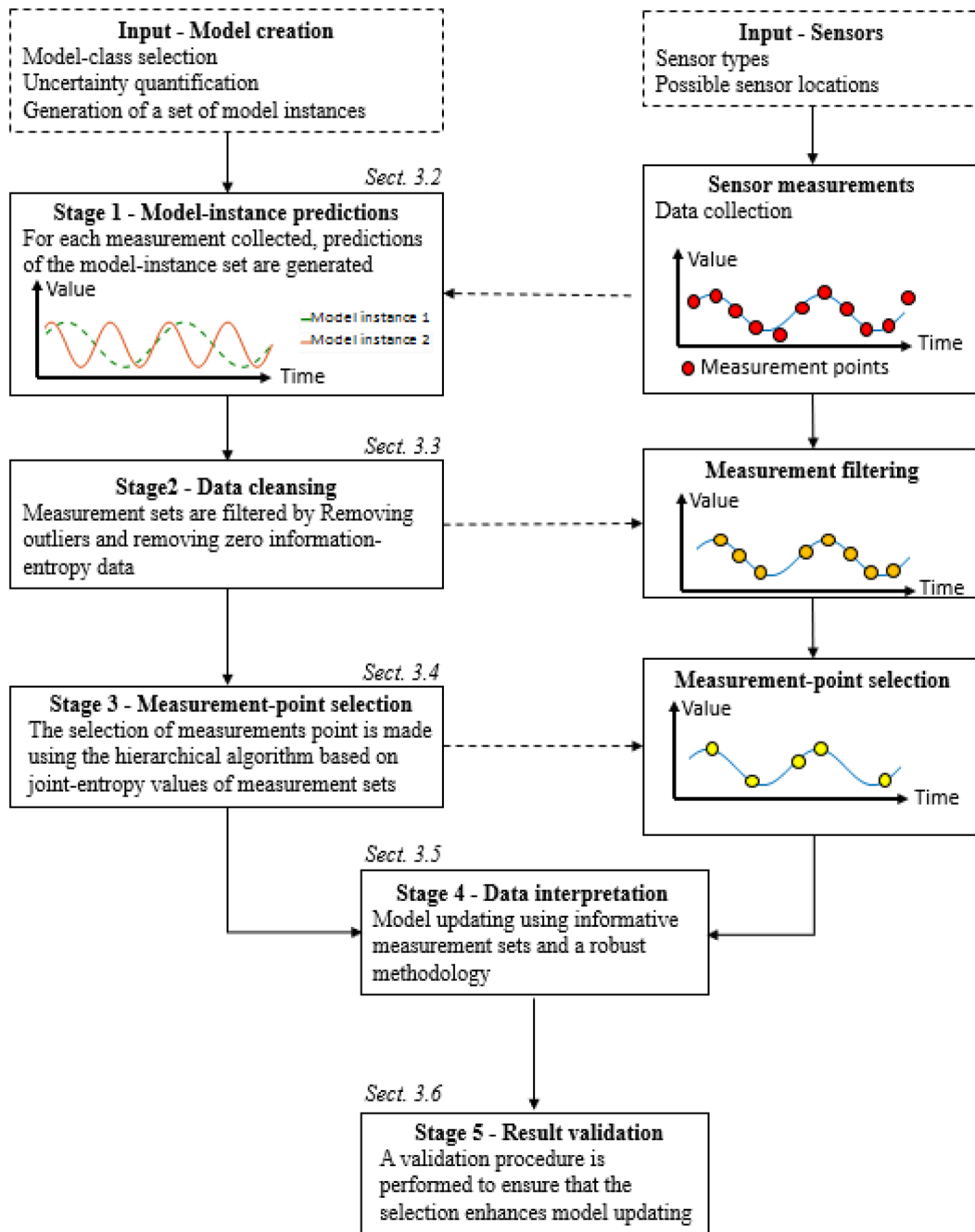

Fig. 1 Flowchart of the methodology for measurement point selection. Relevant paper section numbers are provided for parts of the flowchart

are within ranges. These parameter ranges are defined using engineering knowledge. These model instances are used to obtain predictions for each measurement collected. On the measurement side, field measurements are collected.

In the second stage, the initial data set is filtered using two model-based methodologies. First, outliers are removed. Measurements are defined as outliers if the value is outside the range of predictions, including nonparametric uncertainties. Then, non-informative measurements are removed. Non-informative measurements are defined using information entropy that evaluates the variability of model instance predictions for each measurement. On the measurement side, field measurements are filtered.

The third stage of the methodology involves selecting an optimal set of measurements using the hierarchical algorithm, following an iterative process. This algorithm evaluates the total information gain of a set of measurements, and the smallest set of measurements that maximizes joint entropy is chosen, and the remaining measurement points are discarded. Field measurements are then selected based on hierarchical-algorithm results.

In Stage 4, the model updating is made using the optimal set of measurements, reducing model parameter value ranges. The final stage (Stage 5) involves a validation procedure to ensure that the model updating is enhanced by comparing information gains using the entire data set and the selected measurement set. Using a smaller set of data, the precision of the model parameter value identification must be either equal or better in order to validate the results of the methodology.

3.2 Model class and model instance predictions

In this stage, the numerical model built for the inverse problem is instantiated in multiple model instances. Each instance is a distinct numerical model which has a unique combination of model parameter values. Each model instance has unique prediction values (Fig. 2). Then, the numerical model is used to generate model instance predictions for each measurement collected. These model instance predictions are the main input required for the measurement point selection in the next stages.

3.3 Data filtering
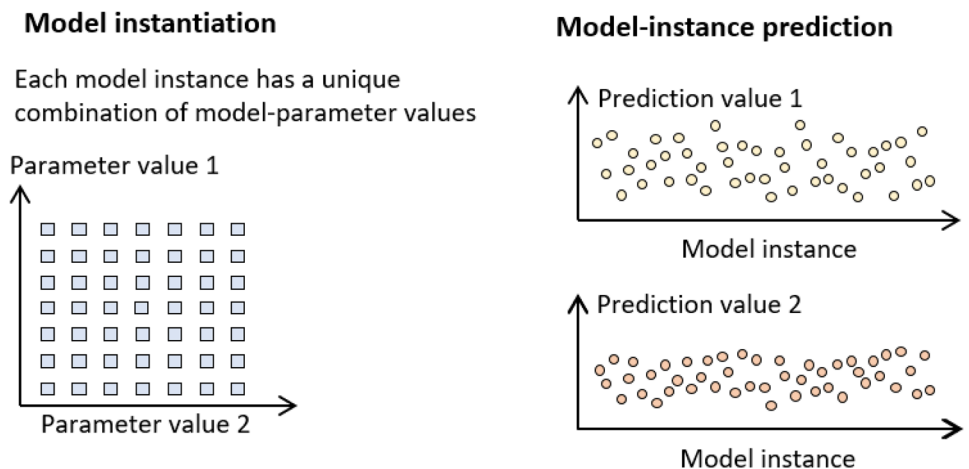
3.3.1 Outlier detection

This stage is divided into two steps. The first step involves using an outlier detection methodology to remove inconsistent data. When continuous data sets are involved, traditional outlier detection methods are designed to detect anomalies in measurements, see for example [54, 55] among others. When traditional methods are appropriate for the case study, they can also be used in this step.

Traditional outlier detection methods are not suitable for examining data sets that consist of non-time-dependent measurements, such as in some civil engineering situations. Also, conventional outlier detection methodologies based on statistical methods and signal processing are not suitable in the context of sparse measurements due to the relatively low number of measurements. Therefore, threshold bound solutions, as proposed below, are recommended [56].
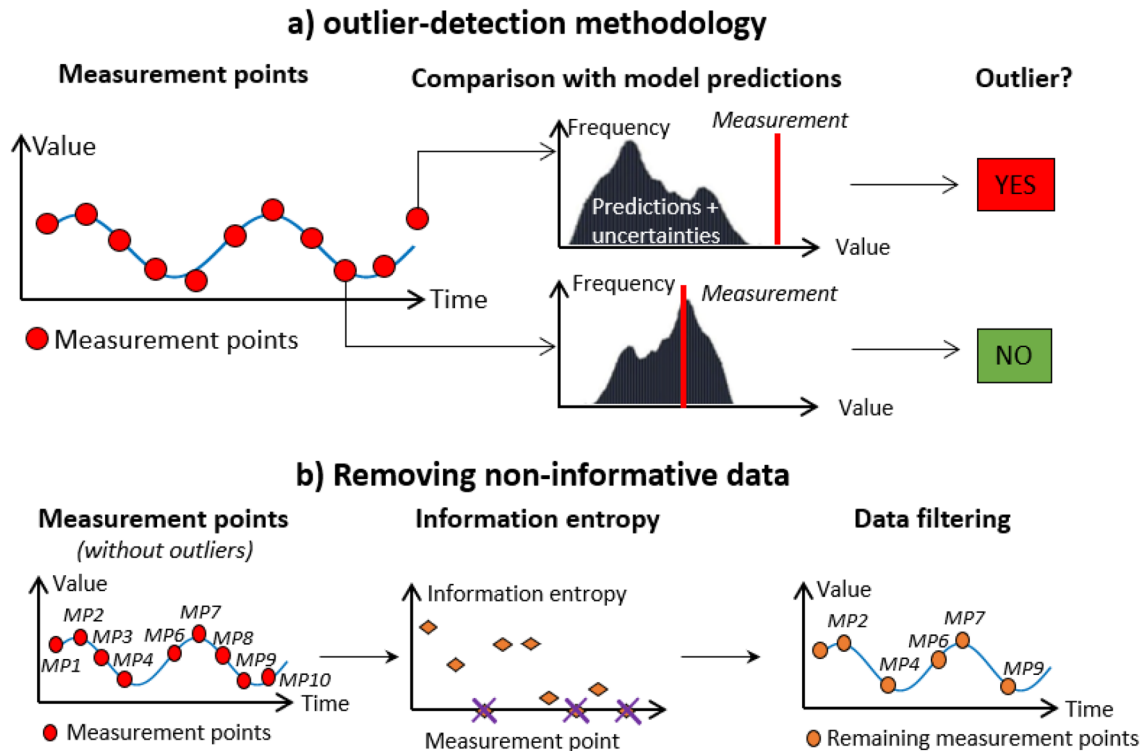
At each measurement point, a distribution is generated by combining the model instance prediction distribution with the combined uncertainty distribution using Monte Carlo sampling. This new distribution covers expected measurement values for a given current model class. A measurement point is judged to be an outlier if the measured value lies outside this prediction range (Fig. 3a). When an outlier is detected, it is then removed from the data set for the next stages.

An important assumption of the outlier detection methodology is that the model class that has been selected is correct. If a large number of the measurements (i.e., most

Fig. 2 Generation of model instance predictions

Model instantiation

Each model instance has a unique combination of model-parameter values

Parameter value 1

Parameter value 2

Model-instance prediction

Prediction value 1

Model instance

Prediction value 2

Model instance

a) outlier-detection methodology

Measurement points   **Comparison with model predictions**   **Outlier?**

b) Removing non-informative data

Measurement points   **Information entropy**   **Data filtering**
(without outliers)

Fig. 3 Data cleansing step. **a** Outlier detection methodology; **b** data filtering using information entropy

measurements on multiple sensors) does not lie in the distribution of model instance predictions, including uncertainties, it is possible that the selected model class is not appropriate. In such situations, the model class must be modified, either through changing the model parameters included in the set or correcting the nonparametric uncertainty estimations, among other measures [57].

When it is established that a sensor has provided several outlier measurements (i.e., at least 20% of outliers), all data collected by this device is excluded, even though some of the data falls within the range of predictions. For example, this situation may happen either when the device has malfunctioned or if it has not been properly installed.

3.3.2 Non-informative measurement point filtering

The second stage involves removing data points that do not provide information. In computer science, information gain is defined as a reduction in uncertainty due to the new availability of data. In this methodology, the information gain is evaluated by the ability of a measurement set to discriminate between model instances based on their predictions and uncertainties.

To evaluate this variability, a model instance prediction histogram is generated at each measurement point i.

The range of model instance predictions is subdivided into $N_{I,i}$ intervals where the interval width is given by the difference between threshold bounds taken from combined uncertainty $U_{i,c}$ (Eq. 2). In the present study, this definition is adapted to account for a unique uncertainty distribution for each model instance. Details are provided in Sect. 4.3.

The probability that the model instance prediction $g_{i,j}$ falls inside the j$^{\text{th}}$ interval is equal to $P\left(g_{i,j}\right) = m_{i,j} / \sum_{j=1}^{N_{I,i}} m_{i,j}$ with $m_{i,j}$ is the number of model instances falling inside this specific interval.

The information entropy $H\left(g_i\right)$ of a measurement point i is evaluated using Eq. (5). As the information entropy measures the disorder in predictions, measurement points associated with large information entropy values represent a high potential for model instance discrimination. In other words, when the value of information entropy is larger than zero, this measurement may help reduce parametric uncertainties and be informative.

By evaluating measurement point information entropy, non-informative data are removed from the set (Fig. 3b). This filtering removes data points where model instance predictions have no variabilities compared with the uncertainties that do not provide additional information. This filtering helps increase the speed of the measurement point selection in the next stage.

$$H(g_i) = -\sum_{j=1}^{N_{I,i}} P(g_{i,j}) \log_2 P(g_{i,j}) \qquad (5)$$

This step has two goals. First, it reduces the computational time of the measurement point selection as these optimization algorithms, as its computational complexity is $O(n^2)$ with respect to the number of measurement points. Secondly, it reveals data points that are not useful, thus improving the understanding of the information collected by the data set.

3.4 Measurement point selection

An optimization algorithm is used to rationally choose a minimum of measurement points that maximize the information gain. The aim is to find the measurement set that will minimize parameter value ranges after monitoring. This strategy involves two principal components: an objective function to evaluate the sets of measurements and an optimization scheme.

3.4.1 Assessing information gain using joint entropy

When systems are monitored, such as bridges, measurements at nearby sensor locations are typically correlated. Therefore, selecting measurement points based on their information entropy values alone leads to measurement sets that provide significant redundant information [40]. To account for the mutual information between measurement points, joint entropy has been introduced as a new objective function for sensor placement [39]. Joint entropy $H(g_{i,i+1})$ quantifies the information entropy between sets of predictions at measurement points. For a set of two measurement points, joint entropy is calculated following Eq. (6), where $P(g_{i,j}, g_{i+1,k})$ is the joint probability that model instance predictions fall inside the jth interval at measurement point i and the kth interval at measurement $i + 1$. In this equation, $k \in \{1, \ldots, N_{I,i+1}\}$ and $N_{I,i+1}$ is the maximum number of prediction intervals at the $i + 1$ measurement and $i + 1 \in \{1, \ldots, n_y\}$ with the number of measurement points n_y.

$$H(g_{i,i+1}) = -\sum_{k=1}^{N_{I,i+1}} \sum_{j=1}^{N_{I,i}} P(g_{i,j}, g_{i+1,k}) \log_2 P(g_{i,j}, g_{i+1,k}) \qquad (6)$$

Due to the redundancy in information gain between measurement points, the joint entropy is less than or equal to the sum of individual information entropies at measurements i and $i + 1$. Equation (6) can be changed to Eq. (7), where $I(g_{i,i+1})$ is the shared information between measurement points i and $i + 1$.

$$H(g_{i,i+1}) = H(g_i) + H(g_{i+1}) - I(g_{i,i+1}) \qquad (7)$$

For a given number of measurements, the set of measurements associated with the largest joint entropy will provide the largest reduction in model parameter values. Therefore, joint entropy is used as the objective function of the optimization scheme.

3.4.2 Optimal set of measurement points using the hierarchical algorithm

Due to a large number of possible combinations of measurement points, a greedy search algorithm is recommended to reduce the computational time, similarly to other sensor placement methodologies [36]. The hierarchical algorithm is a greedy search algorithm that efficiently evaluates joint entropy between sensor locations [39]. This algorithm is used for ranking measurement points based on the joint entropy of sets of measurements.

The ranking of measurement points is then used to select the appropriate number of measurements in the data set (Fig. 4). As long as the joint entropy is increasing, this means that additional measurements will increase the information gain (i.e., reducing model parameter ranges) after data interpretation. When the maximum joint entropy value is reached, this means that the remaining measurements are not helpful and should thus be removed from the data sets. The first set of selected measurements is defined as the minimum number of measurements to reach the maximum joint entropy value.
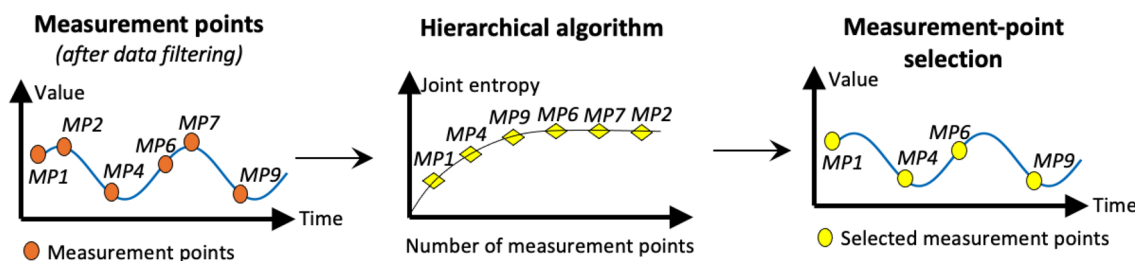

Fig. 4 Measurement point selection using the hierarchical algorithm

An important assumption of the hierarchical algorithm is that histogram intervals are calculated based on the number of measurements in the data sets. Intervals taken in the evaluation of the joint entropy value is directly related to the threshold bounds defined in EDMF (Sect. 2). These threshold bounds are corrected using the Šidák correction [52] which is used to maintain a constant level of probability with respect to the number of measurements. Thus, the number of measurements slightly influences the joint entropy values. This means that joint entropy values depend on this number (Eq. 6). When no results of the data filtering exist, such as during the first run, this number is equal to the initial size of measurement sets. An iterative process is then necessary until the initial assumption of the size of the selected measurement sets is correct. In each iteration, the hierarchical algorithm is run with the number of measurements selected in the previous iteration. Measurements selected in the last iteration constitute the data sets used for model updating.

3.5 Data interpretation

Once measurement points have been selected, these data are then used for model updating. Several methodologies exist. In this methodology, EDMF is recommended as it is an easy-to-use method that provides robust results in the presence of correlated and systematic uncertainties (Fig. 5). In this figure, the process of model instance falsification using EDMF is shown. Model instances in red are falsified as their predictions significantly differ from the measured values. All non-falsified model instances are included in the CMS. Nevertheless, the methodology for measurement point selection presented in this paper is also compatible with other data interpretation methodologies such as traditionally implemented BMU and residual minimization.

3.6 Validation

The last step of the measurement point selection methodology involves a validation procedure for the point selection. Such validation helps guarantee that the refinement of the

data sets does not affect the quality of model updating. It is made using the following two checks:

1. Falsification is performed with each point that has a zero entropy value (Fig. 3). These measurements must not alter the falsification result.
2. Falsification is performed with all data. Using reduced data sets, the precision of the data interpretation results should be equal to or better than using entire data sets.

If both criteria are satisfied, the measurement point selection is validated. This means that the selection of measurement points enhances the model-updating process since the computational time of the data interpretation method is reduced, and the precision of parameter value identification is equal to or higher than using all data sets. When these criteria are not satisfied, it means that either outliers have not been detected or that the model class selection was not appropriate. In such situations, an iteration of the entire methodology is required.

To compare data interpretation results, the falsification performance metric is introduced in Eq. (8) [48]. This metric is bounded between 0 and 1. When the falsification value is close to 1, information gain is significant, while a value of 0 shows that the measurements did not increase knowledge.

$$Falsification\,performance = 1 - \frac{Size\,of\,the\,candidate\,model\,set}{Size\,of\,the\,initial\,model\,set} \tag{8}$$

4 Case study—Bukit Panjang Hawker Centre excavation

4.1 Excavation numerical modeling

The case study is the excavation site of the Bukit Panjang Hawker Centre in Singapore. In this application, monitoring outcomes are used to improve the extrapolation of the retaining wall at subsequent excavation phases. Following

Fig. 5 Model updating using error domain model falsification. It is common for model predictions to cluster either at the upper or lower threshold due to the bias in the uncertainties
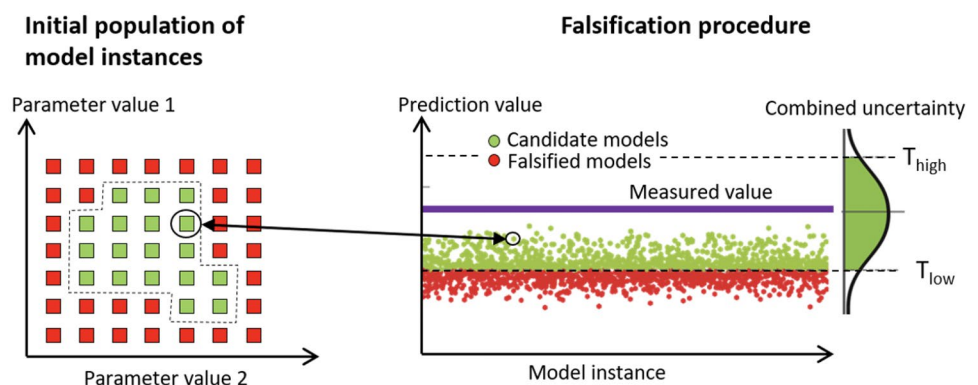
code requirements in Singapore, several inclinometers must be placed on the excavation site. Additionally, each sensor measures the wall-deflection profile at 15 heights during the four excavation phases. In this study, measurements involve 10–15 individual measurements made by the inclinometers during the four excavation phases.

The 10-m deep excavation is approximately 60 by 40 m large and is situated on the Bukit Timah Granite formation (Fig. 6) [58]. Six boreholes were drilled on the excavation site. The 3-m top layer consists of sandy silt and man-made backfill materials. This is underlain by a 10- to 13-m-thick layer of sandy silt, followed by the granitic rock layer at approximately 15 m below the ground surface. On the east side of the excavation site, a 5-m-thick layer of coarse sand is found between the sandy silt and the granitic rock.

4.2 Excavation numerical modeling

Two-dimensional numerical models are traditionally used to predict retaining wall behavior. However, the prediction accuracy is affected by 3D effects such as corner constraints [59]. To improve the quality of the model predictions, 2D and 3D numerical models are built on this case study [60]. The 2D model (Fig. 7a) is used to generate model instance predictions, while the 3D effects are quantified using interpolation of the prediction discrepancies between both models. Inclinometer locations, named S1 to S10, are displayed on the 3D model (Fig. 7b). In both models, soil layers are described using the hardening soil with small strain stiffness model [61], while the rock layer is depicted using the Hoek–Brown model [62]. The initial water table is 2 m below the ground level.

To retain earth, the support system includes diaphragm walls, soldier pile walls, toe pins, and two layers of steel struts and waler beams. The 800-mm-thick diaphragm

Fig. 6 Bukit Panjang hawker center excavation. Courtesy of Lian Soon Construction PTE LTD

walls are modeled as elastic plate members with reduced lateral stiffness due to the presence of construction joints. Toe pins and soldier pile walls are modeled as elastic plate members, while struts and waler beams are modeled as node-to-node anchors and beam elements, respectively. Soil–wall interactions are then modeled using an interface element without thickness. More information about the excavation modeling can be found in [49, 60].

The construction sequence modeled in the finite element analysis involves six phases that are presented in Table 1. The aim of phase 0A is to generate initial ground stresses. The diaphragm wall is "wished-in-place" in phase 0B, and installation effects are assumed to be negligible. Fully coupled flow deformation calculations are performed to account for the combined effects of groundwater flow and time-dependent consolidation. Phases 1 to 4 correspond to the four-stage analysis, where model predictions are compared with field measurements for model updating.

Based on a composite scale sensitivity analysis [63], four model parameters are identified as having the most influence on model predictions, and they are presented in Table 2. Parameter bounds are estimated using engineering judgment. On the four-dimensional space, 1000 model instances are generated using random sampling. Uncertainty sources present in this case study are listed in Table 3. As the 2D model does not consider the corner effects, its predictions are underestimated when compared with 3D model predictions. Therefore, the uncertainty source related to the 2D model simplification has only positive magnitudes. Other sources are estimated based on literature [23, 26, 53] and quantification methods reported in [60].

4.3 Adaptation of the hierarchical algorithm for excavation

The uncertainty magnitude of 3D effects involved in an excavation case study is influenced by model parameter values [60]. Therefore, falsification thresholds differ for each model instance. Consequently, the methodology requires the evaluation of the subset widths at sensor locations in the hierarchical algorithm. For each model instance, the combined uncertainty distribution at each measurement point location is computed. Then, the average of standard deviations $\sigma_{\mathrm{mean},i}$ of model instance combined uncertainty distributions at a measurement point i is calculated. The subset width at this measurement point is equal to $4 \times \sigma_{\mathrm{mean},i}$. This value represents the difference between upper and lower threshold bounds that are defined as two standard deviations from the mean value of predictions, similar to bridge case studies [40]. Finally, the information entropy of a single measurement point is evaluated by Eq. (5).

Fig. 7 Modeling of the excavation site **a** 2D model; **b** 3D model; adapted from [60]
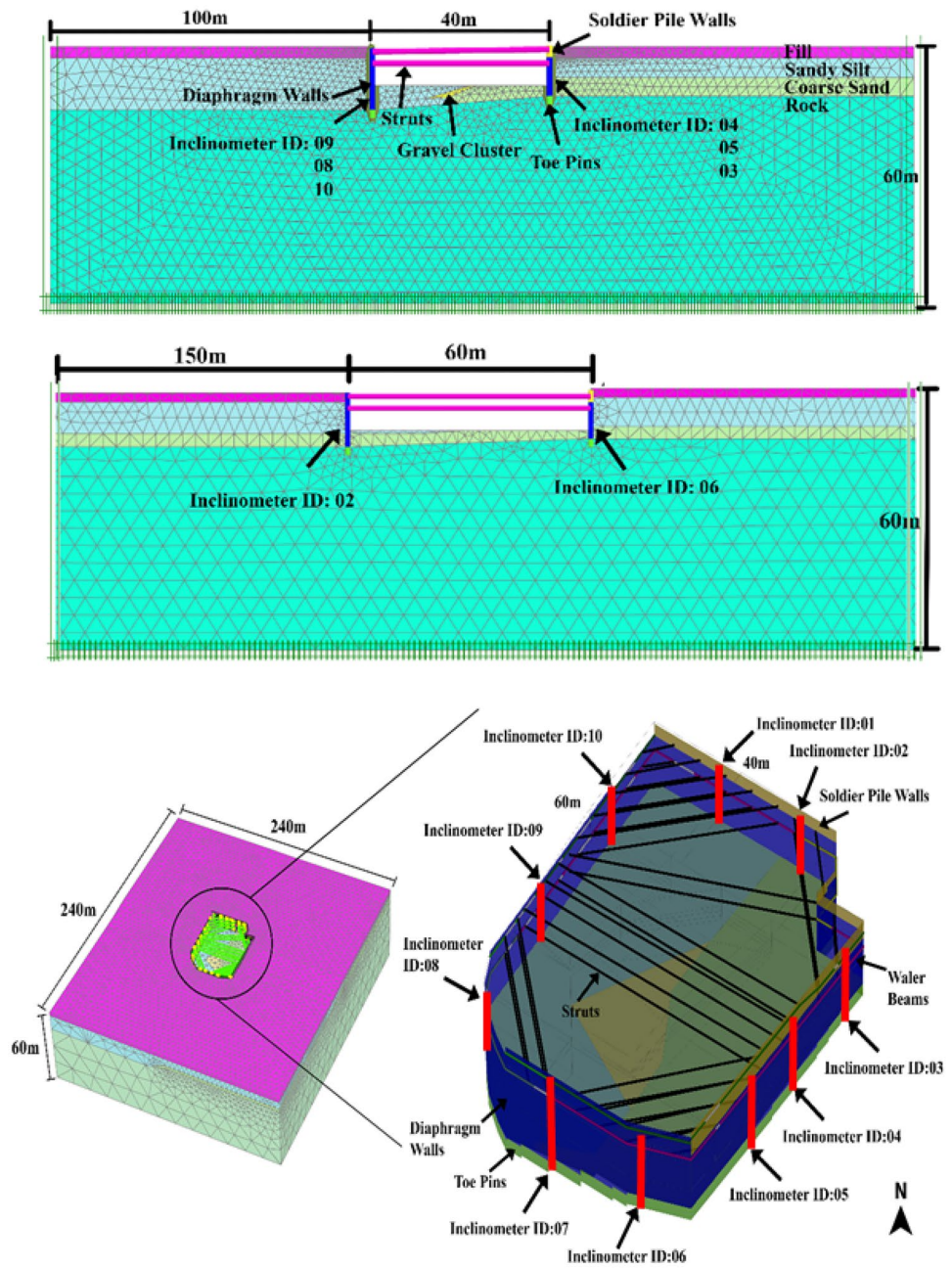
Table 1 Simplified excavation activities modeled in phases; adapted from [60]

Phases	Simplified excavation activities	Duration [days]	Calculation type
0A	Initial condition	–	Gravity loading
0B	Wall installation	–	Plastic
1	Excavation below strut layer 1	20	Fully coupled flow deformation
2	Install strut layer 1	45	Fully coupled flow deformation
3	Excavation below strut layer 2	20	Fully coupled flow deformation
4	Install strut layer 2 and excavation to formation level	30	Fully coupled flow deformation

Table 2 Initial intervals of model class parameters—Bukit Panjang excavation site; adapted from [60]

Parameters	Initial intervals
θ_1—Young's modulus of fill layer (E) [MPa]	[3; 20]
θ_2—Reference Young's modulus of silt layer (E_{50}^{ref}) [MPa]	[5; 50]
θ_3—Reference Young's modulus of sand layer (E_{50}^{ref}) [MPa]	[5; 50]
θ_4—Equivalent wall flexural rigidity [MNm^2]	[3; 10]

Table 3 Uncertainty sources—Bukit Panjang excavation site; adapted from [60]

Uncertainty sources	Magnitudes	Distribution
Inclinometer uncertainties	± 3.5 mm	Gaussian
2D model simplification	+ 4% to + 12%	Gaussian
Response surface approximation	± 2.5 mm	Uniform
Mesh refinement	± 5%	Uniform
Other sources	± 5%	Uniform

4.4 Data filtering

4.4.1 Outlier detection

This step involves data cleansing to remove doubtful measurements. The outlier detection methodology removes data from the measurement sets if the measurements fall outside the 99% interval of the distribution of the combined uncertainties sum with the distribution of model instance predictions.

Results show that 30 measurements have been detected as outliers. These outliers fall into two categories. The first category involves measurements at the bottom of the excavation. In this case, the model predicts zero value inclination and has a distribution of combined uncertainties with a small standard deviation. A small deviation (in absolute values) of the measurements may result in the measurement being qualified as an outlier. This category includes most of the suspicious measurements, representing 20 out of the 28 measurements. The second category includes faulty measurement that occurs randomly.

Figure 8 presents the results of the outlier detection for two measurements. In the first case (Fig. 8a), the measurement falls within the distributions of the model predictions and combined uncertainties, while the second measurement value is clearly outside the expected range (Fig. 8b). Measurement 237 is thus taken to be a reliable measurement,

Fig. 8 Results of the data cleansing. **a** Outlier detection: Measurement 237; **b** Outlier detection: Measurement 435. **c** Distribution of information entropy values of measurement points; **d** Zoom on the distribution of the information entropy values of measurement points (y-axis bounded to 50)
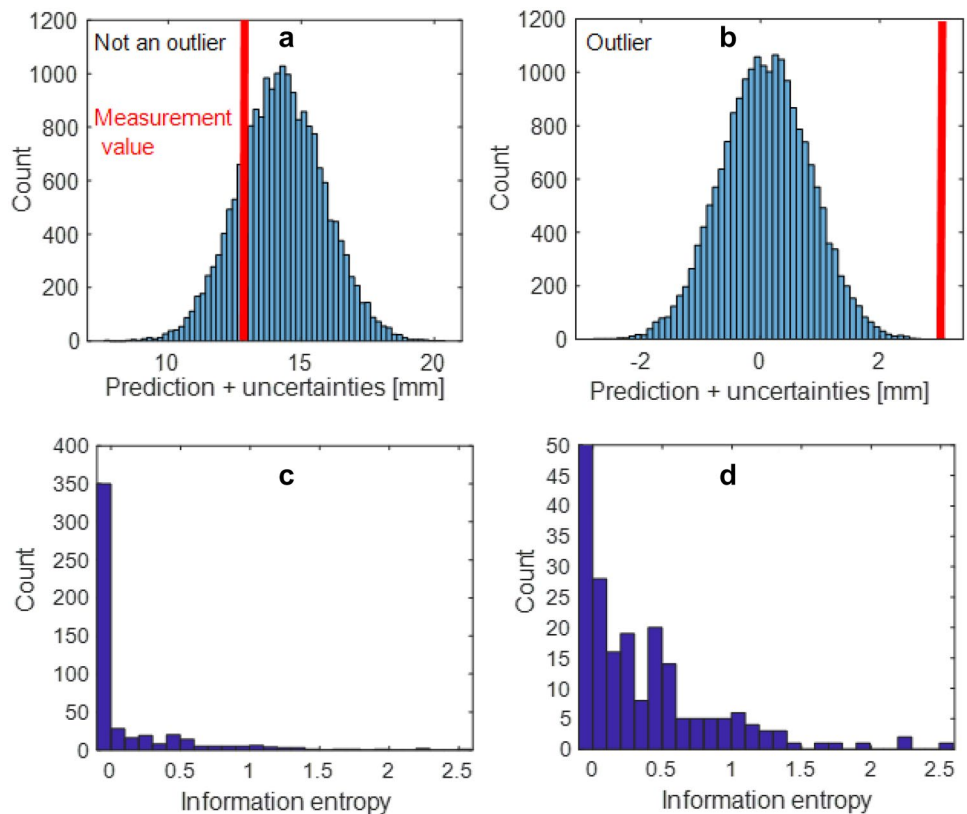
while Measurement 435 is an outlier and is removed from the measurement set.

4.4.2 Information entropy

The second step involves removing data points that are non-informative. This data cleansing is made by evaluating the information entropy of model instance predictions for each measurement point individually.

When the measurement point is associated with an information entropy value equal to zero (Eq. 6), this means that the variability of model instance predictions is very low compared to the combined uncertainties for this data point, meaning that uncertainties have much larger magnitudes than prediction ranges. In such cases, no model instances will be falsified by this measurement, and this measurement is thus not providing any information. This is typically the case when the signal-to-noise ratio of a measurement point is low. Nonetheless, the physical justification of a non-informative measurement point may not be trivial.

Figures 8c, d present the distribution of information entropy values for the 498 measurement points after outlier detection (y-axis bounded to 50). The 320 measurement points have a zero entropy value, meaning that they can be removed from the measurement sets without affecting the information gain during data interpretation. The remaining 148 measurements have non-zero information entropy and are potentially providing useful information. These data points constitute the initial sets of potential measurements for the next step of the process.

4.5 Measurement point selection

4.5.1 Joint entropy assessment

In this stage, the hierarchical algorithm is used to evaluate the joint entropy values of sets of measurements. The aim is to find the smallest number of measurements that maximize the joint entropy (i.e., the expected information gain). As joint entropy values and selected measurement points are affected by the hypothesis of the number of measurements that will be needed (Eq. 6), it is an iterative process.

Figure 9 presents the joint entropy values with respect to the number of measurements. Results of two hypotheses on the number of measurements are shown: 528 measurements (initial number of measurements), and 36 after the iterative process that is proposed in this paper. When the hypothesis involves a smaller number of sensors, the number of useful measurements and the maximum joint entropy values increase. As the Šidák correction increases threshold bound values with the number of measurements (Eq. 5), smaller information gain is possible with more measurements for this hypothesis.
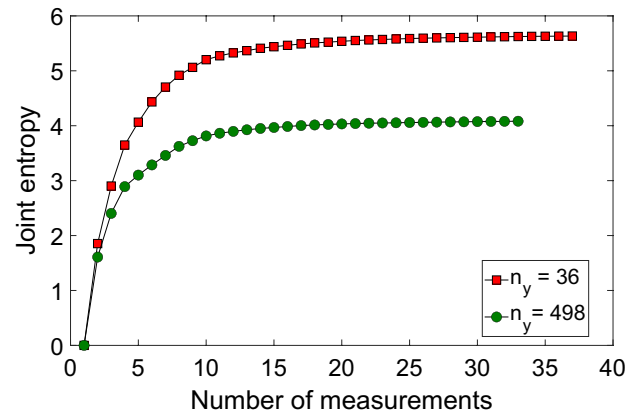
Fig. 9 Joint entropy as a function of number of measurement points for a constant probability of identification. Both data sets show a diminishing increase in joint entropy to nearly zero after 15 measurements. However, the highest entropy is achieved with the smaller data set

Table 4 Size of measurement sets with respect to the step of the methodology

Initial number of measurement points	Measurement set size after outlier detection	Measurement set after information entropy filtering	Measurement set after selection with the hierarchical algorithm
528	498	148	36

4.5.2 Optimal set of measurements

On the initial 528 and 148 measurements after data cleansing, only 36 measurements provide the most useful information (Table 4). These measurements constitute the optimal set of measurements for data interpretation. This corresponds to a reduction of 93% of the initial set and a reduction of 76% compared with the set after data cleansing. This selection must not compromise the information gain, and this will be assessed in the last step of the analysis.

4.6 Result validation

The last step involves validating the measurement point selection by comparing the true information of the selected set of measurements with the entire set after the outlier detection methodology. This step ensures that no information is lost during the data filtering. Field measurements from inclinometers (except I1 and I7) and all phases of excavation are shown in Fig. 10. I1 and I7 are not included due to the small deflection measured at all excavation phases.

Table 5 presents the comparison of the falsification performance and CMS of sets of measurements. Five sets are shown, and they include various numbers of measurements.
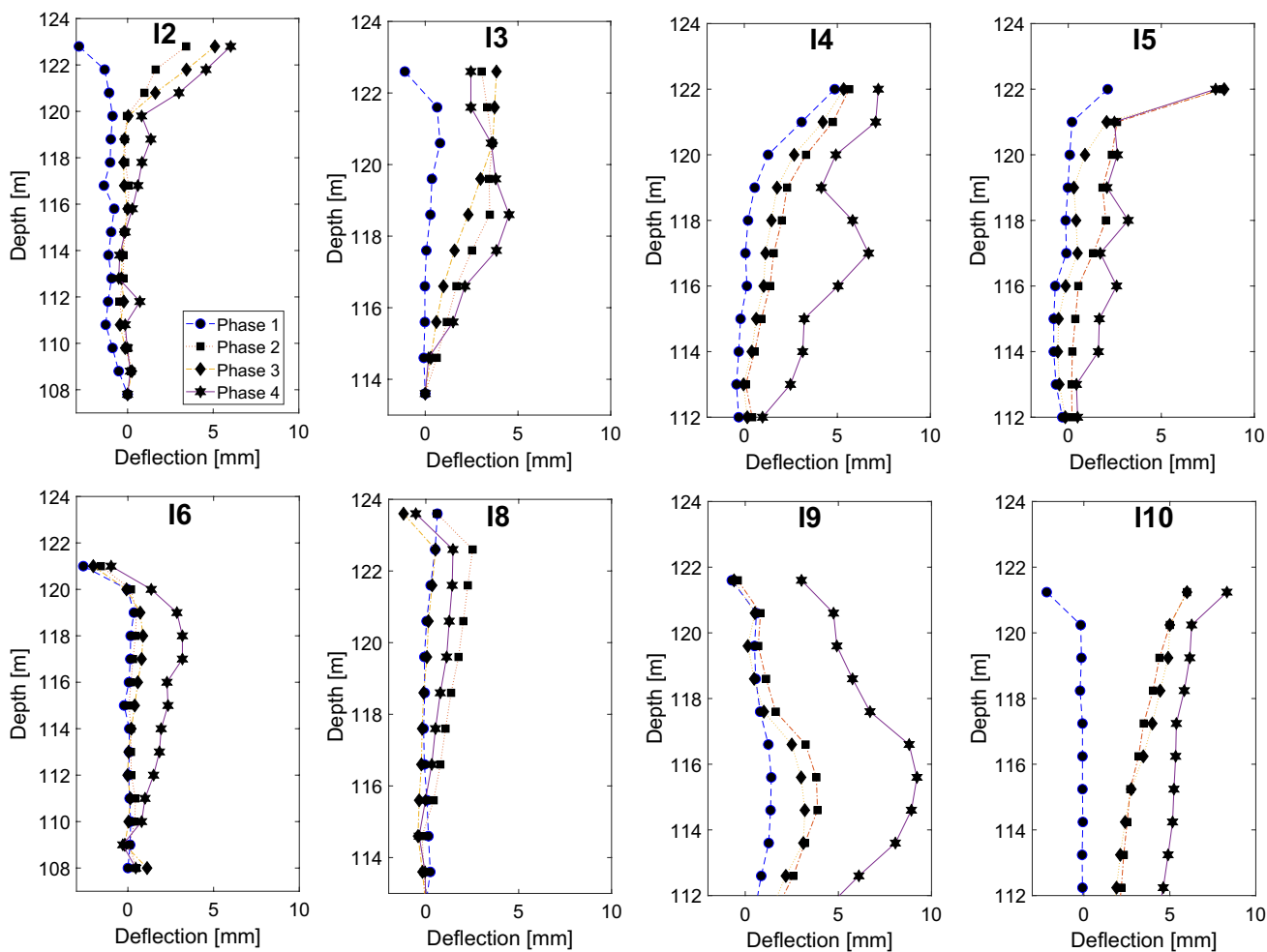
Fig. 10 Deflection measurements by the inclinometers at the four phases of the excavation

Table 5 Falsification performance for several sets of measurement points

Measurement sets	Number of measurements in the set	Candidate model set after falsification	Falsification performance
All measurements after outlier detection	498	8934	0.55
Measurements after data cleansing	148	4420	0.78
Recommendations in (Wang et al. [49])	34	2995	0.85
Proposed methodology	36	2214	0.89
Non-informative data set	350	20000	0

The falsification performance significantly increases with the data selection methodology proposed in this paper. From the initial 20,000 model instances, only 55% are falsified when all measurements are considered in the falsification process. This falsification performance increases to 89% with the proposed methodology, showing a significant increase in information gain. As the Šidák correction also increases threshold bounds for falsification (Eq. 3), adding more data in the measurement set decreases the identification performance as most data do not provide any information.

Compared to the previous methodology [49], a similar data set is selected (36 instead of 34 measurements). However, the information gain is improved as an additional 781 (26%) model instances have been falsified as not the same measurements are included in the set. This result is explained by the iterative process proposed in the methodology that

helps evaluate the information gain from each measurement point more accurately.

A two-step process has been introduced in the methodology section. The first step involves the falsification performance of all non-informative data points. Following the results of the data cleansing section, this set includes 148 measurement points. When the falsification is performed with this set, all initial model instances remain in the candidate model set, confirming the usefulness of these measurements. The second step of the validation process involves comparing the falsification performance of the entire data set after the outlier detection with the set after the measurement point selection. In the present case study, the latter set enables a better falsification performance than the entire data set. Results of the measurement point selection are thus validated. Explanations for this better falsification performance are provided in the next section.

5 Discussion

For a constant probability of identification, the measurement point selection algorithm helps significantly reduce data sets required for structural identification without compromising the falsification performance. In the Bukit Panjang case study, such significant data selection has even improved the falsification performance. This result contradicts the paradigm in information theory "information never hurts" [64, 65].

This contradiction can be partially explained by the data interpretation methodology used. EDMF [23], used in the study, involves the Šidák correction to maintain a given probability of identification in the data interpretation. By using this correction, threshold bounds increase when the number of measurements increases. This increase leads to a reduction in the information gain of measurements with an increase in the number of measurements. The contradiction with the information-never-hurts paradigm has also been observed for the residual minimization method [49]. Additional work should be performed to evaluate the influence of the number of measurements on the information gain when traditionally implemented BMU [30, 32, 37] is used for data interpretation.

Figure 11 presents the influence of the number of measurements on the measurement point performance in terms of information gain for the full-scale case study. Results are shown for the expected information gain, calculated using information entropy (Eq. 5) and observed information gain (Eq. 8) using the best measurement of the case study. The falsification performance of the measurement point decreases with an increase in the number of measurements, showing that adding measurements to the data set reduces the individual information gained by each measurement
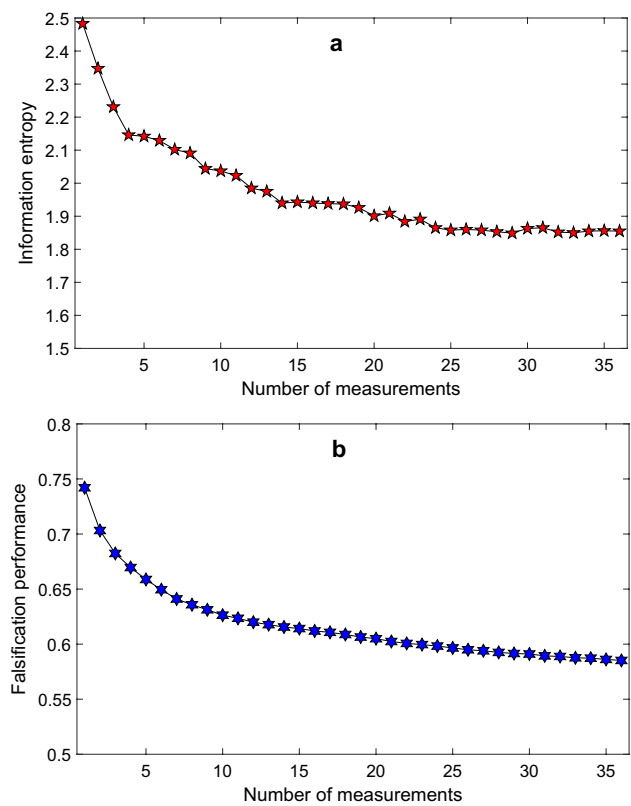
Fig. 11 Information gain of the best measurement point as a function of the number of measurements. **a** Expected information gain (information entropy); **b** observed information gain (falsification performance)

point. The optimal set of measurements is thus a tradeoff between adding a new measurement with unique information and reducing the information gain of other measurements in the set.

The expected information gain (Fig. 11a), measured with information entropy, also decreases with increasing the number of measurements. As the information gain per measurement may vary with respect to the size of the data set, an iterative process to reach the optimal set of measurement points is necessary. This iterative process also leads to a new selection of measurement points as the unique information gain provided by each measurement is more accurately assessed. Even if the total number of measurement points is similar, the falsification performance of the measurement sets may be significantly improved by the iterative process, as shown in Table 5.

The following limitations of the work are recognized. In the case study, the combined uncertainties used in the hierarchical algorithm are calculated as the mean values of combined uncertainties associated with the 1000 model instances in order to reduce the computational time. This simplification may affect the evaluation of the expected measurement point performance, especially in terms of joint entropy

evaluation. Also, this simplification may explain small discrepancies between the results of the hierarchical algorithm and the falsification performance of measurement points. Additionally, Monte Carlo simulations (using 1,000,000 samples) are used to generate threshold bounds which may slightly affect evaluations of expected and observed performances.

This study shows that algorithms that were initially built for measurement system design algorithms, such as the hierarchical algorithm, have the potential to be used for new applications such as measurement point selection. In the context of smart cities and the Internet of Things (IoT), design frameworks are increasingly becoming data-dependent. However, when inexpensive sensor devices are available, the amount of data increases exponentially. In these situations, large data sets are difficult to store, manage and analyze. This often results in weak data interpretation. A systematic and rational method for measurement point selection supports the extraction of valuable information among data sets and leads to the potential for better and more timely decision-making than that occurs in current practice.

6 Conclusions

In this paper, a methodology is proposed to select an optimal set of measurements in large data sets. A full-scale case study has been used to illustrate the proposed methodology. This case study involves updating the numerical model of the excavation of the Bukit Panjang Hawker Centre in Singapore through a model falsification approach. Results have shown that refined data sets provide more information than full data sets for the same probability of identification. Specific conclusions are as follows:

- The methodology for measurement point selection supports engineers in the task of selecting informative measurement points to improve model-updating performance.

  - Reducing the number of measurements using the proposed methodology helps avoid over-instrumentation without reducing the quality of the identification.
  - By comparing identification performance using all measurement points and the subsets, methodologies for measurement point selection are effectively validated using field measurements.

Declarations

References

1. Ashton K (2009) That "Internet of Things" thing. RFID J. 22:97–114
2. Atzori L, Iera A, Morabito G (2010) The Internet of Things: a survey. Comput Netw 54:2787–2805. https://doi.org/10.1016/j.comnet.2010.05.010
3. Catbas F, Kijewski-Correa T, Lynn T, Aktan A (2013) Structural identification of constructed systems. American Society of Civil Engineers, Reston
4. Friswell MI (2007) Damage identification using inverse methods. Philos Trans R Soc Lond A Math, Phys Eng Sci 365:393–410
5. Brühwiler E (2020) UHPFRC technology to enhance the performance of existing concrete bridges. Struct Infrastruct Eng 16:94–105. https://doi.org/10.1080/15732479.2019.1605395
6. Brühwiler E, Denarié E (2013) Rehabilitation and strengthening of concrete structures using ultra-high performance fibre reinforced concrete. Struct Eng Int 23:450–457. https://doi.org/10.2749/101686613X13627347100437
7. Hajiesmaeili A, Pittau F, Denarié E, Habert G (2019) Life cycle analysis of strengthening existing RC structures with R-PE-UHP-FRC. Sustainability 11:6923. https://doi.org/10.3390/su11246923
8. Lynch JP, Loh KJ (2006) A summary review of wireless sensors and sensor networks for structural health monitoring. Shock Vib Digest 38:91–130
9. Taylor SG, Raby EY, Farinholt KM et al (2016) Active-sensing platform for structural health monitoring: development and deployment. Struct Health Monit 15:413–422
10. Frangopol DM, Soliman M (2016) Life-cycle of structural systems: recent achievements and future directions. Struct Infrastruct Eng 12:1–20
11. Cachot E, Vayssade T, Virlogeux M et al (2015) The Millau Viaduct: ten years of structural monitoring. Struct Eng Int 25:375–380. https://doi.org/10.2749/101686615X14355644770776
12. Cremona C, Santos J (2018) Structural health monitoring as a big-data problem. Struct Eng Int 28:243–254. https://doi.org/10.1080/10168664.2018.1461536

13. Wong K-Y (2007) Design of a structural health monitoring system for long-span bridges. Struct Infrastruct Eng 3:169–185

14. Smith IFC (2016) Studies of sensor data interpretation for asset management of the built environment. Front Built Environ 2:2–8

15. Beck JL (2010) Bayesian system identification based on probability logic. Struct Control Health Monit 17:825–847. https://doi.org/10.1002/stc.424

16. Cross EJ, Worden K, Farrar CR (2013) Structural health monitoring for civil infrastructure. Health assessment of engineered structures. World Scientific, Singapore, pp 1–31

17. Moon F, Catbas N, Çatbacs FN et al (2013) Structural identification of constructed systems. Structural identification of constructed systems. American Society of Civil Engineers, Reston, pp 1–17

18. Beven K, Binley A (1992) The future of distributed models: model calibration and uncertainty prediction. Hydrol Process 6:279–298. https://doi.org/10.1002/hyp.3360060305

19. Alvin K (1997) Finite element model update via Bayesian estimation and minimization of dynamic residuals. AIAA J 35:879–886. https://doi.org/10.2514/2.7462

20. Beck JL, Katafygiotis LS (1998) Updating models and their uncertainties. I: Bayesian statistical framework. J Eng Mech 124:455–461. https://doi.org/10.1061/(ASCE)0733-9399(1998)124:4(455)

21. Behmanesh I, Moaveni B, Lombaert G, Papadimitriou C (2015) Hierarchical Bayesian model updating for structural identification. Mech Syst Signal Process 64:360–376

22. Katafygiotis LS, Beck JL (1998) Updating models and their uncertainties. II: model identifiability. J Eng Mech 124:463–467

23. Goulet JA, Smith IFC (2013) Structural identification with systematic errors and unknown uncertainty dependencies. Comput Struct 128:251–258. https://doi.org/10.1016/j.compstruc.2013.07.009

24. Pasquier R, Smith IFC (2015) Robust system identification and model predictions in the presence of systematic uncertainty. Adv Eng Inform 29:1096–1109. https://doi.org/10.1016/j.aei.2015.07.007

25. Pai SGS, Reuland Y, Smith IFC (2019) Data-interpretation methodologies for practical asset-management. J Sens Actuator Netw 8:36. https://doi.org/10.3390/jsan8020036

26. Proverbio M, Vernay DG, Smith IFC (2018) Population-based structural identification for reserve-capacity assessment of existing bridges. J Civil Struct Health Monit. 8:1–20

27. Pai SGS, Smith IFC (2022) Methodology maps for model-based sensor-data interpretation to support civil-infrastructure management. Front Built Environ. https://doi.org/10.3389/fbuil.2022.801583

28. Papadimitriou C, Beck JL, Au S-K (2000) Entropy-based optimal sensor location for structural model updating. J Vib Control 6:781–800

29. Papadimitriou C (2005) Pareto optimal sensor locations for structural identification. Comput Methods Appl Mech Eng 194:1655–1673

30. Ercan T, Papadimitriou C (2021) Optimal sensor placement for reliable virtual sensing using modal expansion and information theory. Sensors. https://doi.org/10.3390/s21103400

31. Gomes GF, de Almeida FA, da Silva Lopes Alexandrino P et al (2019) A multiobjective sensor placement optimization for SHM systems considering Fisher information matrix and mode shape interpolation. Eng Comput 35:519–535. https://doi.org/10.1007/s00366-018-0613-7

32. Yuen K, Kuok S (2015) Efficient Bayesian sensor placement algorithm for structural identification: a general approach for multi-type sensory systems. Earthquake Eng Struct Dynam 44:757–774

33. Goulet J-A, Smith IFC (2012) Performance-driven measurement system design for structural identification. J Comput Civ Eng 27:427–436

34. Kammer DC (1991) Sensor placement for on-orbit modal identification and correlation of large space structures. J Guid Control Dyn 14:251–259

35. Sela L, Amin S (2018) Robust sensor placement for pipeline monitoring: mixed integer and greedy optimization. Adv Eng Inform 36:55–63

36. Papadimitriou C (2004) Optimal sensor placement methodology for parametric identification of structural systems. J Sound Vib 278:923–947

37. Argyris C, Papadimitriou C, Panetsos P (2017) Bayesian optimal sensor placement for modal identification of civil infrastructures. J Smart Cities. https://doi.org/10.18063/JSC.2016.02.001

38. Papadimitriou C, Lombaert G (2012) The effect of prediction error correlation on optimal sensor placement in structural dynamics. Mech Syst Signal Process 28:105–127

39. Papadopoulou M, Raphael B, Smith IFC, Sekhar C (2014) Hierarchical sensor placement using joint entropy and the effect of modeling error. Entropy 16:5078–5101

40. Bertola NJ, Papadopoulou M, Vernay D, Smith IFC (2017) Optimal multi-type sensor placement for structural identification by static-load testing. Sensors 17:2904

41. Bertola NJ, Pai SGS, Smith IFC (2021) A methodology to design measurement systems when multiple model classes are plausible. J Civ Struct Heal Monit. https://doi.org/10.1007/s13349-020-00454-9

42. Lin J-F, Xu Y-L, Law S-S (2018) Structural damage detection-oriented multi-type sensor placement with multi-objective optimization. J Sound Vib 422:568–589

43. Papadopoulou M, Raphael B, Smith IFC, Sekhar C (2015) Optimal sensor placement for time-dependent systems: application to wind studies around buildings. J Comput Civ Eng 30:04015024

44. Civera M, Surace C (2022) An application of instantaneous spectral entropy for the condition monitoring of wind turbines. Appl Sci 12:1059. https://doi.org/10.3390/app12031059

45. Ceravolo R, Civera M, Lenticchia E et al (2021) Detection and localization of multiple damages through entropy in information theory. Appl Sci 11:5773. https://doi.org/10.3390/app11135773

46. Bertola NJ, Smith IFC (2019) A methodology for measurement-system design combining information from static and dynamic excitations for bridge load testing. J Sound Vib 463:114953

47. Bertola NJ, Cinelli M, Casset S et al (2019) A multi-criteria decision framework to support measurement-system design for bridge load testing. Adv Eng Inform 39:186–202. https://doi.org/10.1016/j.aei.2019.01.004

48. Bertola NJ, Costa A, Smith IFC (2020) Strategy to validate sensor-placement methodologies in the context of sparse measurement in complex urban systems. IEEE Sensors J. https://doi.org/10.1109/JSEN.2020.2969470

49. Wang ZZ, Bertola NJ, Goh SH, Smith IFC (2021) Systematic selection of field response measurements for excavation back analysis. Adv Eng Inf 48:101296. https://doi.org/10.1016/j.aei.2021.101296

50. Cao W-J, Zhang S, Bertola NJ et al (2019) Time series data interpretation for 'wheel-flat' identification including uncertainties. Struct Health Monit. https://doi.org/10.1177/1475921719887117

51. Robert-Nicoud Y, Raphael B, Smith IFC (2005) System identification through model composition and stochastic search. J Comput Civ Eng 19:239–247

52. Šidák Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 62:626–633

53. Pasquier R, Smith IFC (2016) Iterative structural identification framework for evaluation of existing structures. Eng Struct 106:179–194

54. Chen X, Cui T, Fu J et al (2016) Trend-residual dual modeling for detection of outliers in low-cost GPS trajectories. Sensors 16:2036

55. Smarsly K, Law KH (2014) Decentralized fault detection and isolation in wireless structural health monitoring systems using analytical redundancy. Adv Eng Softw 73:1–10

56. Proverbio M, Bertola NJ, Smith IFC (2018) Outlier-detection methodology for structural identification using sparse static measurements. Sensors 18:1702. https://doi.org/10.3390/s18061702

57. Pai SGS, Sanayei M, Smith IFC (2021) Model-class selection using clustering and classification for structural identification and prediction. J Comput Civ Eng 35:04020051. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000932

58. Zhang WG, Zhang RH, Han L, Goh ATC (2019) Engineering properties of the Bukit Timah granitic residual soil in Singapore. Underground Space 4:98–108. https://doi.org/10.1016/j.undsp.2018.07.001

59. Finno RJ, Calvello M (2005) Supported excavations: observational method and inverse modeling. J Geotech Geoenviron Eng 131:826–836

60. Wang Z-Z, Goh SH, Koh CG, Smith IFC (2019) An efficient inverse analysis procedure for braced excavations considering three-dimensional effects. Comput Geotech 107:150–162. https://doi.org/10.1016/j.compgeo.2018.12.004

61. Benz T, Vermeer PA, Schwab R (2009) A small-strain overlay model. Int J Numer Anal Meth Geomech 33:25–44. https://doi.org/10.1002/nag.701

62. Cundall P, C Carranza-Torres, and R Hart (2003). A new constitutive model based on the Hoek-Brown failure criterion. FLAC and Numerical Modeling in Geomechanics. In: Proceedings of the Third International FLAC Symposium, Published by: A.A. Balkema, a member of Swets & Zeitlinger. Sudbury, Canada.

63. Calvello M, Finno RJ (2004) Selecting parameters to optimize in model calibration by inverse analysis. Comput Geotech 31:410–424. https://doi.org/10.1016/j.compgeo.2004.03.004

64. Cover TM, Thomas JA (2006) Element of information theory, 2nd edn. John Wiley & Sons, Hoboken

65. Krause A, Singh A, Guestrin C (2008) Near-optimal sensor placements in Gaussian processes: theory, efficient algorithms and empirical studies. J Mach Learn Res 9:235–284