Using entropy to maximize the usefulness of data collection

Y. Robert-Nicoud, B. Raphael and I.F.C. Smith

EPFL-Swiss Federal Institute of Technology

IMAC, ENAC, CH-1015 Lausanne, Switzerland

Yvan.Robert-Nicoud@epfl.ch, Benny.Raphael@epfl.ch, Ian.Smith@epfl.ch

Summary

This paper presents a generic methodology for measurement system configuration when the goal is to identify behaviour models that reasonably explain observations. For such tasks, the best measurement system provides maximum separation between candidate models. In this work, the degree of separation between models is measured using Shannon's Entropy Function. The location and type of measurement devices are chosen such that the entropy of candidate models is greatest. This methodology is tested on a laboratory structure and, to demonstrate generality, an existing fresh water supply network in a city in Switzerland. In both cases, the methodology suggests an appropriate set of sensors for identifying the state of the system.

1 Introduction

The configuration of sensors within diagnostic and control systems remains a task that engineers perform often without systematic and scientific evaluations. Engineers usually make ad-hoc decisions related to location and types of sensors to be used. A more rational and generic methodology for measurement system configuration does not currently exist. De Kleer and Williams (1987) describe an approach to identifying measurements required for performing diagnosis. They use entropy as a measure of probabilities of candidate models in order to identify measurements to be taken. However, their approach is part of a diagnosis methodology and requires measurements from previous sensors in order to locate the next sensor. In civil engineering, installation of sensors and taking measurements are time-consuming tasks. These tasks are not usually performed during diagnostic evaluations; the measurement system has to be configured a-priori such that sensors provide useful data in a wide range of situations.

Isolated proposals for a-priori configuration of sensors are found in areas such as robotics and computer vision. (Cowan et al., 1990, Sakane et al., 1987, Tarabanis and Tsai, 1991, Sedas-Gersey, 1993). Mason and Grun, (1995) have developed an expert system called CONSENS for sensor placement to be used in inspection tasks. In this system, expert photogrammetric knowledge is used to place vision sensors such that the object to be inspected is within the field of view of sensors, dimensions can be computed precisely, and other sets of inspection constraints are satisfied. This approach cannot be applied to other tasks.

Sensor placement strategies in structural engineering are important for system identification tasks. System identification involves determining the state of a system and its parameter values through comparisons of predicted and observed responses (Ljung, 1999, Friswell and Mottershead, 1995). Sensor placement strategies in structural engineering are currently limited to updating vibration models through dynamic measurements. Udwadia (1994) has proposed an approach for determining optimal sensor locations through maximizing an appropriate norm of the Fisher information matrix. Heredia-Zanoni and Esteva (1998) explicitly model uncertainties in structural parameters and seismic ground motion activation through the use of probability density functions. Optimal sensor locations are then determined through minimizing an expected loss function, which has been derived for the case of linear stochastic structural response. Papadimitriou et al (2000) use entropy as a measure of uncertainty in model parameters. The optimal configuration of sensors is then chosen to be the one that minimizes the entropy function. These approaches cannot be easily applied to the case of static measurements when different classes of models that have varying numbers of parameters need to be considered. Furthermore, closed-form mathematical expressions for computing model responses are difficult to obtain.

Measurement system configuration is a discrete combinatorial optimization problem. The number of combinations of sensor types and locations increase exponentially with the number of sensors in the measurement system. Potentially good sensor locations are positions of high entropy in relation to predictions of different candidate models. It is easier to evaluate the entropy of the distribution of model predictions compared to the entropy of probabilities of model parameters. This definition of entropy has not been used in the configuration of measurement systems in previous work.

In this paper, a methodology for measurement system configuration that focuses on the maximization of entropy is proposed. The organisation of the paper is as follows: Section 2 contains a description of the methodology. Results of applying the methodology to a laboratory structure and a water network are presented in Section 3. Comparison with other work and limitations of the methodology are discussed in Section 4, and Section 5 contains the conclusions.

2 Measurement system configuration

The objective of measurement system configuration in this study is to improve the reliability of system identification. Reliability of system identification is poor when many candidate models predict similar values at sensor locations. Therefore, locations and types of measurement devices are chosen such that there is maximum separation between predictions of candidate models. In this work, the degree of separation between models is measured using the entropy function (Shannon and Weaver, 1949). This concept has been developed in the field of information theory and is a measure of “disorder” within a set. There is maximum disorder when predicted values show wide dispersion. An ideal measurement system is the one that results in maximum variation in predictions made by different candidate models at measurement locations. Therefore, the location and type of measurement devices are chosen such that the entropy of the set of model predictions is the maximum.

2.1 Entropy

The entropy function as defined by Shannon and Weaver (1949) is

$$H = -\sum_{i=1}^{m} p_i \cdot \log_2(p_i) \qquad (1)$$

where p_i is the probability of the *i*-th interval of a distribution and *m* is the number of intervals. In the case of a variable with two values (*m*=2) having probabilities p and (1-p) the expression is

$$\mathrm{H} = -\left(\mathrm{p}\log_2 \mathrm{p} + (1-\mathrm{p})\log_2(1-\mathrm{p})\right) \qquad (2)$$

The maximum of this function is H=1.0 at p=0.5, that is when the probability is equally distributed between the two values.

In the case of a variable that has *m* discrete values, the entropy is a maximum when all values have the same probability 1/*m.* Thus entropy is a measure of homogeneity in a distribution. A completely homogenous distribution has maximum entropy.

Equation (1) is used to evaluate the entropy of a distribution of models. For example, suppose that the distribution of model responses at a sensor location is represented by a histogram consisting of 5 intervals. The probability of an interval is defined as the ratio of the number of models lying in the interval (*Ni*) and the total number of models (*Ntot*). The entropy value of the sensor is therefore,

$$H = -\left(\frac{N_1}{N_{tot}} \log_2 \left(\frac{N_1}{N_{tot}} \right) + \frac{N_2}{N_{tot}} \log_2 \left(\frac{N_2}{N_{tot}} \right) + \frac{N_3}{N_{tot}} \log_2 \left(\frac{N_3}{N_{tot}} \right) + \frac{N_4}{N_{tot}} \log_2 \left(\frac{N_4}{N_{tot}} \right) + \frac{N_5}{N_{tot}} \log_2 \left(\frac{N_5}{N_{tot}} \right) \right) - (3)$$

2.2 A methodology for measurement system configuration

The number, the type and the location of sensors determine the capacity of a measurement system to discriminate between candidate models. The capacity of a particular system is evaluated using the entropy function. The best system is the one in which the total entropy is the maximum. This can be formulated as a discrete optimisation problem. Suppose there are q possible sensor locations, then there are $N=2^q-1$ possible combinations of sensor placements. Sensors at each location could be of different types. There are q boolean variables in the optimisation formulation. Each variable indicates whether the sensor is present at a specified location. A solution consists of a set of values for all the variables and can be evaluated using the entropy function. Since this discrete combinatorial optimisation problem is difficult to solve, a "greedy" algorithm has been developed and is summarised below. A "greedy" algorithm is characterized by the selection of the best immediate alternative for each incremental step. Selections that consist of accepting a less attractive alternative for a better overall final solution are not allowed in this algorithm.

The user inputs the set of potential sensor locations and a range of possible hypotheses. A population (sample) of models is generated randomly (or deterministically) using assumptions that are specified by the user. These models are analysed using the finite element method and predicted responses at all possible sensor locations are computed. Entropy is calculated using the distribution of predicted responses. The sensor that results in maximum entropy is chosen. Sets of models that cannot be separated using data from this sensor are used for the identification of subsequent sensors and the process is repeated. This process results in a list of combination of sensors that are ordered by increasing entropy. Precision of sensors is also taken into account in the methodology and therefore, sensors having different precisions (costs) may be considered. The methodology is described in more detail in the following.

2.2.1 Definition of the frequency histogram

The entropy corresponding to a sensor location is calculated using the distribution of values predicted by a sample of models at the sensor location. Since the behaviour of a structure is difficult to predict a-priori, a population of models is randomly generated using a set of assumptions specified by the user. (Models could also be generated deterministically by considering all combinations of assumptions when the set of possible assumptions is small). These models are analysed using the finite element method and the predicted responses at all possible sensor locations are calculated. A histogram is created for representing the distribution of model responses at each sensor location. The characteristics of the histogram are represented by the following variables

X_{1j}	The minimum value predicted by the sample of models at the sensor location j
X_{2j}	The maximum value predicted by the sample of models at the sensor location j
Dj	Maximum deviation in values at the sensor location j, X_{2j} - X_{1j}
Pj	Precision of the measurement of the sensor at location j
Iwj	The width of each interval in the histogram. Iwj >= Pj

b The acceptable maximum number of models within an interval. If an interval contains more than b models, the group is considered to be non-identifiable.

αj Dj/Pj, coefficient used to compare the efficiency of the sensor location j

nb Maximum number of intervals

Ej Extended width of the histogram

2.2.2 Construction of the frequency histogram

The frequency histogram is constructed for each sensor location independently. In order to compare the histograms having different ranges of values for two different sensor locations, the following procedure is used:

1. Calculate the maximum deviation in Dj values at each sensor location.
2. Calculate α_j = Dj/Pj for each sensor location and determine the maximum value $\alpha_{j\ max}$. (that is, at the sensor location where there is maximum deviation in comparison with the precision).
3. Calculate the interval width for each sensor location Iwj using the relationship

   $Iwj = a \times Pj$

   Where $a = \alpha_{j\ max}/nb$ with the condition that a=1 if a < 1. This relationship ensures that the interval width is always greater than or equal to the precision of the sensor. The parameter a is used to keep the same ratio of interval width to precision for all sensors.
4. Calculate the extended width of the histogram using $Ej = Iwj \times nb$. The extended width ensures that the number of intervals is the same for all the histograms.
5. Calculated the entropy for each histogram

For comparing two sensors of different type, the entropy calculated by this procedure is weighted by the ratio of the extended width to the precision of the sensor.

2.2.3 A procedure that focuses on maximum entropy

A greedy algorithm is used to maximise entropy where, at each stage, the sensor location that corresponds to maximum entropy is selected. Even though this procedure may miss the optimal sensor configuration, it avoids the exponential complexity associated with the evaluation of all possible combination of sensor locations. The algorithm is shown in pseudo-code below.

```
User specifies b, the acceptable maximum number of models in an interval that cannot be separated. The number of intervals, nb, and the maximum number of iterations, numIterations, are also specified by the user

A list of sets of models to be separated, modelList, is created. The set of randomly generated models is added to the list.

Initialise the set of chosen sensors, sensorList, to null

Initialise iteration counter, count, to zero

Repeat while modelList is not empty and count is less than numIterations {

    Select the first set from modelList, let it be currentSet. CurrentSet is removed from modelList.

    Create the histogram for each sensor using the models in currentSet. Calculate the entropy for each histogram
```

```
        Select the histogram with maximum entropy.  Add the corresponding sensor to
        *sensorList*.  The sets of models corresponding to intervals containing more than *b*
        models are added to *modelList*.

        Re-order *modelList* such that the set containing maximum number of models is at
        the top of the list.

        Increment count

    }

    Report the group of models that cannot be separated further.
```

At the end of the process, it is possible to evaluate the performance of the measurement system using the number of models that cannot be separated further.

3 Evaluation and results

The methodology for system identification and measurement system configuration has been tested on a laboratory structure and a water network. These are summarised in the discussion below.

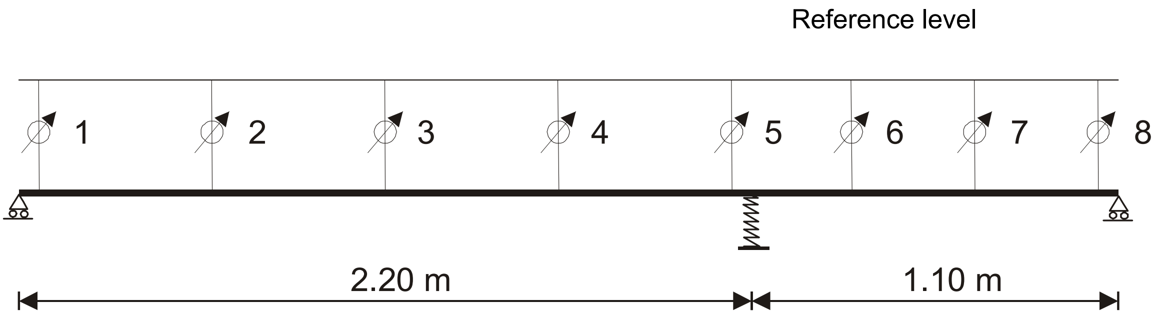


Figure 1 Position of sensors on timber beam supported on springs

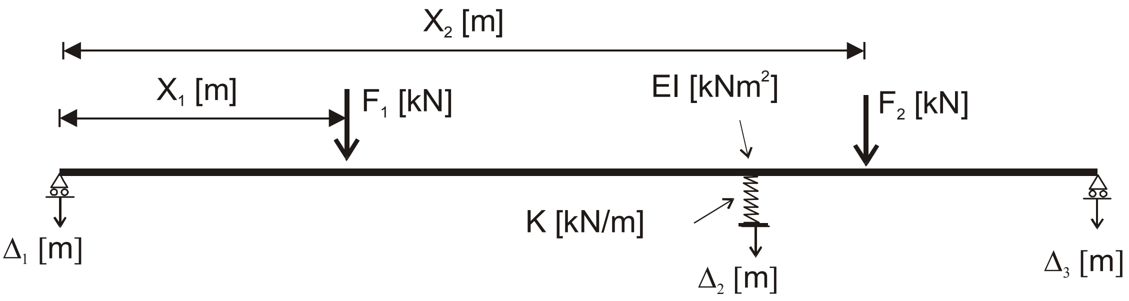


Figure 2 System identification variables

3.1 Timber beam supported on springs

A timber beam supported on springs was constructed in the laboratory. Eight inductive sensors were used to measure vertical displacements at different locations. These were uniformly distributed over the length of each span of the beam (Figure 1). Positions and magnitudes of applied loads along with the characteristics of the structure such as the material properties and support conditions were treated as unknown variables (Figure 2).

A sample consisting of 1455 models was used to evaluate the potential of the measurement system for the identification of good models that correspond to reality. The order of sensors suggested by the measurement system configuration methodology is given in Table 2.

	Sensors		
Rank	Number	Position [m]	Entropy
1	4	1.62	8.82
2	3	1.1	8.69
3	5	2.14	8.52
4	2	0.58	8.20
5	6	2.5	8.10
6	7	2.87	7.74
7	1	0.06	6.91
8	8	3.24	6.78

Table 1. Entropy of sensors

The sensor closest to the midspan, 4, has the maximum potential to separate candidate models. This is followed by the other two sensors near the midspan, 3 and 5. When all eight sensors were considered in the measurement system, the order of addition of sensors suggested by the methodology is given in Table 2.

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

Table 2. The order of sensors for maximum entropy

The number of models that cannot be identified for each situation is shown in Figure 3. Two cases were studied, one in which the minimum width of the interval was chosen as 0.01 and the second in which the minimum width is chosen as 0.2 (low precision sensors). In both cases the separability of models does not improve significantly after 3 sensors.

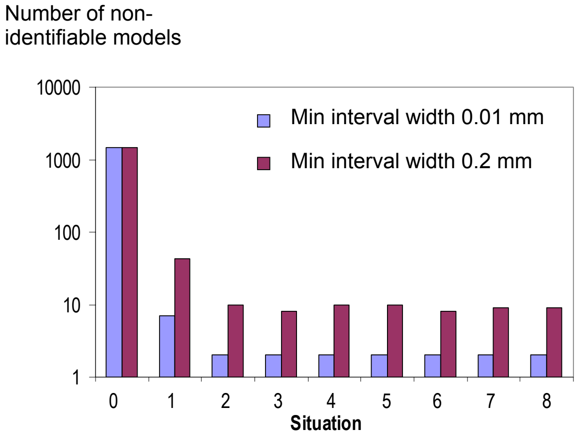


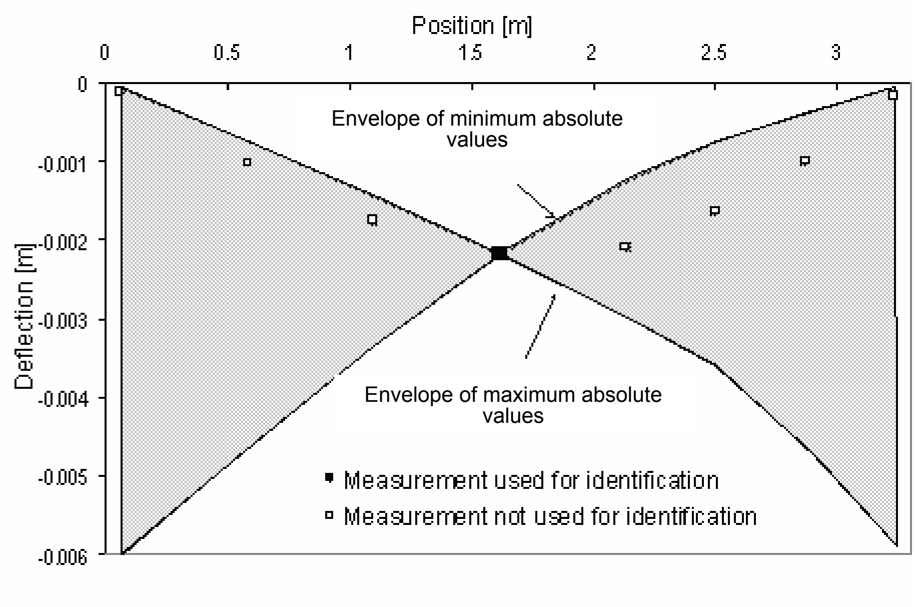
Figure 3 The number of models that cannot be identified for each situation

Loads were applied on the structure and measurements were taken from all the sensors in order to test the validity of the model identification capability that is predicted by the measurement system configuration methodology. Initially measurement from a single sensor (sensor 4) was used to identify models. The set of models that predicted responses close to the measured value included those that corresponded to reality as well as those that involved wrong support conditions and loading. The envelope of the deflected curves of all the candidate models are shown in Figure 4a. All models match the deflection at the location of sensor 4, but differ significantly at sensor locations that were not used in model identification. With three sensors, the candidate models reasonably matched measurements at all sensor locations including those that were not used in system identification (Figure 4b). The results in Figure 4b are not much different if eight sensors are used.

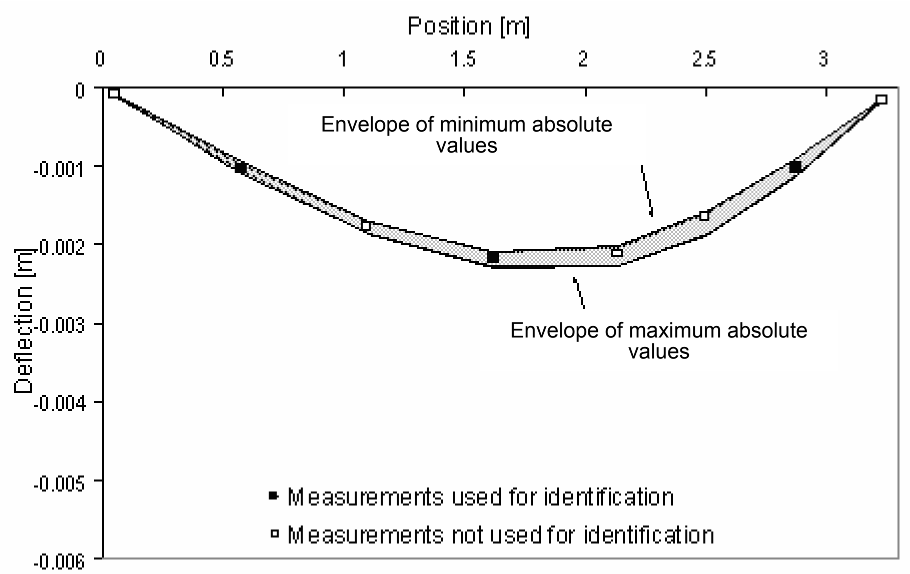
In conclusion, the measurement system configuration module correctly determines the sensors that are required for good system identification. The methodology is able to correctly identify the state of the system provided that a minimal number of measurements are available.

3.2 Leak detection in a water network

The generality of this approach is now studied through an application outside of structural engineering. Leaks in networks typically result in losses of about 20-30% of water in fresh water distribution systems of towns and cities. Detection of leaks is difficult since pipes are buried under ground. An approach to leak detection involves measurement of noise created by leaks (Hunaidi 2000). However, a rational methodology for determining the location of installation of measurement devices does not exist. The objective of the present study is to illustrate the generality of the methodology, by applying it to the configuration of measurement systems for leak detection in water networks. Since rigorous mathematical models of noise propagation do not exist, a simplified model is used in which the intensity of noise is assumed to drop inversely proportional to the square of the distance from the source. Thus the predicted level of noise at a sensor location is calculated by computing the minimum distance between a possible location of leak and the location of the sensor. An algorithm for computing the shortest path between two points in a graph is used.



a) When a single sensor was used to identify models.



b) When three sensors were used to identify models

Figure 4 Envelope of predicted responses of candidate models.

A water network is represented as a graph consisting of nodes and elements. An element has attributes, fore-node, back-node and length. A leak could exist on any element and is characterised by the element number and the distance from the fore-node. A sensor might be installed at any point along the length of an element. A model consists of one or more leaks in the network. The locations of leaks and their intensities are unknown variables to be determined through system identification.

The network of the town Martigny in Switzerland was used to test the methodology. A sample consisting of 1000 models was employed for the measurement system configuration. Eighty four potential sensor locations were chosen. Their ordering according to decreasing entropy was found to be as given in Table 3.

Rank	1	2	3	4	5	6	7	8	9	10	…	84
Sensor	0	5	6	15	67	1	23	4	66	7	…	61
Entropy	3.64	3.64	3.57	3.57	3.57	3.2	3.2	3.06	3.06	3.01	…	0.52

Table 3. Ordering of sensors according to entropy

Sensor 0 has the highest entropy, followed by Sensors 5, 6 and finally Sensor 61. According to the methodology, the highest ranking sensor, 0, is initially chosen. The number of models that cannot be separated using measurement from this sensor is 789. These models were then used for the selection of the second sensor. Sensor 25 had the highest entropy using the sample of 789 non-separable models and therefore this was chosen as the second sensor. The number of models that cannot be separated using the combination of sensors 5 and 25 is 685. The process is repeated using these 685 models. The combination of first 20 sensors that are identified using this procedure is given in Table 4. (S is the sensor that is added at the N-th stage)

N	1	2	3	4	5	6	..	14	15	16	17	18	19	20
S	0	25	77	10	69	54	..	81	64	31	36	68	3	42
M	789	685	584	504	454	413	..	217	203	190	177	165	153	142

Table 4. The number of non-identifiable models (M) vs the number of sensors (N).

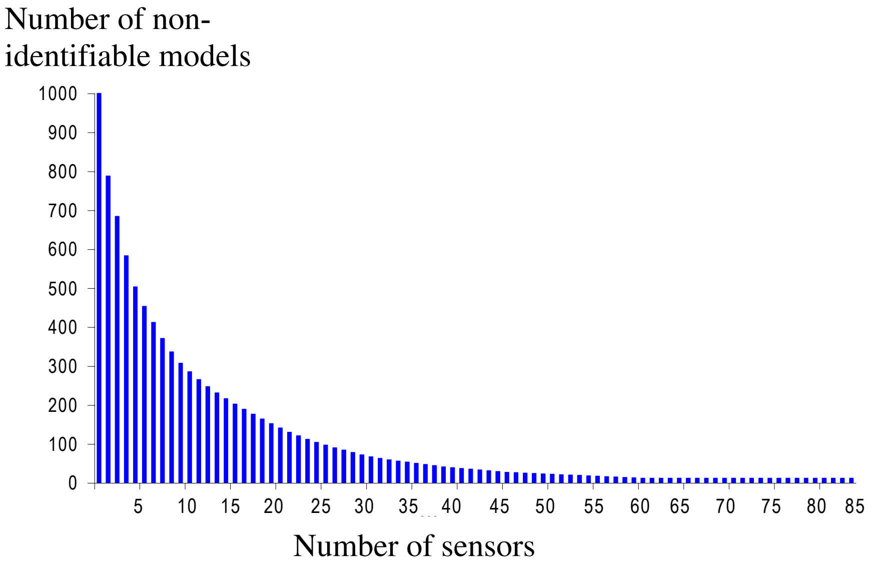


Figure 5 **The number of models that cannot be separated decreases asymptotically with the number of sensors.**

The number of models that cannot be separated using a combination of the best *k* sensors (k=1 to 84) is shown in Figure 5. The number of models that cannot be separated decreases asymptotically with the number of sensors.

In order to test the capability of the measurement system that has been configured by the methodology, two leaks were assumed to be present in elements 7 and 43, at distances 110.0 and 49.0 m. from the respective fore-nodes. The intensity of noise at the location of the leak was assumed to be 60 dB. In order to simulate measurements, the intensities of noise at the locations of sensors were computed using these parameters assuming that the intensity of noise drops inversely proportional to the square of the distance from the source. Two measurement systems were tested for their capacity to identify the state of the network: A – consisting of the best 9 sensors, B – consisting of the best 20 sensors. System identification was performed by minimising the deviation of predicted responses from measured values. The locations of leaks that were identified using the first measurement system are on elements 7 and 37, whereas the actual leaks are on elements 7 and 43 (Figure 6). On the other hand, the correct locations were identified upon increasing the number of sensors to 20 using measurement system B.

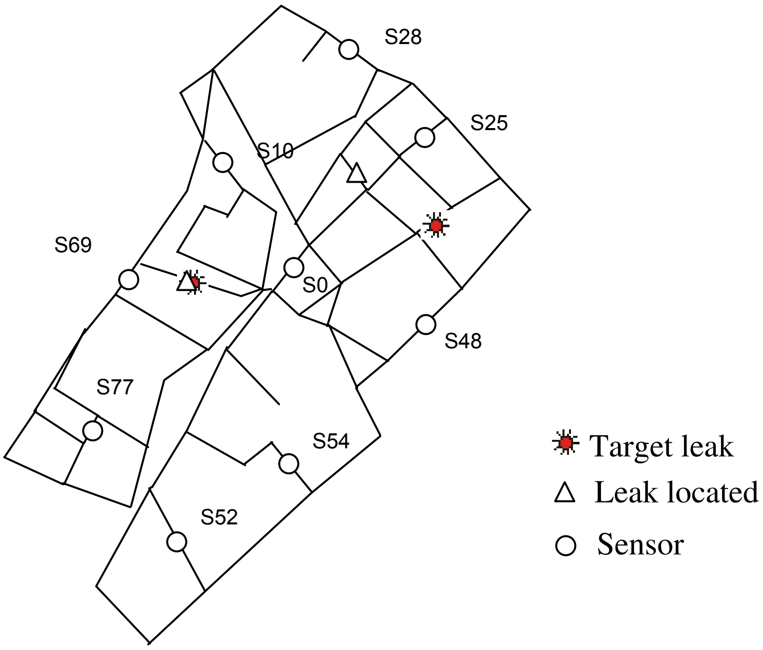


Figure 6 Leaks that have been identified using Measurement System A

4 Discussion

The methodology for measurement system configuration has been tested on a number of other case studies that are not included in this paper (Robert-Nicoud 2003). In all cases, correct system identification has been found possible with a limited number of sensors.

The present work differs from that of Papadimitriou et al (2000), who use entropy as a measure of uncertainty in model parameters – therefore, the optimal configuration of sensors corresponds to the one having minimum entropy. In the present work, entropy is used to measure the dispersion in the values of predicted responses at a sensor location. Therefore, the best configuration corresponds to maximum entropy since this configuration results in maximum separation between models.

Limitations of this procedure include the following

- The sample size of models is defined a-priori and therefore, all possible models are not tested for identifiability.
- Users define the range of hypotheses that determine the initial sample of models
- The set of possible sensor locations are specified by users. This requires a knowledge of sensors and their potential locations
- The optimal number of sensors is not provided directly by the algorithm. Incremental improvements provided by addition of sensors might, for example, be insignificant beyond a certain number of sensors.
- The entropy calculation depends, albeit weakly, on the total number of intervals of the histogram.

These limitations are not problematic since the methodology is meant for decision support and not for autonomous configuration of measurement systems. Engineers use their knowledge and experience to specify valid modelling assumptions and potential sensor locations. When assumptions are suspected to influence results, users can easily observe such effects through multiple execution cycles.

5 Conclusions

The following points summarise the conclusions of this study.

- Entropy is a useful concept for evaluating the information content at sensor locations and has the potential to provide systematic and rational support for the configuration of measurement systems.
- The state of the system can be identified correctly provided that a minimal number of measurements are available.
- The measurement system configuration methodology helps determine an appropriate set of sensors for good system identification.

The methodology is already proving to be a valuable tool for engineers who are involved in the task of monitoring and maintenance of engineering systems. The amount of data that is collected is limited to the most useful for the task, thus eliminating expenses related to unnecessary data collection and interpretation.

ACKNOWLEDGEMENT

This research is funded by the Swiss National Science Foundation (NSF) and the Commission for Technology and Innovation (CTI).

6 References

Cowan C., Model base synthesis of sensor location, International conference on Robotics and Automation, April 24-29, 1988, pp. 900-905.

De Kleer J. and Williams B. C., Diagnosing Multiple Faults," Artificial Intelligence, Vol 32, Number 1, 1987.

Friswell M.I. and Mottershead J.E., Finite Element Model Updating in Structural Dynamics, Kluwer, 1995.

Heredia-Zavoni E. and Esteva L., Optimal instrumentation of uncertain structural systems subject to earthquake ground motions, *Earthquake engineering and structural dynamics*, 27, pp. 343-362, 1998.

Papadimitriou, J.L. Beck and S.K. Au, Entropy-Based Optimal Sensor Location for Structural Model Updating, C. Journal of Vibration and Control, 6, 781-800, 2000.

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

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

Sakane S., Sato T., Kakikura M., Automatic planning of light source placement for an active photometric stereo system, IEEE workshop on Intelligent robots and systems, IROS 90, 1990, pp. 559-566.

Scott O.M. and Grun A., Automatic sensor placement for accurate dimensional inspection, Computer vision and image understanding, Vo. 61, 3, 1995, pp. 454-467.

Sedas-Gersey S.W., Algorithms for automatic sensor placement to acquire complete and accurate information, PhD thesis, Robotics Instiute, CMU, Pittsburgh, 1993.

Shannon C., Weaver W., The Mathematical Theory of Communication, University of Illinois Press, 1949.

Tarabanis K., Tsai R.Y., Allen P.K., Automatic sensor planning for robotic vision tasks, Proceedings of the 1991 IEEE conference on Robotics and Automation, 1991, pp. 76-82.

Udwadia F.E., Methodology for optimum sensor locations for parameter identification in dynamic systems, *Journal of engineering mechanics*, ASCE, 120, p. 368-390, 1994.