SMaRT-OnlineWDN D6.1: Evaluation of existing source identification approaches

Jochen Deuerlein, Olivier Piller, Idel Montalvo Arango, Denis Gilbert, Hervé Ung

SPONSORED BY THE

Federal Ministry of Education and Research

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

Deliverable 6.1

Evaluation of existing source identification approaches

Dissemination level: Public

WP6

Online Source Identification and Risk Management

12th May 2013

SMaRT-Online WDN

Online Security Management and Reliability Toolkit for Water Distribution Networks

ANR reference project: ANR-11-SECU-006
BMBF reference project: 13N12180

Contact persons
Fereshte SEDEHIZADE Fereshte.Sedehizade@bwb.de
Olivier PILLER Olivier.Piller@irstea.fr

WP 6 – Online Source Identification and Risk Management

D6.1 Evaluation of existing source identification approaches

List of Deliverable 6.1 contributors:

From 3S [WP6 leader]
Jochen Deuerlein (deuerlein@3sconsult.de)
Idel Montalvo (montalvo@3sconsult.de)

From Irstea
Olivier Piller (olivier.piller@irstea.fr)
Denis Gilbert (denis.gilbert@irstea.fr)
Hervé Ung (herve.ung@irstea.fr)

Work package number	6	**Start date:**	01/01/2013
Contributors	3S Consult	Irstea	
Person-months per partner	1,5	1	

Keywords
Inverse problem, contaminant source, backtracking, water quality model,

Objectives
To investigate and study existing approaches for the solution of the source identification problem. New in this context is the fact that the problem is solved online based on near real-time data of the current hydraulic state of the network and water quality measurements. The desired outcome of the task is that the most promising concept and algorithms are identified and that a first draft design of the software architecture is developed.

Summary

This document describes the existing state of the art of solving the source identification problem. First, different approaches are briefly described and strength and weaknesses are discussed. At the end a roadmap for source indentification module implementations in the SMaRT-Online^{WDN} project is derived based on the review of existing methods.

Contents:

1. Problem description

In case of a contamination event that is detected by one or more sensors the first step of remedial actions includes the identification of the contamination source by backward calculation. After the source is known forward transport calculations can help to estimate the dissemination of the contaminant within the network which are the basis for response actions like isolation of contaminated network parts from the rest of the system and targeted public warnings for minimizing the exposure of the population. In general, neither the time nor the location of the contamination event are known.

The objective of existing source identification algorithms is to solve the inverse problem of determining time an location of contaminant sources after alarms by one or more sensors have been released. The general mathematical problem can be formulated as follows:

$$f(X^*) = \min_X \left\{ \sum_{i=1}^{n} \sum_{t=1}^{T} [y_i(t) - y_i'(t)]^2 \right\}$$

subject to

$$y_i(t) = F(X_i, t)$$
$$X_{min} \leq X \leq X_{max}$$

where X^* is the optimal solution, $y_i'(t)$ = measured concentration at monitoring location *i* at time *t* and X_{min} and X_{max} are bounded limits of the sources' release concentrations (Guan, et al., 2006). The equality constraint includes system hydraulics and transport of quality parameters. In the most general form they consist of a system of partial differential equations. For discretization and simplification different techniques have been used and will be discussed in the sequel.

Given the fact that only a limited number of sensors can be installed in a real distribution network the problem is ill-posed in nature. A large number of different approaches has been published during the last decade. In contrast to forward simulations where all the model parameters are assumed to be known and the state variables of the network are calculated based on the given boundary conditions, in inverse problems some of the state variables are measured by sensors and the parameters are calibrated to minimize deviations between measured and calculated values. A common approach for solution is the formulation of a least squares minimization problem in order to best fit the parameters in order to match the observed state values by calculation results. However, because the problem is underdertermined (the unknowns outnumber the observations) there may exist an infinite number of possible solutions.

Another problem of applying existing techniques to real world problems is that the number of potential contamination event is generally assessed as $n \cdot n_T$ where n is the number of nodes (possible injection locations) and n_T is the number of time steps that define the past time horizon we want to explore. In addition, uncertainties concerning the actual water demands and other model parameters like valve states, network topology and roughness can lead to inaccurate estimates of pipe flow velocities that are calculated by use of hydraulic simulation models. The flow velocity is the driving force of advective transport of the contaminant in the pipes.

In summary the difficulties that arise in the context of solving the source identification problem in real time are (De Sanctis, et al., 2010), (Propato, et al., 2010):

- inherent nonuniqueness of the solution due to limited sensor data available compared to the large number of potential contaminant source locations in a real drinking water distribution system and to graph complexity.
- measurement errors and model uncertainties (especially actual demands)
- computational effort that significantly increases with network size.

2. Methods

1.1. Overview

For solution of the source identification problem there are several categories of formulations and solving methods. Traditionally, SI was treated as inverse parameter estimation problem. The parameters are the node time pairs of possible contaminations. At this stage stochastic modeling was limited to measurement errors solely. Methods of this class endeavor to determine potential contaminant sources given perfect sensor responses and a transport model. There is no distinction between the solutions. For formulation of the source identification problem a weighted least squares minimization problem was formulated where the objective function consist of the errors between calculated concentrations and measurements at the sensors. As constraints for calculation of the sensor concentrations a hydraulic water quality model (e. g. EPANET) was used. However, shortcomings are that the reaction kinetics have to be known and that the resulting optimization problem is a large scale nonlinear problem. In order to transfer the infinite-dimensional problem into a finite dimensional, one time-discretization is applied and different optimization techniques were applied for solution.
Second method types aim to classify these potential sources with weights or probabilities in order to aid the decision-making. Ultimately, a unique solution is sought depending on the objective. In order to deal with non-uniqueness of the solution regularization techniques are applied or in two-step algorithms the first calculated candidate locations are further classified by use of statistical methods.
Finally, the third class of methods uses the alarm classifier accuracy to calculate the probability of a contamination at a connected component given the sensor responses.
In the following, first, the linear input/output model that is used by most of the source identification algorithms by considering the linear or the first order reaction kinetics is described then it is followed by a short discussion of selected methods for source identification.

1.2. Background: Concentration Input/Output linear transport model

If the reaction in the bulk flow (transport in the liquid) and with the pipe wall (including interaction with deposit/biofilm) is conservative or reactive in the first order, several authors (*e.g.*: (Boccelli, et al., 1998)) derivate that nodal concentration outputs may be predicted by a linear equation. The aim of tracking methods is to transform the system of partial differential equations describing water quality in the network in a set of algebraic equations that can be easier used as constraints for optimization methods.
For control of chlorine booster stations, (Constans, et al., 2003) introduce an appropriate linear model to constrain water quality optimization problems. Firstly, the method performs an acyclic reduced subgraph that concentrates on flows between disinfectant sources (tanks and injections) and supernodes (strongly connected components). Then, the transport-reaction matrix is worked out with a characteristic method. Also for design of feedback control algorithm, (Propato & Uber, 2004) applied a perturbation method to the transport-reaction module in Epanet to determine the influence or response coefficients. The linear I/O model proposed by (Shang, et al., 2002) differs from the previous one in that system response is computed in reverse time. Their particle (water parcel) backtracking algorithm (PBA) tracks a large number of water parcels simultaneously. The PBA may be classified as a Lagrangian method. So, the following linear equation may be used to predict the output concentrations at any node and time:

$$C^{out} = TC^{in} \quad (1)$$

Where T is the transport matrix (from/to rates that are dimensionless); C^{in} is the input concentration vector of size (number of source nodes)-by-(number of time steps over the

calculation duration); C^{out} is the output concentration vector of size (number of non-source nodes)-by-(number of time steps). The coefficients of the given linear system are calculated tracking the different paths a particle will take through the system from an output node to a source. This describes the water quality at any node and any time as a linear function of the input concentration at the source nodes (Input – Output –Model).
The PBA is used as basis of many approaches and was implemented as extension to EPANET (Rossman, 2000): EPANET-BTX: Particle Backtracking extension to EPANET (Shang & Uber, 2009). The transport matrix T of the PBA is calculated by hydraulic network simulations and describes the different paths that a particle can take from a source to an output node.

1.3. Class 1: traditional parameter estimation methods

The SI problem can be formulated as a nonlinear, infinite dimensional optimization problem subject to algebraic, ordinary differential, and partial differential constraints (Laird, et al., 2005). Input parameters are the flow profiles calculated by hydraulic simulation and measured concentrations at sensor locations. As Output, the time-dependent concentration along pipes and at junctions and the mass input at junctions as function of time are sought. For solution, direct sequential and direct simultaneuous methods were proposed. The origin tracking algorithm introduced by (Laird, et al., 2005) reformulates the water quality equations into a set of algebraic equations. In contrast to PBA it handles the pipes sequentially and describes the time delay between boundary concentrations and connected nodes for each single pipe. In order to force a unique solution a regularization term is added to the objective function.
The method was tested for a network with 469 nodes. Even for this small model the number of unknowns in the NLP already reached 210.000. As a consequence the method can not be applied to the large-size network models of Smart-Online^{WDN} with more than 50,000 nodes. Other problems of this method are: 1.) It is assumed that concentration measurements of the contaminant are available; 2.) Assumption of perfect sensors; 3.) Comparably large number of sensors required.
Other authors use stochastic optimization methods for solution of the parameter estimation problem. (Preis & Ostfeld, 2008) combine the hydraulic and water quality simulation software EPANET with a Genetic Algorihm (GA). The objective function consists of a least squares function of measured and calculated concentrations. For the measurements also imperfect sensors are taken into account. The calculation time for a test network with less than 1,000 nodes was about one hour. (Liu, et al., 2011) use an Evolutionary Algorithm (EA) for adaptive (based on updated observations) and continually search for optimal solutions of a modified least squares function. For speeding up the calculations 10 parallel 2.2 GHz processors were used for a network with 1,574 junctions (Micropolis Network). The resulting calculation time is reported as 5 minutes. For SMaRT-Online^{WDN} stochastic search algorithms are not favored since the networks are much bigger. Gradient-kind algorithms will be preferred assuming that calculating the derivatives is not time-consuming.
Several authors (e.g., (Propato, et al., 2007); (Propato, et al., 2010)) propose to apply the Eq (1) to contaminant source identification. A simplified equation (Eq. (2)) is derived where the number of rows is reduced to sensor nodes with significant contaminant concentration and where columns correspond to potential pairs (node positions by time):

$$MC^{potential} = C^{sensor} \quad (2)$$

Where M is the transport matrix (dimensionless) expressing the relationship between potential source nodes and positive sensor responses; $C^{potential}$ is the input concentration vector of size (number of potential source nodes)-by-(number of time steps over the calculation duration); C^{sensor} is the sensor concentration vector of size (number of sensor nodes)-by-(number of time

steps). We may point out that M is in general rectangular with more columns than rows, that is: the system Eq. (2) is undetermined with an infinite number of potential source solutions.
(De Sanctis, et al., 2010) use PBA for identification of all possible contaminant source locations. For alarm generation, binary sensor information is introduced. The authors claim that formulating the SI problem as an inverse water quality problem is difficult for three reasons:
1) ill-posedness (sparse sensor grid in contrast to huge number of possible sources, e.g. hydrants, house connections, 2.) problem size (number of possible sources times number of time steps within detection time), 3.) assuming existence of perfect quality sensors that are capable of measuring the concentration of all relevant substances (do not exist in reality).
It is followed by a discussion of different regularization methods and their usefulness for the solution of the SI problem. As a consequence of regularization, a unique solution of the mathematical problem could be wrong. Multisource contamination is also possible. The total source status matrix S is introduced where $S_{i,j}$ has three different states: G(green): safe, R (red): unsafe, possible source of contaminant: W (white): no impact on sensors.

1.4. Stage 2: Weighting the potential source solutions

(Propato, et al., 2007) then (Propato, et al., 2010) developed a two-step approach. First, linear algebra is employed to rule out potential contaminant injections. The result is Eq. (2) with contaminant concentrations predictions. Second, an entropic-based Bayesian inversion technique, the Minimum Relative Entropy method, solves the problem for the remaining variables. This formulation allows for the less committed prior distribution with respect to unknown information. It can include model uncertainties and measurement errors. This implies to be able to model for a specific conservative contaminant that can be measured by the warning detection system.
As mentioned above several authors suggested a Weighted Least-Squares (WLS) formulation approach. For example, (Laird, et al., 2006) added a second step to their earlier approach (Laird, et al., 2005) for better selection of most likely events. For that purpose a mixed-integer quadratic programming problem is formulated that is based on the solution of the first step which is actually the NLP explained above. The variables belonging to active mass input constraints (that means that the mass input is zero) are eliminated from the search space and the MIQP is formulated only for the rest of nonzero variables. The solution may be constrained to be a single-source contamination or even some source solution may be disregarded. This formulation is well suited for detecting specific contamination with a concentration measurement error that follows a normal distribution: in this case, the maximum likelihood estimation problem is equivalent to the Least-Square problem.
(Guan, et al., 2006) use an Ordinary Least-Squares formulation to identify the release-history of few potential contaminant nodes. The Epanet model is used to make the concentration predictions and for estimation of the derivatives. Their finding is that in any case original contaminations are retrieved. An explanation may be the particular attention paid to the placement of sensors (by Engineering judgment). They proved that their approach is robust even in presence of normal measurement errors.
(Liu, et al., 2011) infer the probabilities of being a contaminant source from a logistic regression model. Their approach comprises a training phase where Epanet is used. Measurement error has relatively small effect on the prediction while demand uncertainty may lead the true source to be not detected in 4% of cases. They report a loss of efficiency when binary responses with high detection threshold (low detection power – high beta) are used in place of chemical-specific contaminant concentration. Logistic regression corresponds to the maximum entropy classifier for independent observations.

(Preis & Ostfeld, 2006) introduced a hybrid approach for contamination source identification in water distribution systems using a coupled model trees – linear programming scheme. The data-driven technique used is model trees. The reason for using model trees instead of artificial neural networks is to accelerate the linear programming phase where linear equation classification rules generated by the model trees are used to solve the inverse problem. However, it is expensive to use in terms of its required computational resources, although most of its computational expense is due to stage 1 (i.e., building the model trees) which is run off line.

1.5. Stage 3: Probability of contamination given non-perfect sensors

(Dawsey, et al., 2006) were the first to propose a Bayesian Belief Network (BBN) for combining evidence to better characterize contamination events using non-perfect sensor detections. The methodology uses distribution system simulations and conservative transport to estimate conditional prior probabilities for contaminant introductions. A BBN is developed that integrates sensor data with other information such as operation changes. More research should undertake to consider the full spatial and temporal characteristics of the sensor data and distribution system model.

Recently, (Perelman & Ostfeld, 2010) first simplify the hydraulic network to a directed acyclic graph (DAG). The DAG nodes are supernodes (clusters) of the initial network graph that reflect the connectivity and the flows the best. The Epanet software is used both for EPS hydraulic simulation and for the transport prediction of a conservative contaminant. Then, a Bayesian network is built with leaf nodes representing potential detection at a cluster, and a source node. The conditional probability table, given the source of contamination, is worked out based on the deterministic contaminant Epanet scenarios. As a result, several statistical inferences may be achieved.

(Wang & Harrison, 2013) use a Bayesian approach to address problems like multiple sources and stochastic behavior in nodal demands and sensor errors, uncertainty in multiple source characterization parameters (e.g., magnitude, start time, and location), and prior knowledge

1.6. Other methods

(Salomons & Ostfeld, 2011) used reverse hydraulic modeling to determine the potential contaminant sources with time. Consumers become sources, sources become consumers, and the flow quantities and directions are reversed. Following, a tracer is injected at the location and time of the detection and a water quality simulation is performed using the reversed flows for a duration defined by the user.

For sensor placement, (Tryby, et al., 2010) propose to select sensor placement at model nodes that will improve the inverse problem solving of Eq. (2). They formulate a single nonlinear integer programming to solve to maximize the eigenvalues of M'M over all possible monitoring designs. The singular value decomposition of M is not used but rather the simple operator trace and the Euclidian matrix norm to evaluate this single-objective.

3. Conclusions and SI-Roadmap for SMaRT-Online

The SI problem has been tackled by a large number of authors who used different approaches: Some of them are promising whereas others seem to be not suited for real-time applications. In particular, stochastic optimization techniques are supposed to be not applicable due to the need of large number of simulation runs that can not performed for the size of networks that are studied in SMaRT-Online^{WDN}. Non-uniqueness of the solution is an underlying and regularization methods based on prior information are mathematical techniques that cannot

guarantee that the calculated global solution also exactly identifies the real source location in the physical system. Also, the influence of prior information should be studied. Multisource contamination is also a possibility that should not be disregarded. Thererfore non uniqueness is not regarded as so much important and enumeration method will be preferred at first step. However exploration and classification of candidate solutions using additional sources of information are worth to be considered.
For the real-time application in SMaRT-Online^{WDN} the following roadmap is proposed:

- Application of two-stage method with a first enumeration step for calculating all candidate locations independent of the reaction kinetics of particular subtances. In the second exploration step a probability calculation for ranking of the candidate solutions is proposed as in (Ung, et al., 2013).The uniqueness of the solution is not imperatively required.
- Consideration of water demands and input flow of contaminant as uncertainties.
- Consideration of binary sensor signal since the contaminant and its reaction kinetics are not known.
- Extensive use of simplification methods:
  - calculation of non observable network parts
  - aggregation of input scenarios that lead to identical sensor alarms (example: combination of pipes in series without intermediate sensor)
- Subsequent online update of I/O-matrix instead of recalculation in case of an alarm.

The selection of the method that will be finally used is not part of this document and is described in deliverable 6.2.

4. References

Baranowski, T. M. & LeBoeuf, E. J., 2006. Consequence Management Optimization fro Contaminant Detection and Isolation. *Journal of Water Resources Planning and Management*, July/August, 132(4), pp. 274 - 282.

Baranowski, T. M. & LeBoeuf, P. E., 2008. Consequence Management Utilizing Optimization. *Journal of Water Resources Planning and Management*, July/August, 134(4), pp. 386 - 394.

Boccelli, D. L. et al., 1998. Optimal Scheduling of Booster Disinfection in Water Distribution Systems. *Journal of Water Resources Planning and Management*, March/April, Band 2, pp. 99 - 111.

Constans, S., Brémond, B. & Morel, P., 2003. Simulation and Control of Chlorine Levels in Water Distribution Networks. *Journal of Water Resourcs Planning and Management*, March/April, 129(2), pp. 135 - 145.

Dawsey, W. J., Minsker, B. S. & VanBlaricum, V. L., 2006. Bayesian Belief Networks to Integrate Monitoring Evidence of Water Distribution System Contamination. *Journal of Water Resources Planning and Management*, July/August, Band 4, pp. 234 - 241.

De Sanctis, A. E., Shang, F. & Uber, J. G., 2010. Real-Time Identification of Possible Contamination Sources Using Network Backtracking Methods. *Journal of Water Resources Planning and Management*, July/August, Band 4, pp. 444-453.

Guan, J., Aral, M. M., Maslia, M. L. & Grayman, W. M., 2006. Identification of Contaminat Sources in Water Distribution Systems Using Simulation-OPtimization Method: Case Study. *Journal of Water Resources Planning and Management*, July/August, 132(4), pp. 252 - 262.

Laird, C. D., Biegler, L. T., Bartlett, R. A. & van Bloemen Waanders, B. G., 2005. Contamination Source Determination for Water Networks. *Journal of Water Resources Planning and Management*, March/April, 131(2), pp. 125-134.

Laird, C. D., Biegler, L. T. & van Bloemen Waanders, B. G., 2006. Mixed-Integer Approach for Obtaining Unique Solutions in Source Inversion of Water Networks. *Journal of Water Resources Planning and Management*, July/August, 132(4), pp. 242-251.

Liu, L., Ranjithan, S. R. & Mahinthakumar, G., 2011. Contaminat Source Identification in Water Distribution Systems Using an Adaptive Dynamic Optimization Procedure. *Journal of Water Ressources Planning and Management*, March/April, Band 2, pp. 183 - 192.

Perelman, L. & Ostfeld, A., 2010. *Bayesian Networks for Estimating Contaminant Source and Propagation in a Water Distribution System Using Cluster Structure.* Tucson, ASCE Water Distribution Systems Analysis, pp. 426 - 435.

Preis, A. & Ostfeld, A., 2006. Contamination Source Identification in Water Systems: A Hybrid Model Trees-Linear Programming Scheme. *Journal of Water Resources Planning and Management*, July/August, 132(4), pp. 263 - 273.

Preis, A. & Ostfeld, A., 2008. Genetic Algorithm for source characterization using imperfect sensors. *Civil Engineering and Environmental Systems*, March, 25(1), pp. 29 - 39.

Propato, M., Sarrazy, F. & Tryby, M., 2010. Linear Algebra and Minimum Relative Entropy to Investigate Contamination Events in Drinking Water Systems. *Journal of Water Resources Planning and Management*, 7/8, pp. 483-492.

Propato, M., Tryby, M. E. & Piller, O., 2007. *Linear Algebra Analysis for contaminat source identification in water distribution systems.* Tampa, Florida, ASCE, World Environmental Water Resources Congress, pp. 1 -10.

Propato, M. & Uber, J. G., 2004. Linear least-squares formulation for operation of booster disinfection systems.. *Journal of Water Resources Planning and Management*, January/February, 130(1), pp. 53 - 62.

Rossman, L. A., 2000. *EPANET2 users mannual.* [Online] Available at: http://nepis.epa.gov/Adobe/PDF/P1007WWU.pdf [Zugriff am 28 5 2013].

Salomons, E. & Ostfeld, A., 2011. *Identification of Possible Contamination Sources Using Reverse Hydraulic Simulation.* Tucson, Water Distribution System Analysis Symposium, ASCE, pp. 447 - 453.

Shang, F. & Uber, J. G., 2009. *EPANET Backtracking Extension (BTX) User's Manual (Version 1.0).* [Online] Available at: http://sourceforge.net/p/epanet-btx/code/HEAD/tree/trunk/doc/epanet-btx.pdf [Zugriff am 28 5 2013].

Shang, F., Uber, J. G. & Polycarpou, M. M., 2002. Practical back-tracking algorithm for water distribution system analysis. *Journal of Environmental Engineering*, 128(5), pp. 441 - 450.

Tryby, M. E., Propato, M. & Ranjithan, S. R., 2010. Monitoring Design for Source Identification in Water Distribution Systems. *Journal of Water Resources Planning and Management*, 136(6), pp. 637 - 646.

Ung, H., Piller, O., Gilbert, D. & Mortazavi, I., 2013. *Inverse Transport Method for Determination of Potential Contamination Sources with a Stochastic Framework.* World Environmental Water Resources Congress 2013: Showcasing the Future, ASCE, pp. 798 - 812.

Wang, H. & Harrison, K. W., 2013. Bayesian Update Method for Contaminat Source Characterization in Water Distribution Systems. *Journal of Water Resources Planning and Management*, 1/2, 139(1), pp. 1-13.