Optimal Pressure Measurement Layout Design in Water Distribution Network Systems

Kálmán Klapcsik[1*], Roxána Varga[1], Csaba Hős[1]

Abstract

This paper addresses the problem of locating the optimal pressure measurement points in a hydraulic system to help system management, calibration/validation of hydraulic models and measurement planning. Two approaches are discussed in the present work. The first method splits the hydraulic system by means of community concept borrowed from graph theory and uses merely the topology of the network. The resulting subsystems will have minimum number of external and maximum number of internal connections and leaves the choice of locating the single pressure measurement location per subsystem to a higher-level decision. The second technique is based on the sensitivity analysis of the hydraulic network and places the measurement points at the most sensitive locations, while trying to preserve the spatial diversity of the layout, i.e. preventing the accumulation of the measurement points within a small area of high sensitivity. The performance of both techniques is demonstrated on real-size hydraulic networks. The proposed sampling layouts are compared to classic D-optimality, A-optimality and V-optimality criterion.

[1] Department of Hydrodynamic Systems,
Faculty of Mechanical Engineering,
Budapest University of Technology and Economics
H-1521 Budapest, P.O.B. 91, Hungary

*Corresponding author, e-mail: kklapcsik@hds.bme.hu

1 Introduction

Pressure data collection in water distribution systems (WDS) is essential for proper management of the system. The knowledge of pressure values generally drives the operational actions, such as leakage, infiltration and demand control [1]. An urban water distribution system usually contains hundreds or thousands of nodes and pipes, but the limited resources do not allow the deployment of pressure loggers at every potential measurement location (typically fire hydrants). Hence, one has to decide where to mount the finite number of pressure sensors, out of the vast number of potential locations. The measurement layout should be optimal in the sense that it gives the "most possible" information about the actual state of the network, however, defining the optimality condition (i.e. the objective to maximise) is non-trivial.

Numerous researchers addressed the problem of optimal sampling layout design. One important branch of the researches focuses on capturing the presence of contamination (e.g. due to terrorist attack) as early as possible, see [2]. Another approach aims to use the measurement for the calibration of a hydraulic simulation model, notably pipe roughness values and/or demand patterns. For such purposes, Walski [3] suggested to monitor the pressure at nodes with high base demand and on the perimeter of the skeletonised network. Lee and Deininger [4] developed a unique coverage-based approach to select sampling points to monitor water quality. Yu and Powell [5] describe the problem as a multi-objective optimization, and used dynamic covariance matrix analysis to locate good sampling points. Bush and Uber [6] derived three methods based on the analysis of the Jacobian matrix from the D-optimally criterion. De Schaetzen [7] proposed three new algorithms, two out of these were based on the shortest path concept while the third approach was based on the maximization of Shannon's entropy. Kapelan et al. [8] developed three methods based on D- and A-optimality criterion and on the model prediction uncertainty, which was previously used by Lansey et al. [9]. Behzadian et al. [10] used multi-objective genetic algorithm and adaptive neural network to design a good sampling layout.
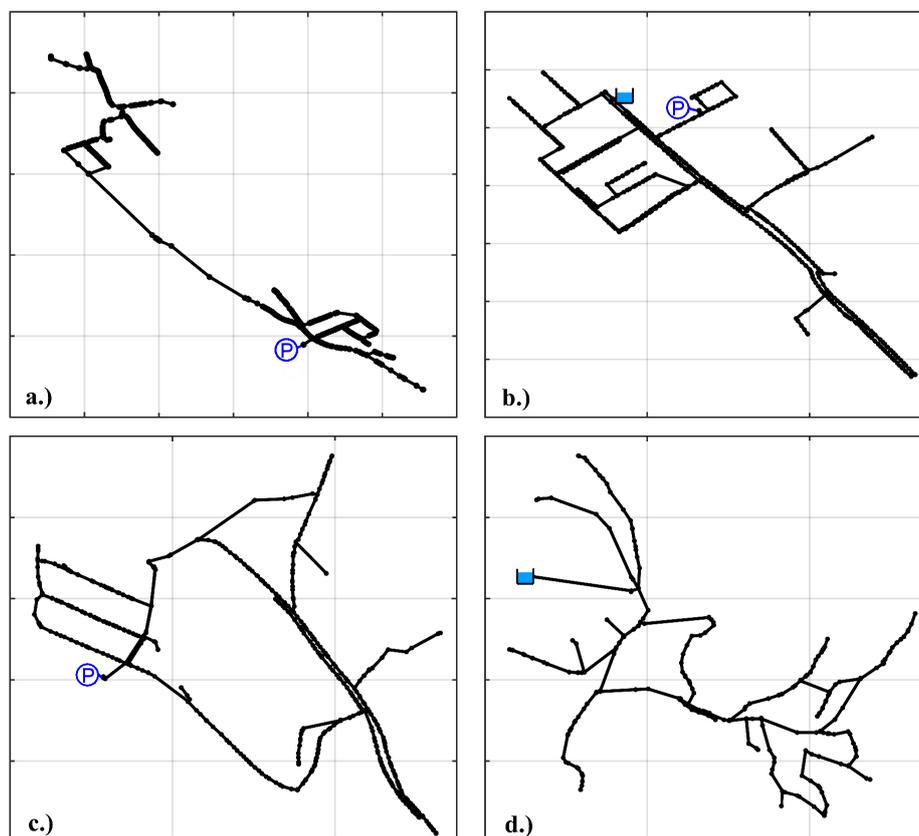
Fig. 1 The test networks, zones of Sopron Waterworks, Hungary. The blue circles with letter P denote the feeding pump stations (modelled with constant pressure) and the pool symbols are the reservoirs.

This paper utilizes graph theory to split the WDS into sub-graphs (communities) that are isolated in the sense that they have maximum number of inner connections but minimal number of inter-community connections. After the split, the most sensitive node of each community is chosen for pressure logging. We compare these results against the layouts obtained by the objectives developed in [7] using sensitivity and Shannon's entropy, and the classic FOSM [6] models, respectively. We report on the performance of these techniques in the case of four real WDSs with moderate size (few hundreds of nodes).

One of the strengths of the proposed approach is that it provides measurement districts and leaves the final decision to be made at higher level. Further, in contrast to the classical FOSM based approaches, instead of computing sensitivities for grouped roughness parameters; all pipes were treated separately. When using pipe cluster with grouped parameters, the number of pipes in the groups may affect the final sampling design. Since, parameter groups with plenty of pipes may be overrepresented against parameter groups containing only a few pipes. In addition, by means of our in-house software "STACI", we are able to obtain the elements of the sensitivity matrix analytically as described in [11]. The drawback of our approach is the slow computation of communities. The partitioning is obtained by maximizing a measure called modularity that is known to be a combinatorial NP problem [12], thus the computational effort escalates with the network size. However, we implemented a genetic algorithm that can accelerate the partitioning, and makes the community finding feasible for these medium-sized networks.

The investigated real-life systems belong to the WDS of Sopron, a town Hungary and are typically providing potable water for a few thousands of citizens. The networks are depicted in Fig. 1, where the blue circles with letter P and the pool symbols denote the feeding pump stations (modelled with constant pressure) and the reservoirs, respectively. Networks a, b, c, and d, contain 488, 700, 549, and 283 nodes, and 461, 634, 505 and 257 pipes, respectively. The topology and the pipeline data (length, diameter, and coordinates) were exported directly from the GIS system of the Sopron Waterworks. The demand data represents an average monthly value and was extracted from the Waterworks' billing database.

The rest of the paper is organized as follows. We start with presenting the mathematical toolbox: community structure of graphs, the hydraulic simulation software, the sensitivity analysis, Shannon's entropy, evolutionary algorithms, and FOSM models. In the next section, we present the proposed approaches in case of a hypothetical test WDS. Then we investigate the performance of these techniques in case of real-life medium-sized networks. Finally, the test results are reported and we summarize the study.

2 Mathematical tools

2.1 Dividing the network into disjoint networks

We start off with presenting a technique that is capable of splitting the WDS into subsystems with minimum number of external connections (between the subsystems) and maximum internal connectivity (within one subsystem). Note that this technique uses the *topology only*, (i.e. no information on pipe length, diameter, flow rate, pressure, etc. is used). Thus the results will be independent of the hydraulic state of the system, which might be useful in the case of WDSs with highly varying hydraulic parameters (e.g. during morning/evening peaks and nightly low-load periods) but is clearly disadvantageous for WDSs where the pressures and flow directions are relatively constant over a long time interval. Using purely the topology, we lose pieces of engineering information, since e.g. the main supply pipes with high base demand and costumer pipes with low base demand are considered with equal relevance. It is worth mentioning that using weight-based modularity allows us to consider specific network characteristics [13] and design communities for specific purposes.

Let us now describe the algorithm. The technique is based on the slightly modified Newman-algorithm [14-17] from the theory of large networks. The main idea is to define strongly connected disjoint regions in the network which are weakly connected. This is achieved by maximizing a measure, called the modularity, which is denoted by Q; larger Q values represent better partitioning. Maximizing Q is usually achieved by 'fine-tuning' algorithms (see [18]) however, we found that even the fine tuning in itself did not give satisfactory result. Hence, after the fine-tuning stage we apply a genetic algorithm to reach proper partitioning.

To formulate Q, the definition of the graph representing the hydraulic system is needed. For this, the nodes of the graph are the network nodes (the possible measurement points) of the system, and two nodes in the graph are connected only when they are connected in the WDS via a pipe (or by another edge element, e.g. valve, pump, etc.). For the mathematical formulations and the numerical computations the adjacency matrix $\mathbf{A}$ of the graph and the degree k_i of the ith node are needed. An $A(i, j)$ element of $\mathbf{A}$ is 1 if nodes i and j are connected in the graph and 0 otherwise. Since the graph is undirected, A_i is symmetric. The degree k_i of node i is the number of edges connected to it and can easily be determined from the adjacency matrix: $k_i = \sum_j A(i, j) = \sum_j A(i, j)$. The total number of the edges in the graph can also be calculated from A: $N_e = 1/2 \sum_i \sum_j A(i, j)$.

With the above definitions the modularity Q of a given partition is defined as

$$Q = \frac{1}{2N_e} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2N_e} \right] \delta_{c(i)c(j)}. \tag{1}$$

Here, δ is the Kronecker delta, and $c(i)$ is the community (subgraph) of node i, i.e. for node pairs of the same community

we have $\delta = 1$ and for pairs of different communities we have $\delta = 0$. The second term behind the sum $(k_i k_j / (2N_e))$ is simply the probability that nodes i and j are connected. The division by $2N_e$ is only for the normalization of the value of Q. Therefore Q is large if the number of connections within a community is larger than the average degree of the nodes over the whole network. Now the aim of the algorithm is to find a partitioning of the network which maximizes the modularity Q. Since this problem is known to be an NP-hard problem [12], it is suspected that the best division cannot be found in polynomial time, i.e. the computational time will escalate as the graph size increases.

If a network is divided into just two communities, one can define the vector $\mathbf{S}$ to represent the nodes in the two groups. Let $S(i) = 1$ when node i is in the first group, and $S(i) = -1$ otherwise. With $\mathbf{S}$ the modularity Q defined by (1) can be rewritten in a more convenient way [17]

$$Q = \frac{1}{4N_e} \mathbf{S}^T \mathbf{B} \mathbf{S}, \tag{2}$$

where $\mathbf{B}$ is the modularity matrix, whose elements are

$$B_{i,j} = A_{i,j} - \frac{k_i k_j}{2N_e}. \tag{3}$$

In the present paper the maximization of Q is achieved in three steps. First, using Newman's spectral algorithm an initial division of the network is given. Second, this division is further improved with a so called fine-tuning step employing the Kerrighan-Lin algorithm [18]. Finally, in order to further increase the value of Q a genetic algorithm is used in which the fine-tuning method is regularly used. This way we split the graph into two subsystems. If more subsystems are needed, we restart the computations on one of the previously found communities.

Now our aim is to find the $\mathbf{S}$ that maximizes Q and contains only 1 or -1 values. One way of achieving this is to use the largest eigenvector of $\mathbf{B}$: assuming that $\mathbf{S}$ can be written as a linear combination of the orthonormal eigenvectors u_j of $\mathbf{B}$, i.e. $\mathbf{S} = \sum_j \alpha_j u_j$ and using that $\alpha_j = \mathbf{S}\mathbf{u}'_j$, we have

$$Q = \frac{1}{4m} \left[\sum_j \alpha_j \mathbf{u}'_j \right] \mathbf{B} \left[\sum_j \alpha_j \mathbf{u}_j \right] = \frac{1}{4m} \sum_j \left(\mathbf{S}\mathbf{u}'_j \right)^2 \beta_j, \tag{4}$$

where β_i is the eigenvalue of $\mathbf{B}$ corresponding to the eigenvector $\mathbf{u}_i$. From the right-hand side of (4) it is straightforward to see that a good guess of the division would be to choose S according to the term of the largest eigenvalue β_n. To do that we write $S(i) = -1$ if $u_n(i) < 0$, and $S(i) = 1$ otherwise.

Our numerical experiments with real-life WDSs show that the above method does not give a good division (see Fig. 2a for such example), but it can be used as an initial guess, that needs to be further improved. Newman suggested that to improve the quality of the division (i.e. to further increase Q) one should use the Kernighan-Lin algorithm [18], often referred to as
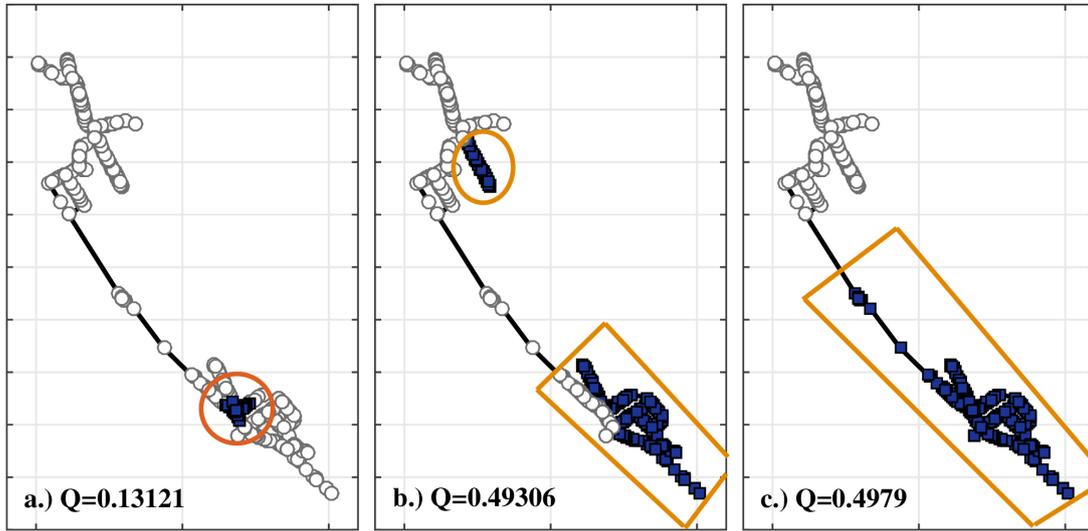
Fig. 2 Result of the community splitting a.) without fine-tuning, b.) with fine-tuning and c.) with genetic algorithms.

fine-tuning method. This technique systematically moves each node *at the boundary of the two subsystems* to the opposite community and the new value of Q is computed. When all the nodes have been moved, we pick the one that has resulted in the largest change in Q. In every step only one node is moved and the new value of Q is calculated, and one node can only be moved once. When every node has been moved the configuration that has the largest Q is kept, but only if it is larger than the original Q. This whole procedure is repeated until no further increase in Q can be reached (see also [19]).

We found that even applying the above fine-tuning method did not provide an acceptable result in many cases (see Fig. 2b) but is extremely time-consuming for large networks. In order to overcome this problem and to speed up the computation, we embedded the above technique into a genetic algorithm (GA). For the GA computations, binary coding was used; i.e. the gene length was equal to the number of nodes N_n and each gene values (-1, 1) represent the community, where a given node is concerned. The computations were performed using the built-in GA solver of MATLAB with default settings. Apart from the population size, which was set to 10 individuals. The computations were performed on up to 100 generations, then the fine-tuning process was ran on the best individual. If the separation was not complete, the above steps were repeated until we obtained two disjoint region. The result of such an optimization is depicted in Fig. 2c.

2.2 Hydraulic simulations

In this section we briefly summarize the hydraulic model and its solution technique. We used the in-house software "STACI" developed in C++ at the Dept. of Hydrodynamic Systems, which, similar to EPANET [20], solves the nodal mass conservation equations, given by

$$\sum_j \delta_j \dot{m}_j = D_i, \text{ for all nodes} \tag{5}$$

and the general edge equations

$$h_s - h_e = f\left(\dot{m}_j\right), \text{ for all edges.} \tag{6}$$

In Eq. (5), $\dot{m}_j$ is the mass flow rate (kg/s) through the jth edge connected to the actual node and D_i (kg/s) is the demand at the ith node. As all edges are directed (to define the positive direction of flow), $\delta_j = 1$ if the actual edge terminates at the node (i.e. positive flow of the jth edge is an inflow to the node) and $\delta_j = -1$ if it starts at the edge (i.e. positive flow of the jth edge is an outflow from the node).

Equation (6) describes a general edge element connecting two nodes, e.g. a pipe, a pump or a valve. h_s and h_e are the heads $\left(p/(\rho g)\right)$ at the starting and terminating nodes of the pipe, respectively, and $\dot{m}_j$ is the mass flow rate through the edge. The edge equations are formulated in terms of head (mwc, metres of water column) rather than in terms of pressure (Pa) because this way, after combining them with the nodal equations (5), the resulting system is well-conditioned (numerically).

Combining the nodal and edge equations – (5) and (6) – we end up with a system of $N_n + N_e$ nonlinear algebraic equations ($N_{n,e}$ are the number of nodes and edges, respectively) for the N_n number of unknown nodal heads and N_e number of edge mass flow rates. Formally, this can be re-written as

$$\mathbf{0} = \mathbf{F}(\mathbf{x}). \tag{7}$$

where $\mathbf{F}$ is a $\left(N_n + N_e\right) \times 1$ (column) vector of the equations and the $\mathbf{x} = \left(h_1, h_2, \ldots h_{Nm}, \dot{m}_1, \dot{m}_2, \ldots \dot{m}_{Ne}\right)^T$ is the unknown vector.

The above system is solved by means of the damped Newton-Raphson technique. During numerical implementation, rather than inverting the Jacobian, we solve the linear system

$$\mathbf{J}(\mathbf{x}_n)\delta\mathbf{x} = \omega\mathbf{F}(\mathbf{x}_n) \qquad (8)$$

for $\delta\mathbf{x} = \mathbf{x}_{n+1} - \mathbf{x}_n$, where the relaxation parameter $0 < \omega \leq 1$ is adopted during the iteration process. Moreover, we employ the sparse structure of the Jacobian and use the UMFPACK sparse linear solver (for details, see [21]) to obtain the solution efficiently.

2.3 Sensitivity analysis

Sensitivity analysis provides information on to what extent the hydraulic parameters (nodal pressures and pipe flow rates) change if a parameter is varied, i.e. we wish to compute $\partial\mathbf{x}/\partial\mu_i$, where μ_i is the actual parameter. In our case, we considered the pipe roughness and the nodal demand sensitivities, since these are the key uncertain parameters, when modelling a WDS. Roughness is usually estimated using engineering tables and proposed relationships based on pipe diameter, material and age. Water demand is inherently a stochastic variable [30] and it depends on many factors, such as climate, household size, urban density, water use policies and price.

Notice that instead of using finite differences (as in [6, 22, 23]), there is a more convenient way to perform the computation. As the hydraulic solver will solve

$$\mathbf{0} = \mathbf{F}(\mathbf{x}(\mu_i), \mu_i), \qquad (9)$$

(where we emphasized the dependence of the solution vector $\mathbf{x}$ on μ_i and also the explicit appearance of μ_i in the hydraulic equations), upon differentiating the above equation with respect to μ_i, we obtain

$$0 = \frac{\partial\mathbf{F}}{\partial\mathbf{x}}\frac{\partial\mathbf{x}}{\mu_i} + \frac{\partial\mathbf{F}}{\partial\mu_i} \quad \rightarrow \quad -\mathbf{J}\frac{\partial\mathbf{x}}{\partial\mu_i} = \frac{\partial\mathbf{F}}{\partial\mu_i}, \qquad (10)$$

where the Jacobian was already computed during the Newton iteration and $\partial\mathbf{F}/\partial\mu_i$ is the derivative of the system with respect to the sensitivity parameter. If the sensitivity parameter is the demand, only the corresponding continuity equation (5) must be differentiated, while if we compute roughness sensitivities, Eq. (6) must be differentiated that can also be performed analytically. In both cases, these will be the only non-zero elements of the $\partial\mathbf{F}/\partial\mu_i$ vector.

In what follows, we use the following notations for the roughness sensitivities:

$$S_{\lambda,ij} = \left|\frac{\partial p_i}{\partial\lambda_j}\right| \quad i\in[1, N_n] \quad j\in[1, N_e]. \qquad (11)$$

and for demand $q = \dot{m}/\rho$ variation:

$$S_{q,ij} = \left|\frac{\partial p_i}{\partial q_j}\right| \quad i, j\in[1, N_n]. \qquad (12)$$

where p_i and q_i are the pressure and demand at the ith node, λ_j is the friction coefficient of the jth pipe, N_n and N_e are the number of the nodes and pipes, respectively.

The nodes with high sensitivities are affected mostly by the hydraulic behaviour of the network; thus, when trying to identify measurement points one might choose the most sensitive ones [22]. Let $\mathbf{X}_k$ denote the discrete set of k monitoring point locations, then the objective function to maximize (see also [7]) is

$$F_1(\mathbf{X}_k) = \sum_{j=1}^{N} a_j, \quad a_j = \max_{i\in X_k}(S_{\zeta,i,j}), \quad j\in[1, N], \qquad (13)$$

where ζ is the parameter (friction coefficient λ or demand q) and N equals to N_e (number of pipes, in the case of roughness sensitivity) or N_n (number of nodes, in case of demand sensitivity). To put it simple, for k sensor location candidate nodes, we pick the corresponding rows from the sensitivity matrix and then pick the *column-wise* maximum values. This ensures that if two nodes are "very" sensitive to the same pipe roughness, only one of them will be picked and hence the present pipe will not be overrepresented in the measurement. An upper bound $F_{1,\max}$ can be defined, by allowing all the nodes to be chosen ($k=N_n$) and the lower bound is $F_{1,\min} = 0$.

Fig. 3 and Fig. 4 depict the relative roughness and demand sensitivity (the sensitivity values of a given node summed over all parameters, relative to the highest one) of the test systems from Fig. 1, respectively. As it can be seen, the sensitivities (isolevels) correlate reasonably: the high and low sensitivity regions are located in the same segment of each network.

As expected, being constant-head nodes, the sensitivity of the inlet node (pumping station modelled by constant pressure) and of the reservoir (prescribed water level) are zero. It is clearly seen that there are *regions* of high sensitivity, meaning that if we used purely (13) to find the measurement locations (e.g. simple ranking methods such as in [22]), we would end up with all the sensors accumulating in a relatively small area, which is not useful, as explained in [3]. Therefore, in the next section we introduce a second objective that ensures the *diversity* of the measurement locations.

2.4 Shannon's entropy

Shannon introduced the entropy [24, 25] as a probabilistic measure of uncertainty or information. To understand the information aspect of entropy consider a random vector $\mathbf{X}$ containing N elementary events. Each $x_j\in\mathbf{X}$ event has a probability of occurrence $\rho(x_j) = \rho_j$ and the union of all x_j gives the certain event. In information theory, the Shannon entropy defined as

$$H(\mathbf{X}) = -\sum_{j=1}^{N}\rho_j\cdot\ln\rho_j, \qquad (14)$$

which is closely related to the diversity in ecology: ρ_j is often the proportion of individuals belonging to the jth species in the dataset of interest. Then the Shannon entropy quantifies the uncertainty of predicting the species of an individual that
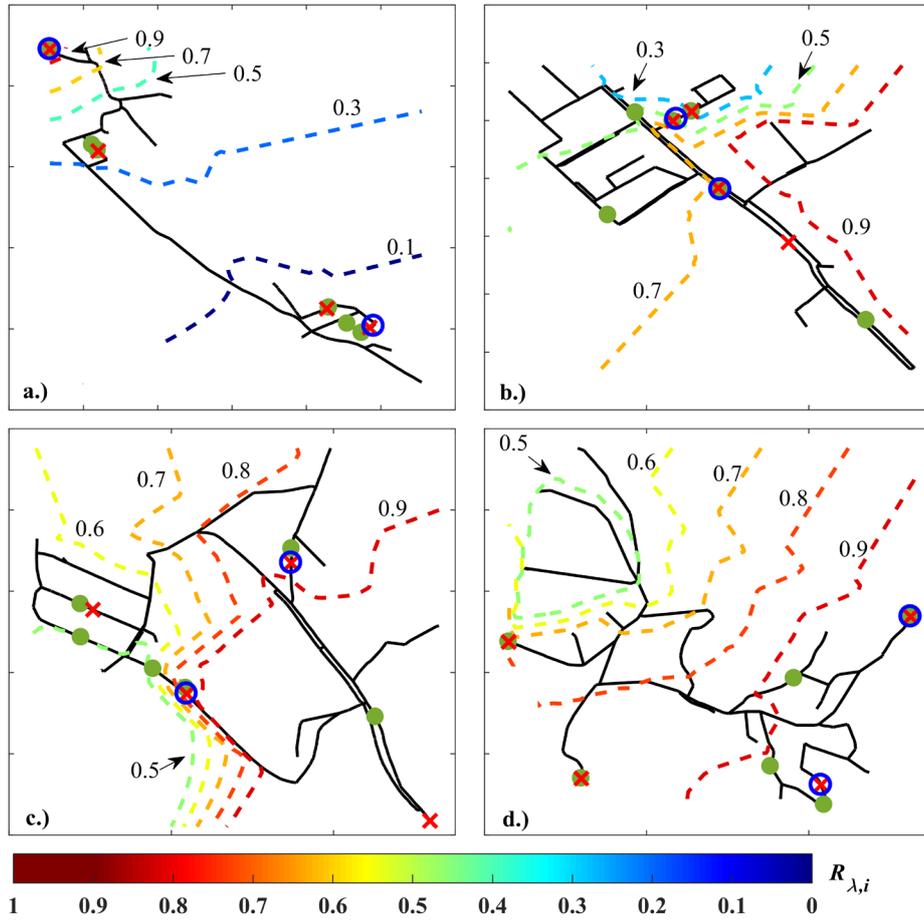
Fig. 3 The relative roughness sensitivities (isolevels), and the optimal layouts of 2 (blue circels) 4 (red crosses) and 6 (green dots) devices based on roughness sensitivity optimization.

is taken at random from the dataset [26-28]. Now, the second cost function is based on the entropy function (14) and ensures the diversity of the measurement locations (see [7] for details)

$$F_2(\mathbf{X}_k) = -\sum_{j=1}^{N} \rho_j \ln \rho_j, \quad \rho_j = \frac{a_j}{\sum_{j=1}^{N} a_j}. \quad (15)$$

Jayne's maximum entropy principle [26, 27] proposes to choose the distribution function, which maximizes Shannon's entropy and consistent with a given set of constraints. The entropy function (15) defines the sampling set $\mathbf{X}_k$ with the most evenly spread of pipe roughness or nodal demand. The entropy reaches its maximum value in case of uniform probability $\rho_j = 1/N$, thus an upper bound can be defined as $F_{2,\max} = \ln(N)$. The lower bound $F_{2,\min}$ is assumed to be zero.

2.5 Optimization by means of evolutionary algorithms

We are now in the position of formulating the optimization problem, which will maximize the sensitivity condition (13) and the Shannon entropy (15) using a multi-criterion distance based technique to find candidate solutions. Although this is a typical multi-objective problem, we shall handle the problem as a single-objective one, where the relative importance of the

two objectives is set by weights (see e.g. [7]). Let w_i denote the weight of the ith objective and $F_{i,\max}$ its maximum value, then the objective to minimize is

$$f(\mathbf{X}_k) = \sqrt{\sum_{i=1}^{2} w_i \left(\frac{F_i(\mathbf{X}_k) - F_{i,\max}}{F_{i,\max}} \right)^2}. \quad (16)$$

This programming approach defines the closest solution (by some distance measure) to an ideal (usually infeasible) solution $(F_{1,\max}, F_{2,\max})$. The objective is to define the sampling set that has the minimum values of (16). It is worth mentioning that the problem to identify k measurement locations among N_n nodes in the WDS has

$$\binom{N_n}{k} = \frac{N_n!}{k!(N_n - k)!} \quad (17)$$

possible candidates. For example, a system with 549 nodes (WDS c. in Fig. 1 with 5 pressure sensors, this gives approx. 408 billion possible layouts. This means that the complete enumeration is impossible for real-size networks; which, together with the fact that the decision variables are integers (ID of the node), led us to the use of a genetic algorithm (GA) again to solve the problem
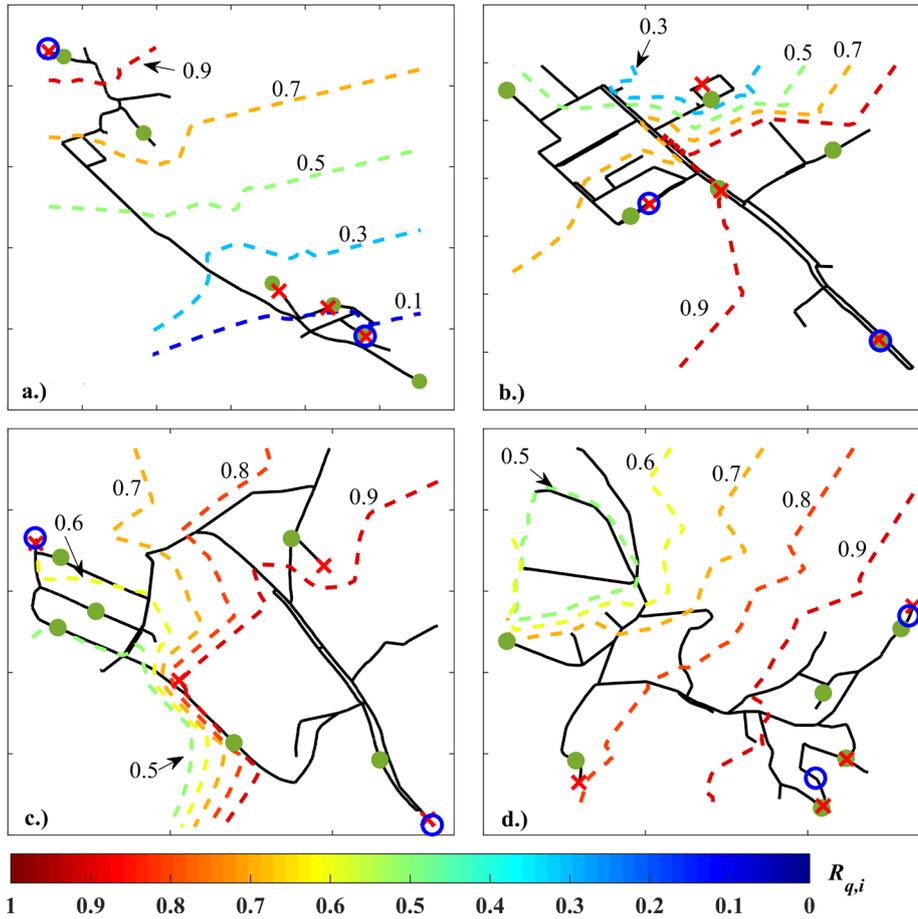
Fig. 4 The relative demand sensitivities (isolevels), and the optimal layouts of 2 (blue circles) 4 (red crosses) and 6 (green dots) devices based on demand sensitivity optimization.

efficiently. Here, the gene is defined as the discrete set of the k monitoring points: $\mathbf{X}_k = \left[x_1, x_2 \ldots, x_k \right]$, where x_p is a monitoring point location (row of a given sensitivity matrix). Each monitoring point can be chosen only once, thus $x_p \neq x_q$, $\forall p, q \in \left[1, k \right]$. The GA computations were performed up to 500 generations, with the following parameters: population size of 50, stochastic uniform selection, Gaussian mutation where the scale and shrink values were 1, and scattered crossover was set.

2.6 Parameter estimation and calibration of hydraulic models

Before any hydraulic model can be used, they need to be calibrated. The identification of good sampling points in a WDS is necessary to ensure that the accuracy of the calibrated model is reasonable. The design of the good sampling layout throughout the calibration process was used widely in previous papers, see e.g. [6, 8, 10, 29]. In the present paper we used this approach as a reference to compare with the above described sampling design approach.

During the measurements, errors (i.e. uncertainties) in pressure propagate to the calibration parameters (diameter, roughness, etc.) resulting uncertain model prediction. The goal is to determine the sampling set with the lowest uncertainty in

model parameters or predictions. The estimation of the parameter and prediction uncertainties is based on the First Order Second Moment (FOSM) [6] theory. The first order approximation of the parameter matrix:

$$\mathbf{Cov}_a = \sigma^2 \cdot \left(\mathbf{J}^T \mathbf{J} \right)^{-1} \tag{18}$$

where σ^2 is the error variance of the pressure measurements, $\mathbf{J}$ is the Jacobian matrix of the derivatives:

$$\frac{\partial y_i}{\partial a_k}, \quad i \in \left[1, \ldots k \cdot N_t \right], \quad j \in \left[1, \ldots N_a \right], \tag{19}$$

where y_i are the model predictions (pressures), a_k is the estimated model parameter (here, the cluster of pipes with the same roughness coefficient); k, N_t and N_a is the number of devices, observations in time, and estimated parameters, respectively.

Now, the main objective is to minimize the uncertainty of estimated parameters using A-optimality (minimize)

$$F_A = \frac{1}{N_a} \sum_{i=1}^{N_a} Cov_{a,ii}^{1/2}, \tag{20}$$

or D-optimality (maximize) criterion

$$F_D = \left(\det \left(\mathbf{Cov}_a^{-1} \right) \right)^{1/(2N_a)}. \tag{21}$$

By using the calibrated parameters with uncertainty (18) in the hydraulic simulations, the errors propagate to the prediction in pressures. The model prediction variance-covariance matrix can be estimated by means of the first order approximation:

$$\mathbf{Cov}_z = \mathbf{J}_z \cdot \mathbf{Cov}_a \cdot \mathbf{J}_z^T \qquad (22)$$

where J_z is the Jacobian matrix of the derivatives at the model prediction of interest:

$$\frac{\partial z_i}{\partial a_k}, \quad i \in [1,...n_z], \quad j \in [1,...n_a], \qquad (23)$$

where z_i is the model prediction value (e.g.: pressures at pre-defined locations) and N_z is the number of such predictions. Then, V-optimality criterion is used to minimize the model prediction uncertainty:

$$F_V = \frac{1}{N_z} \sum_{i=1}^{N_z} Cov_{z,ii}^{1/2}. \qquad (24)$$

Note that, in order to estimate the parameters, the number of observations $N_0 = k \cdot N_t$ must be greater than the number of the calibrated parameters N_a, otherwise the matrix $(\mathbf{J}^T\mathbf{J})$ will be badly conditioned. This means that we cannot use the same Jacobian (or sensitivity matrix) as in Eq. (11). If we would handle all pipe roughness in a WDS as a single parameter, the number of observations would be greater than the number of pipes N_e, resulting an infeasible method. For example, if the network contains 500 pipes and one has to deploy 5 devices, at least a one hundred observation in time is required to solve the calibration problem. Therefore, a common practice to reduce the number of calibrated parameters is to group pipe roughness coefficients [10], and increase the number of demand conditions [8]. Then, the elements of Jacobian matrix are approximated by using finite difference method [9].

3 Motivating example

The aim of this section is to demonstrate the above described methodologies in case of an artificial network depicted in Fig. 5. The network contains $N_n = 8$ nodes and $N_e = 10$ pipes. The system is fed by a pump modelled by constant head (45 m) and by a reservoir (bottom level 160 m, water level 2m) connected to node 1 and 3, respectively. The network node and pipe data are given in Table 1 and Table 2, respectively.

The first step of the approach is to divide the network into subsystems. Here, we will divide the network into only two communities. To obtain an initial guess, the eigenvalue approach (4) is applied. By computing the eigenvectors of the modularity matrix **B**, based on the resulting **S** vector the two communities, contain nodes 1, 2, 3, 4, 5 and nodes 6, 7, 8, respectively. The corresponding modularity is $Q = 0.155$. The modularity value will be increased by using the fine-tuning method. Node 4, 5, 6 and 7 are located at the boundary of the subsystems. These nodes are systematically moved to the opposite community

then the new Q value is computed. When every node is picked, the node that has resulted the largest increase in the Q value is moved to the opposite community. Moving node 4 then node 5 increases the modularity value up to $Q = 0.195$ then $Q = 0.22$, respectively. Further Q increase cannot be reached for this network. Thus, the final communities are composed by nodes 1, 2, 3 and 4, 5, 6, 7, 8, respectively. It should be mentioned, the initial communities are connected by four edges, while the finals are connected only by two edges. Thus, after the fine tuning the external connections are decreased that was one of the conditions of the subdivision.
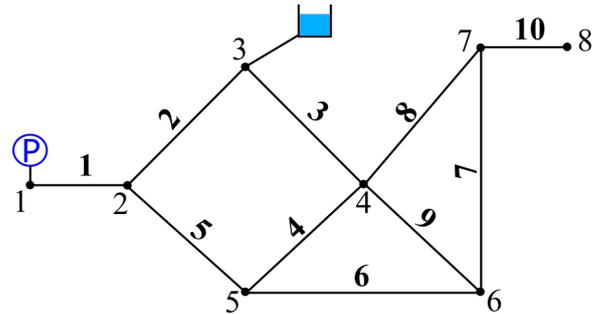
Fig. 5 An artificial test network.

Table 1 Node Data

Node	Elevation (m)	Demand (m3/h)
1	118	0
2	125	40
3	133	30
4	123	25
5	124	20
6	117	15
7	112	15
8	111	20

Table 2 Pipe Data

Pipe	Diameter (mm)	Length (m)	H-W Coeff.
1	250	200	120
2	200	400	130
3	150	400	110
4	150	800	110
5	200	800	130
6	150	1000	110
7	200	8200	130
8	200	8000	130
9	200	200	130
10	250	500	120

Table 3 Normalized roughness sensitivity matrix

Node \ Pipe	1	2	3	4	5	6	7	8	9	10	$\sum_{j=1}^{10} R_{\lambda i,j}$
1	0	0	0	0	0	0	0	0	0	0	0
2	0.4021	0.1151	0.0597	0.0168	0.0353	0.0146	2e-4	2e-4	4e-4	0	0.6443
3	0	0	0	0	0	0	0	0	0	0	0
4	0.1938	0.0555	1	0.2465	0.5174	0.2137	30e-4	31e-4	53e-4	0	2.2383
5	0.3070	0.0879	0.4891	0.1377	0.8197	0.1194	17e-4	17e-4	30e-4	0	1.9671
6	0.1974	0.0565	0.9841	0.2345	0.5269	0.2346	0.0108	0.0110	0.0191	0	2.2748
7	0.1956	0.0565	0.9920	0.2405	0.5221	0.2241	0.6437	0.6673	69e-4	0	3.5484
8	0.1956	0.0560	0.9920	0.2405	0.5221	0.2241	0.6437	0.6673	69e-4	6e-4	3.5923

Table 4 Normalized demand sensitivity matrix

Node \ Demand	1	2	3	4	5	6	7	8	$\sum_{j=1}^{8} R_{qi,j}$
1	0	0	0	0	0	0	0	0	0
2	0	0.0437	0	0.0211	0.0334	0.0215	0.0213	0.0213	0.1622
3	0	0	0	0	0	0	0	0	0
4	0	0.0211	0	0.3533	0.1728	0.3476	0.3505	0.3505	1.5957
5	0	0.0334	0	0.1728	0.2736	0.1759	0.1743	0.1743	1.0043
6	0	0.0215	0	0.3476	0.1759	0.3681	0.3577	0.3577	1.6284
7	0	0.0213	0	0.3505	0.1743	0.3577	0.9668	0.9668	2.8375
8	0	0.0213	0	0.3505	0.1743	0.3577	0.9668	1	2.8707

Now, one can identify the sampling locations in the sub-networks. The simplest way is to pick the most sensitive nodes from both communities. The results of the sensitivity analysis with respect friction coefficient and nodal demand are summarized in Table 3 and Table 4, respectively. The sensitivity values are normalized with the maximum sensitivity value for each case. The more sensitive node are those with higher row-wise sum (last columns), since these are affected mostly by the hydraulic behaviour of the network. It is worth mentioning that the pipes and nodes with higher sum per column has greater impact on the hydraulic state of the WDS. The first community contains node 1, 2, and 3, however only node 2 can only be proposed as a measurement point. Node 1 and 3 are constant-head nodes, connected to a pump station (prescribed pressure) and reservoir (prescribed water level). Thus, as expected, these are zero sensitivity nodes. In addition, these two nodes do not affect the hydraulic behaviour, since their sum per column is zero in case of demand sensitivity, see Table 4. The second community allows more freedom to consider sensitive nodes as potential measurement locations. However, the most sensitive nodes are picked with respect pipe roughness and nodal demand which are node 8 and node 7, respectively.

To demonstrate the calculation of the objective that maximizes sensitivity and diversity, consider the case of deploying three measurement devices based on the roughness sensitivity matrix. Let $\mathbf{X}_k = [2,5,6]$ be e discrete set of devices. We pick the corresponding rows of the sensitivity matrix (Table 4) then pick the *column-wise* maximum (underlined in the table) values a_j. These maximum values are summarized to obtain the value of the first objective (13) that is $F_1(\mathbf{X}_k) \cong 2.83$. Now, each ρ_j can be calculated from a_j using Eq. (15), then the value of second objective $F_2(\mathbf{X}_k) \cong 1.62$ can be obtained. With using equal cost function weighting $w_1 = w_2 = 0.5$, and $F_{1\max} \cong 4.19$ and $F_{2\max} = \ln(10) \cong 2.30$, the combined objective (16) is $f(\mathbf{X}_k) \cong 0.31$. The above computation is the same in terms of demand sensitivities. Searching for the minimum results in the optimal layout of three devices, that is the set of nodes 2, 5 and 8; and the corresponding objective is $f(\mathbf{X}_k) \cong 0.11$. Using demand sensitivities, similar node set is identified as optimal measurement points, however, instead of node 8, we pick node 7 as the third sample point.

The classical FOSM based approaches requires to use grouped pipe roughness coefficients. From the above Pipe Data (see Table 2) three aggregate groups of pipes can be composed for H-W roughness coefficients 110, 120 and 130.

The elements of Jacobian are the nodal-pressure (19) derivatives with respect the roughness parameters, which was approximated by finite differences. The roughness value of each pipe in a given group was increased by 10 percent, then the pressures (heads) were computed at the nodes by means of

"STACI". After that, the approximate values of the derivatives was calculated using forward finite difference method. The obtained H-W-roughness parameter-group sensitivity matrix for one loading condition (given in Table 1) is summarized in Table 5. Since these methods are based on the calibration of the unknown pipe roughness groups, for the feasible calibration at least three observations are required. Therefore the minimum number of pressure sensors for one loading condition is three. The optimal layouts of these three sensors are the set of nodes 2, 4 and 5 using A-Optimality (20) criterion, while D-Optimality (21) and V-Optimality (24) criterions suggest the same node set 2, 4, and 8 as optimal locations.

Table 5 Jacobian matrix of grouped pipe roughness coefficients

Node	H-W 110	H-W 120	H-W 130
1	0	0	0
2	4e-4	0	0
3	0	0	0
4	0.0197	28e-4	62e-4
5	35e-4	44e-4	0.0101
6	0.0196	28e-4	66e-4
7	0.0197	28e-4	0.0240
8	0.0197	34e-4	0.0240

4 Real Life Case Studies

In this section we use the above described methodologies to locate the best sampling locations on the test WDSs shown in Fig. 1. First we investigate the effect of the number of sensors on the cost function (16) with equal weights. Then, we compare the different optimization methods: (a) the roughness and (b) demand sensitivity based optimizations of the objective function (16), (c) A-optimality (20), (d) D-optimality (21), (e) V-optimality (24) criterion, respectively. We also investigate the effect of cost function weighing with varying w_1 between

0 and 1, while the sum of the weights was kept constant ($w_1 + w_2 = 1$). Finally, the results of partitioning are presented, and within each zone the node with the highest sensitivity is picked as a sampling location.

4.1 Optimal layouts with equal weights

Fig. 6 depicts the value of the objective function (16) as a function of the number of sensors k, using equal weights. Subplots a. and b. correspond to the roughness and demand sensitivity approaches, respectively, while the red square, blue circle, green downside and black upside triangles denote WDSs a., b., c. and d., respectively, from Fig. 1. The figure shows a convergence in both cases for each WDS: beyond a certain limit, the objective of the optimal solution cannot be improved significantly, i.e. adding more sensors above a limit provides only a little amount of extra information, giving rise to an *optimal sensor number*.

The optimal layout for different number of measurement devices with the same cost function weighting ($w_1 = w_2 = 0.5$) are shown in Fig. 3 and Fig. 4 based on roughness and demand sensitivities respectively, for 2 (blue circles), 4 (red crosses) and 6 (green dots) devices. Comparing the corresponding maps in the figures one can observe slightly different locations in the case of the two sensitivity matrices. In the case of systems a., c. and d. the sensors are placed in similar locations, however, for system b. the proposed sampling points are different. It is interesting to note that the technique is robust in the sense that upon adding new sensors, the old ones did not move significantly.

4.2 Comparison of different optimization methods

The resulting layouts of different optimization methods are compared in Fig. 7 using four pressure loggers. The blue and red triangles denote the layouts obtained by means of roughness (11) and demand (12) sensitivity approach with maximization of Shannon entropy (16) using equal ($w_1 = w_2 = 0.5$)
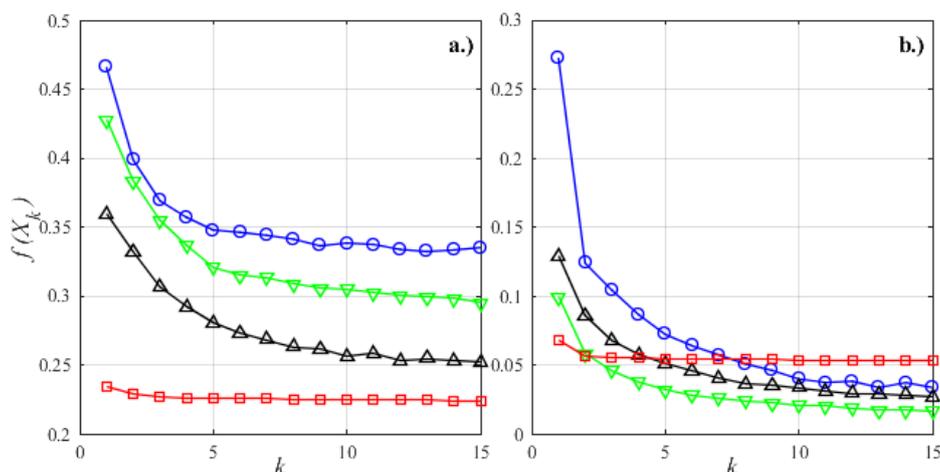
Fig. 6 The optimal values of the objective function with respect the number of measurement devices. Red square-WDS a., blue circle-WDS b., green downside triangle-WDS c., black triangle-WDS d., see Fig 1. Subplot a.) and b.) correspond to the roughness and demand sensitivity approaches, respectively.
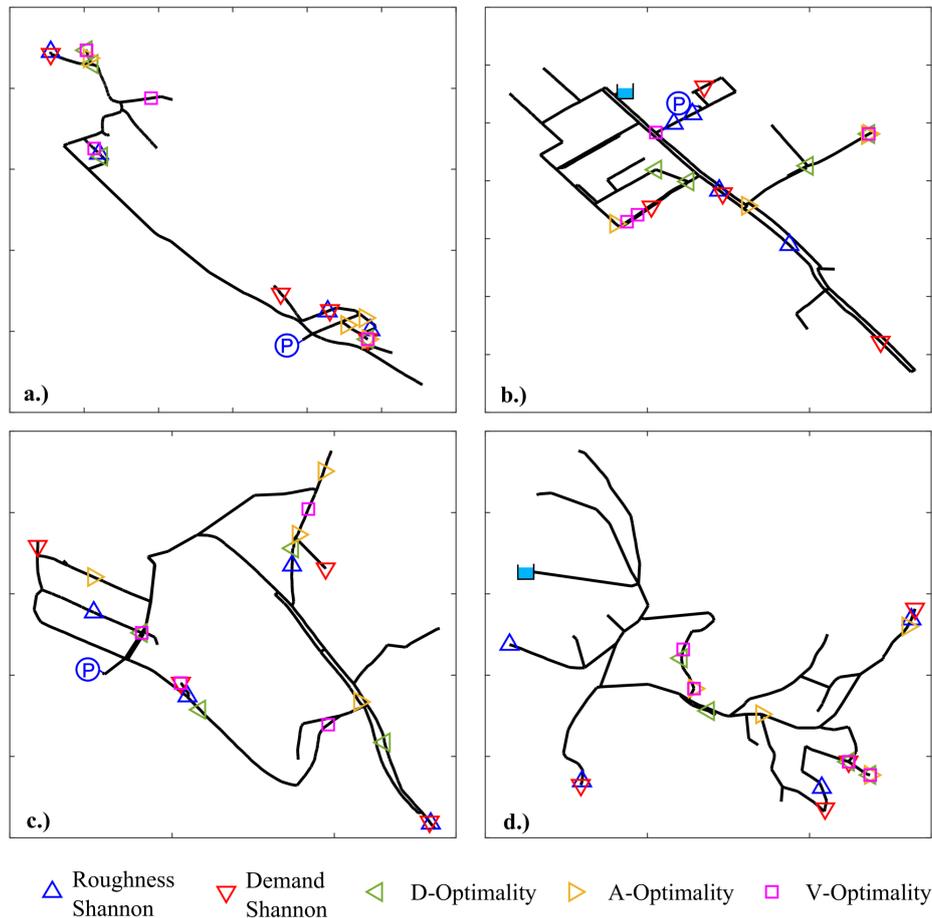
Fig. 7 Comparison of the layouts obtained by means of different methodologies.

cost function weighting, while the green, yellow and magenta markers denote the measurement locations obtained by means of D-optimality (21), A-optimality (20) and V-optimality (24) criterion, respectively.

The Jacobian matrix (19) of the derivatives of Network a., b., c., and d., was calculated using N_a = 4, 5, 6 and 4 grouped pipe roughness coefficients, respectively, while the number of observations in time N_t was increased to five by artificially prescribing demand scenarios considering the scaling properties [30] of water demand statistics. The standard deviation of pressure loggers was assumed to $\sigma = 1$m.

In general, the comparison shows good agreement between the different methodologies in case of all investigated WDSs. The located sampling points spread evenly across the networks, however, slight differences between measurement layouts can be recognized.

One of the reasons is probably that there are usually many combinations of near optimal solution (i.e. sampling layout) with similar fitness values in a large-size optimization problem. We can also conclude that the distribution of the sampling points is relatively uniform, but in some cases the FOSM based approaches result almost identical locations, (see e.g.: Network b., A-optimality.). This is probably due to that the number of observations $N_o = k \cdot N_t = 4 \cdot 5 = 20$ (spatial

and temporal) is higher, than the minimally required for calibration process.

Finally, the measurement locations are usually located away from the pump stations, reservoirs (see Fig. 1), and transmission pipes, respectively. This confirms the suggestion of Walski [3] that has also been verified in previous studies [8, 10].

4.3 The effect of cost function weighting and community structure

The effect of cost function weighting was also investigated. The first weight w_1 was varied between 0 and 1 such that the sum of the weights was kept constant ($w_1 + w_2 = 1$). In Fig. 8 and Fig. 9 the optimal layouts of 4 measurement devices with three different cost function weightings ($w_1 = 1$, 0.5, 0) are given for roughness-based and demand-based sensitivity matrices, respectively, denoted by brown, blue, and red triangles. From the result of computations, it was observed regardless of the sensitivity approach, that the technique is insensitive to the weighing in the range of $w_1 = 0.1…0.9$. In this range the resulting layouts are almost equivalent or very similar to the one with cost function weighting $w_1 = 0.5$. It should be noted that the demand-based approach tends to place the devices to dead ends and also, in the case of $w_1 = 0$ (purely Shannon-based objective) one can observe poor results as all the measurement
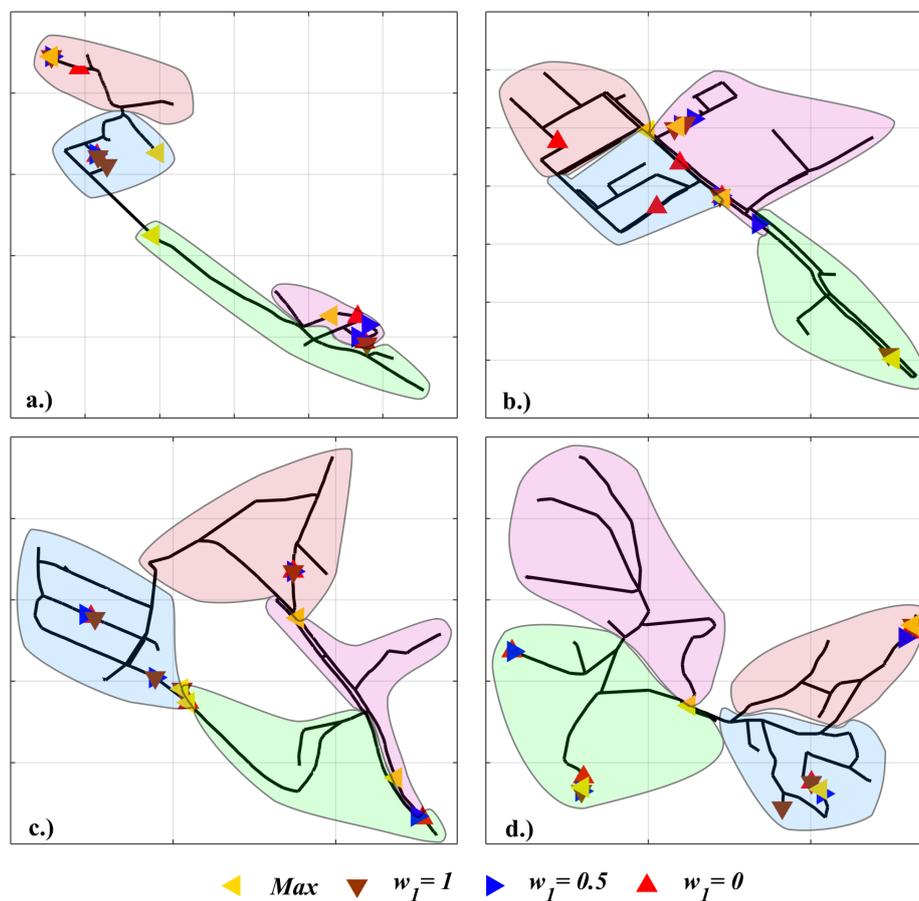
devices are deployed very close to each other and along the main supply route, see Fig. 9. By comparing Fig. 8 and Fig. 9 we can conclude that both sensitivity approaches result in a similar distribution of measurement devices.

Finally, the partitioning of the four test networks into four disjoint sectors was performed with the method of modularity maximization explained earlier. The results are shown in Fig. 8 and Fig. 9 with differently coloured domains corresponding different communities, and the node with the highest roughness and demand sensitivity in each community is denoted by yellow triangles ("*Max*"). It should be noted the sensitivities increase almost gradually in a given direction. This actually explains why the nodes at the boundaries of communities are picked in a few cases. Although it is the simplest approach to assign one pressure logger (per community), it is also possible to use other methodologies to locate one or more sampling point in each measurement zone.

Comparing the partitioning with the results of the different function weighting one can conclude that for both the roughness and demand sensitivity approaches it is not true that in every sector one sensor is placed. However, in the case of equal cost functions (blue triangles) the sensors are almost evenly spread between the sectors.

5 Summary

In this paper the problem of optimal measurement layout design was investigated on real-life water distribution systems. First, a method was presented that provides a partitioning of the network based only on its topology. The described algorithm is a slightly modified version of the Newman algorithm augmented with fine tuning embedded in a genetic algorithm to maximize modularity. Then the measurement layout design problem was formulated as a single-objective optimization, where the objective was to maximize the linear combination of the overall sensitivity and Shannon's entropy. Roughness and demand sensitivity approaches were used for the optimization, which were performed by means of a standard genetic algorithm.

As a result a convergence was observed in the sense that upon adding new sensors, beyond a certain limit (different for each investigated network) the optimal value of the objective function was not improved significantly. This means that there exists an optimal number of pressure sensors, which provides a reasonable compromise between measurement effort and accuracy. This method was found to be robust, regardless of the sensitivity approach: upon adding a new sensor, it did not change the location of the previously mounted sensors. The effect of
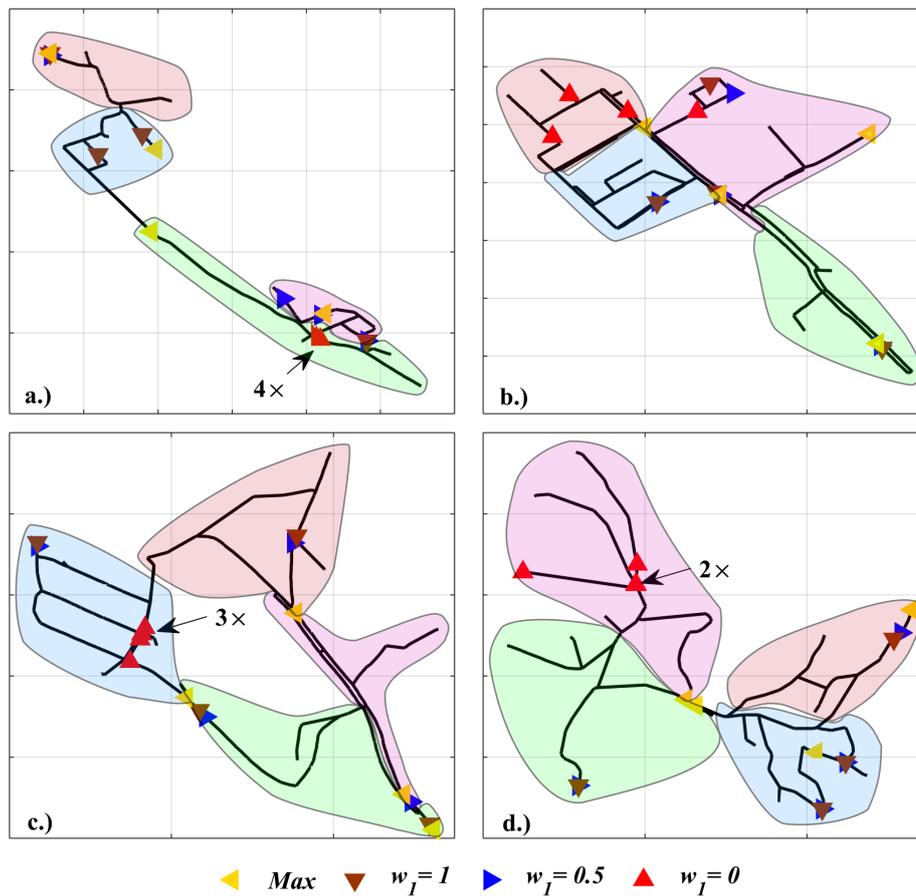
Fig. 9 The optimal measurement layouts of four sensors as a result of demand sensitivity approach with 3 cost function (16) weightings, and the communities with the most sensitive nodes.

cost function weighing was also investigated. The results show that the distribution of the pressure sensors are similar in a wide range of cost function values ($w_1 = 0.1...0.9$).

Finally, the partitioning of the test networks into four sectors was compared with the distribution of four pressure sensors with different cost functions. In the case of equal cost functions, in almost every case, three sensors out of the four were placed in different sectors.

Acknowledgement

References

[1] Ridolfi, E., Servili, F., Magini, R., Napolitano, F., Russo, F. "Artificial neural networks and entropy-based methods to determine pressure distribution in water distribution systems." *Procedia Engineering*. 89, pp. 648-655. 2014.
https://doi.org/10.1016/j.proeng.2014.11.490

[2] Xu, J., Fishcbeck, P. S., Small, M. J., VanBriesen, J. M., Casman, E. "Identifying sets of key nodes for placing sensors in dynamic water distribution networks." *Journal of Water Resources Planning and Management*. 134(4), pp. 378-385. 2008.
https://doi.org/10.1061/(ASCE)0733-9496(2008)134:4(378)

[3] Walski, T. M. "Technique for calibrating network models." *Journal of Water Resources Planning and Management*. 109(4), pp. 360-372. 1983.
https://doi.org/10.1061/(ASCE)0733-9496(1983)109:4(360)

[4] Lee, B. H., Deininger, R. A. "Optimal locations of monitoring stations in water distribution system." *Journal of Environmental Engineering*. 118(1), pp. 4-16. 1992.
https://doi.org/10.1061/(ASCE)0733-9372(1992)118:1(4)

[5] Yu, G., Powell, R. S. "Optimal design of meter placement in water distribution systems." *International Journal of Systems Science*. 25(12), pp. 2155-2166. 1994.
https://doi.org/10.1080/00207729408949342

[6] Bush, C., Uber, J. "Sampling design methods for water distribution model calibration." *Journal of Water Resources Planning and Management*. 124(6), pp. 334-344. 1998.
https://doi.org/10.1061/(ASCE)0733-9496(1998)124:6(334)

[7] De Schaetzen W. B. F., Walters, G. A., Savic, D. A. "Optimal sampling design for model calibration using shortest path, genetic and entropy algorithms." *Urban Water*. 2(2), pp. 141-152. 2000.
https://doi.org/10.1016/S1462-0758(00)00052-2

[8] Kapelan, Z., Savic, D., Walters, G. A. "Multiobjective sampling design for water distribution model calibration." *Journal of Water Resources Planning and Management*. 129(6), pp. 466-479. 2003.
https://doi.org/10.1061/(ASCE)0733-9496(2003)129:6(466)

[9] Lansey, K. E., El-Shorbagy, W., Ahmed, I., Araujo, J., Haan, C. T. "Calibration assessment and data collection for water distribution networks." *Journal of Hydraulic Engineering*. 127(4), pp. 270-279. 2001. https://doi.org/10.1061/(ASCE)0733-9429(2001)127:4(270)

[10] Behzadian, K., Kapelan, Z., Savic, D., Ardeshir, A. "Stochastic sampling design using a multi-objective genetic algorithm and adaptive neural networks." *Environmental Modelling & Software*. 24(4), pp. 530-541. 2009. https://doi.org/10.1016/j.envsoft.2008.09.013

[11] Bhave, P. R. "*Optimal Design of Water Distribution Networks*." Alpha Science International, Pangbourne England. 2003.

[12] Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., Wagner, D. "On Finding Graph Clusterings with Maximum Modularity." In: Graph-Theoretic Concepts in Computer Science, 33rd International Workshop, WG 2007, Dornburg, Germany, June 21-23, 2007, pp. 121-132. https://doi.org/10.1007/978-3-540-74839-7_12

[13] Giustolisi, O., Ridolfi, L. "New Modularity-Based Approach to Segmentation of Water Distribution Networks." *Journal of Hydraulic Engineering*. 140(10), 04014049. 2014. https://doi.org/10.1061/(ASCE)HY.1943-7900.0000916

[14] Newman, M. E. J. "Fast algorithm for detecting community structure in networks." *Physical Review E*. 69(6), 066133. 2003. https://doi.org/10.1103/PhysRevE.69.066133

[15] Newman, M. E. J. "Finding community structure in networks." *Physical Review E*. 74(3), 036104. 2006. https://doi.org/10.1103/PhysRevE.74.036104

[16] Newman, M. E. J. "Modularity and community structure in networks." *Proceedings of the National Academy of Sciences*. 103(23), pp. 8577-8582. 2006. https://doi.org/10.1073/pnas.0601602103

[17] Newman, M. E. J. "Spectral methods for network community detection and graph partitioning." *Physical Review E*. 88(4), 042822. 2013. https://doi.org/10.1103/PhysRevE.88.042822

[18] Kerninghan, B. W., Lin, S. "An efficient heuristic procedure for partitioning graphs." *Bell System Technical Journal*. 49(2), pp. 291-307. 1970. https://doi.org/10.1002/j.1538-7305.1970.tb01770.x

[19] Sun, Y., Danila, B., Josi, K., Bassler, K. E. "Improved community structure detection using a modified fine-tuning strategy." *EPL (Europhysics Letters)*. 86(2), 28004. 2009. https://doi.org/10.1209/0295-5075/86/28004

[20] Rossman, L. A. "EPANET 2 Users Manual." U.S. Environmental Protection Agency report EPA/600/R-00/057, 2000.

[21] Davis, T. A. "Algorithm 832: Umfpack v4.3-an unsymmetric patter multifrontal method." *ACM Transactions on Mathematical Software*. 30(2), pp. 196-199. 2004. https://doi.org/10.1145/992200.992206

[22] Morosini, A. F., Costanzo, F., Veltri, P., Savic, D. "Identification of measurement points for calibration of water distribution network models." *Procedia Engineering*. 89, pp. 693-701. 2014. https://doi.org/10.1016/j.proeng.2014.11.496

[23] Greco, M., Giudice, G. "New approach to water distribution network calibration." *Journal of Hydraulic Engineering*. 125(8), pp. 849-854. 1999. https://doi.org/10.1061/(ASCE)0733-9429(1999)125:8(849)

[24] Shannon, C. E. "A mathematical theory of communication." *The Bell System Technical Journal*. 27(3), pp. 379-423. 1948. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[25] Shannon, C. E. "A mathematical theory of communication." *ACM SIGMOBILE Mobile Computing and Communications Review*. 5(1), pp. 3-55. 2001. https://doi.org/10.1145/584091.584093

[26] Jaynes, E. T. "Information theory and statistical mechanics." *The Physical Review*. 106(4), pp. 620-630. 1957. https://doi.org/10.1103/PhysRev.106.620

[27] Jaynes, E. T. "On the rationale of maximum-entropy methods." *Proceedings to the IEEE*. 70(9), pp. 939-952. 1982. https://doi.org/10.1109/PROC.1982.12425

[28] Kesavan, H., Kapur, J. N. "The generalized maximum entropy principle." *IEEE Transactions on Systems, Man and Cybernetics*. 19(5), pp. 1042-1052. 1989. https://doi.org/10.1109/21.44019

[29] Kapelan, Z. S. Savic, D. A., Walters, G. A. "Optimal sampling design methodologies for water distribution model calibration." *Journal of Hydraulic Engineering*. 131(3), pp. 190-200. 2005. https://doi.org/10.1061/(ASCE)0733-9429(2005)131:3(190)

[30] Vertommen, I., Magini, R., da Conceição Cunha, M., Guercio, R. "Water Demand Uncertainty: The Scaling Laws Approach." In: Ostfeld, A. (ed.) *Water Supply System Analysis-Selected Topics*. (pp. 105-129.), InTech, https://doi.org/10.5772/51542