HAL
open science

SMaRT-OnlineWDN D6.2: Adaption, integration and extension of existing concepts for online source identification

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

To cite this version:

Olivier Piller, Jochen Deuerlein, Idel Montalvo Arango, Denis Gilbert, Hervé Ung. SMaRT-OnlineWDN D6.2: Adaption, integration and extension of existing concepts for online source identification. [Research Report] irstea. 2014, pp.19. hal-02599891

HAL Id: hal-02599891

https://hal.inrae.fr/hal-02599891

Submitted on 16 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Deliverable 6.2

Adaption, integration and extension of existing concepts for online source identification

Dissemination level: Public

WP6

ONLINE-SOURCE IDENTIFICATION AND RISK MANAGEMENT

30 September 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

D6.2 Adaption, integration and extension of existing concepts for online source identification

List of Deliverable 6.2 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	6.2		**Start date:**	01/02/2013
Contributors	Irstea	3S Consult	IOSB	
Person-months per partne	5	4	1	

Keywords
Inverse problem, contaminant source, backtracking, water quality model, simplification, sensitivity analysis

Objectives

To identify sources of contamination using all the information in an online context. The method should use reliable hydraulic information stored in a historical database. It should be able to treat subsequently released alarms in near real-time. For exploring solutions and acceleration of the identification process a dynamical graph simplification, both spatially and temporally, should be tested. Finally, a sensitivity analysis should be undertaken to explore the influence of the uncertainty in the demand distribution.

Summary

Water Distribution Networks (WDNs) are critical infrastructures that are exposed to deliberate or accidental chemical, biological or radioactive contamination. The project *SMaRT-Online*^{WDN} aims to develop methods and software solutions 1) to detect contamination from non-specific sensors, 2) to maintain an online water quantity and water quality model that is reliable and 3) to use the past model predictions to backtrack the potential sources of contaminations. This deliverable aims to adapt, integrate and extend existing source identification concepts in the online context.

From the deliverable 6.1 and the analysis of existing methods, it was decided that a two-stage method should be adapted for online context. First, an enumeration step for calculating all candidate locations is undertaken. The reaction kinetics of particular substances is not considered. Secondly, an exploration algorithm calculates probabilities for ranking of the candidate solutions. The first step method should be adapted to use more reliable recent hydraulic information stored in a historical database. It is able to treat subsequently released alarms in near real-time. This deliverable details how this method is adapted in the online context.

The adaptation consists of acceleration of the first step by an adjoint method and a development of multi-stage refinement to reuse the solution from the previous times. It is developed in the Sir 3S solution with sliding windows for storing past velocity data. The source identification problem in an online context may use new responses (every 5 minutes) from sensors to improve the potential source solution with successive problem runs. As a consequence, it is adaptive and multistage. Then, for exploring solutions and accelerating the identification process, a dynamical graph simplification, both spatially and temporally, is suggested. The time simplification aggregates all the time steps to retain only the potential candidates. The graph simplification consists of applying tree/core decomposition to select best locations, representative of any contamination spreading. Finally, a sensitivity analysis is proposed to explore the influence of the uncertainty in the demand distribution.

List of Figures:

Contents:

1 The two-step enumeration/exploration method

Here we detail the principals of the two-step methods for finding the potential contaminant source nodes.

1.1 First enumeration step by reverse time solving of the transport equation

A transport model combining transport through pipes then (perfect or incomplete) mixing at nodes is developed in this project. Many contaminant components (in this project biological or chemical) may be waterborne and the transport mechanisms are multiple and complex. An alarm alert is raised based on the analysis of abrupt changes in normal water quality measurements and training from historical databases as expert knowledge. The identification of the product may take several hours and we may want to localise the source of contamination first. For this it is impossible to consider all the phenomena (reaction, dispersion, diffusion) as they are solution specific. A conservative hypothesis is to assume that the contamination is non-reactive and is transported by plug-flow. This is the worst-case scenario considering the contaminant concentration but of course this simplification may lead to significant uncertainty that we will have to estimate.

The advection equation through pipes is:

$$\frac{\partial C}{\partial t} + U(t)\frac{\partial C}{\partial x} = 0, \tag{0}$$

Where C is the solute concentration and U the pipe water speed, and perfect mixing is used at junction nodes or imperfect mixing at double-Tee and cross junctions depending on the flow conditions.

One solution method for backtracking positive or negative responses from sensors to potential sources is to solve Eq. (0) by a method of characteristics (MOC) that uses transit time to backtrack the solution through the network graph. It is also possible (cf Neupauer, 2011) to solve the following adjoint transport equation in reverse time with appropriate Boundary Conditions and Initial Conditions:

$$\frac{\partial \psi}{\partial \tau} - U(\tau)\frac{\partial \psi}{\partial x} = 0, \tag{0}$$

Where ψ is the adjoint state, and $\tau = t_f - t$ is the backward time. The advantage of the latter is that the transport direct solver for Eq. (0) may be used with no change (the sensor becomes the sources, the velocity is reversed). In that case, ψ corresponds to the sensitivity of the concentration to a source release. Other phenomenon like the hydrodynamic dispersion may be added to the adjoint method. It is also possible to consider imperfect mixing as well.

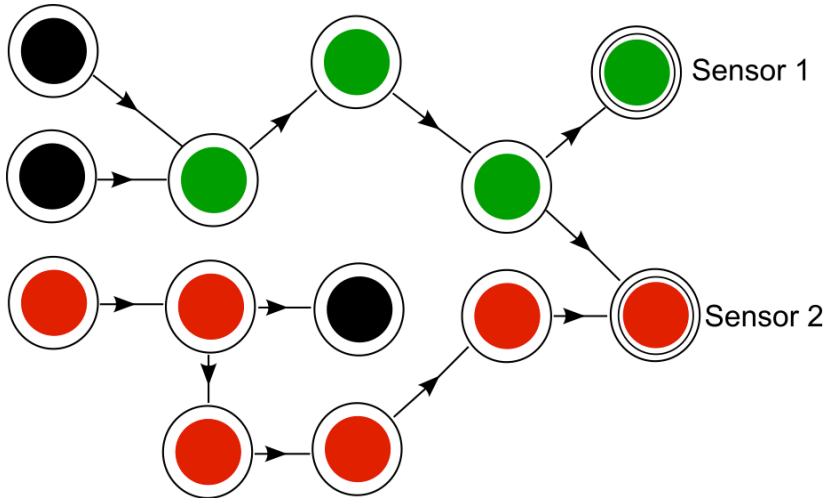

Figure 1: Concept of inverse transport on both positive and negative responses of sensors

As illustrated in Figure 1, the backtracking algorithm (MOC or solving the adjoint state) uses the fact that sensors have either positive or negative responses to investigate the possibility for the nodes to be a source of contamination. For this example, only 6 nodes (coloured in red) are potential contaminations. The backtracking time duration is 3 hours which is insufficient to go back up to the 2 nodes (North-west in black). Besides, 4 nodes are coloured in green because of the backtrack of the negative response of sensor #1.

Additional to the enumeration of the potential candidate for contamination, the backtracking algorithm also explicates the relationships between this potential candidates at different times with each sensor at the different times. These relationships are gathered in an input/output matrix.

1.2 Construction of the I/O matrix

The I/O matrix is constructed by the backtracking algorithm. Each line corresponds to a sensor at a time step and each column corresponds to a potential node at a precise time step. Figure 2 recapitulates the different steps of the algorithm.

Firstly, the alarms are generated corresponding to a scenario of contamination. Then, for each alarm, sensor and time a recursive function is used to do a backtracking on all inflow pipes and find the list of possible contamination nodes. Finally, all results are gathered in the I/O matrix as well as the column and line lists of corresponding nodes and times.

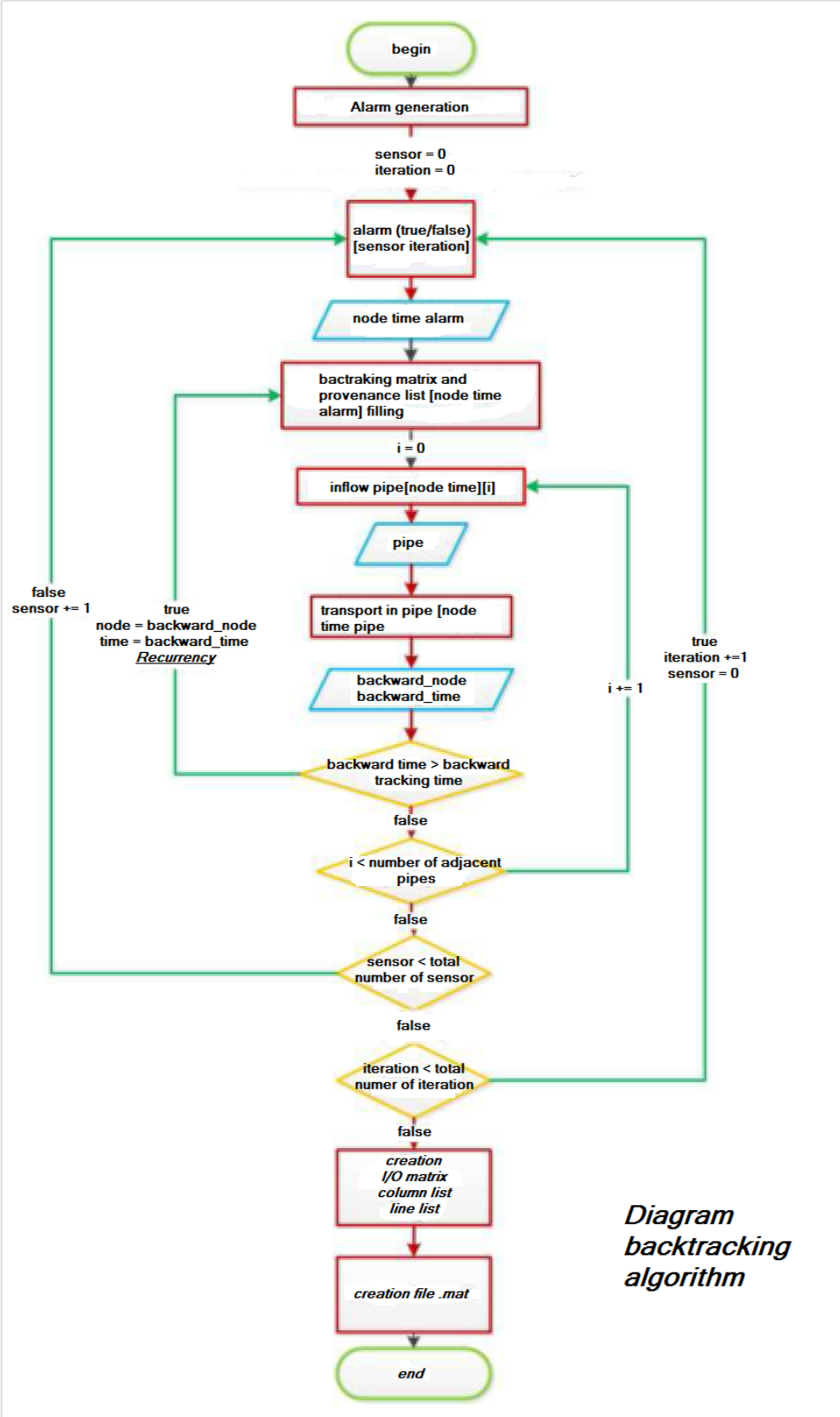

Figure 2: Flowchart for recursive assembling of the I/O matrix

The I/O matrix (hereafter referred to as A) is only composed of 0 and 1 and meets the following equations:

$$\mathbf{AU} = \mathbf{1}_M \tag{0}$$

With $(A)_{ij} \in \{0,1\}$ and therefore $U \in \Omega := [0,1]^N$

This equation is used for the ranking of the potential nodes of contamination developed in the next part. An alternative, may be employed using the results of the adjoint method:

$$\mathrm{U} = \mathrm{U}_0 \tag{0}$$

1.3 Exploration by using two stochastic optimisations

Two methods are used for the ranking of the potential sources of contamination: 1) the minimum relative entropy and 2) the least-squares method associated with the Tikhonov regularization.

1.3.1 The minimum relative entropy (MRE) method

The MRE method consists of minimizing the relative entropy under constraints. The concept comes from information theory and extends the discrete entropy defined by Shannon (1948). The relative entropy, also known as the Kullback-Leibler divergence, is a non-symmetric measure of the difference between two probability density functions (pdf) p and q. It gives the loss of information when p is approximated as q. The MRE problem reads:

$$\begin{aligned} &\min \mathrm{H}_e(q) := \int_\Omega q(\mathbf{U}) \ln\left(\frac{q(\mathbf{U})}{p(\mathbf{U})}\right) d\mathbf{U} \\ &\text{subject to:} \\ &\int_\Omega q(\mathbf{U}) d\mathbf{U} = 1 \\ &\mathbf{A}\,\mathrm{E}_{\mathbf{U}}(q) = \mathbf{1}_m, \;\; \mathrm{E}_{\mathbf{U}}(q) = \int_\Omega q(\mathbf{U})\mathbf{U} d\mathbf{U} \end{aligned} \tag{0}$$

Where H_e is the objective to minimize; $\mathbf{U}$ is the unknown contaminant intensity vector with $\mathbf{U}_i$ between 0 and 1; $\mathbf{U}$ is a random variable with joint pdf q; q is to be determined and p is a prior pdf, and $\mathrm{E}_{\mathbf{U}}(q)$ is the expectation of $\mathbf{U}$ with joint pdf q.

The objective function compares the unknown pdf q to the prior information p. The first constraint states the q joint pdf must integrate to one. The second constraint deals with the positive information given by the sensor responses. It takes the form of the expectation of $\mathbf{U}$ that should be solution of Eq. (0) or (0).

If the last constraint is not informative or empty: the expectation of $\mathbf{U}$ with pdf p is a feasible solution, and q equal to p is a global MRE solution with relative entropy zero ($\ln(1) = 0$). Eq. (0) may possess multiple solutions. The MRE solution chooses one of them to be the expectation of $\mathbf{U}$. The solution is the mean value for each potential node and its confidence interval for a given alpha level.

The problem Eq. (0) is resolved by a Lagrangian approach: the expression of q may be found dependent of the Lagrange variables μ and λ while differentiating the Lagrangian of the problem. We get the following explicit formulation:

$$\hat{q}(\mathbf{U},\mu,\lambda) = p(\mathbf{U})\exp(-1-\mu-\lambda^T\mathbf{AU}) \tag{0}$$

Where μ is the Lagrange multiplier of the q pdf integrity constraint; and $\boldsymbol{\lambda}$ is the Lagrange multiplier vector associated with the I/O matrix constraint (the information from the sensor responses).

The $\boldsymbol{\lambda}$ multipliers are then worked out as the solution of the dual problem:

$$\max G(\lambda) := -1-\mu(\lambda)-\lambda^T\mathbf{1}_M, \text{ with } \mu(\lambda) = \ln\left(\int_\Omega p(\mathbf{U})\exp(-\lambda^T\mathbf{AU})d\mathbf{U}\right)-1 \tag{0}$$

Where G is the dual function (strictly concave), and $\mu(\boldsymbol{\lambda})$ is fixed by the integrity constraint with q. This problem is not constrained. This is an advantage to calculate the optimum $\boldsymbol{\lambda}$ multiplier as a zero of the gradient of G, this gradient being $AE_U\left(\hat{q}(\mu(\lambda),\lambda)\right)-1_M$.

The $\boldsymbol{\lambda}$ solution with maximization of the dual problem permits the calculation of the optimal q Eq. (0) and $\mathbf{E_U}$ the expectation of $\mathbf{U}$ with q is deduced; it is the average value of all values that $\mathbf{U}$ can take.

1.3.2 The Least-Squares (LS) method

Tikhonov regularization is one common method to solve ill-posed Least-squares problems. It has been used by Laird *et al.* (2006) to convexify the LS problem. The suitable LS formulation reads:

$$\min c(\mathbf{U}) := \tfrac{1}{2}\left\|\mathbf{AU}-\mathbf{1}_M\right\|_2^2 + \alpha\left\|\mathbf{U}-\mathbf{U}_0\right\|_2^2 \tag{0}$$

Where $\mathbf{U}_0$ is a prior estimate; α is a Tikhonov regularization coefficient; and the second term corresponds to the regularization. This problem may be constrained with a positivity integrity constraint or with bound constraints.

As the criterion c is strongly convex for sufficiently large α, the problem (0) still has a unique real solution.

LS problem Eq. (0) can be solved by a Gauss-Newton algorithm by choosing the initial variable equal to the prior solution. The second term is weighted by the regularization coefficient chosen as small as possible but large enough for the problem to have a unique solution. The optimal α can be tuned with the L-curve method testing an increasing sequence of α. shows the values taken by the norm of $\mathbf{u}$-$\mathbf{u}_0$ versus the norm of $\mathbf{Au}$-$\mathbf{1_M}$ for different values of α. The larger the value of α, the more the regularization term gains importance and the norm of $\mathbf{u}$-$\mathbf{u}_0$ tends to zero. The aim is to find α such that the norm of $\mathbf{Au}$-$\mathbf{1_M}$ is the closest to zero.

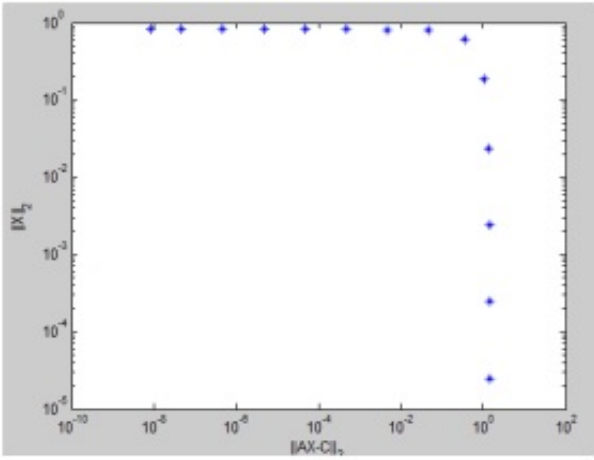

Figure 3: Example of L-curve for the LS weighted solution of the potential contaminations

1.4 Adaptive multistage problem

The source identification method by enumeration/exploration, as presented in the 3 first paragraphs, may be adapted to real time monitoring. Every time step, new information comes online and this can help to refine the solution given by preceding calculations. The backtracking algorithm Eq. (0) may be used to not rerun the overall calculation. Additionally, the optimization method may take into account preceding solutions (as additional constraints for example).

The recursive algorithm for assembling the I/O matrix (Figure 2) may handle the insertion of new information. For new positive responses, each time step of binary responses can be backtracked and added to the preceding matrix (new rows and columns). On the contrary, negative observation may lead to removing some potential solutions (some columns of the the I/O matrix).

For the exploration by MRE it is also possible to use the preceding solution by updating the prior joint pdf function and using the new input/output matrix. Alternatively for the LS formulation Eq. (0) both the I/O matrix A and the prior estimate may be updated.

2 Integration in the Sir 3S OPC server

For online application of the proposed method the flow velocities of the last few hours have to be known. There exists a trade off between calculation time, memory consumption and accuracy due to the following reasons:

The most efficient approach in terms of calculation time would be to store the calculated flow velocities for all calculated time steps and all pipes. However, the large amount of data prevents from saving these data in memory so that the flow vectors have to be stored in a database losing efficiency again. An alternative is to store only selected time steps either by constant time steps or dependent on the rate of change. It is easy to see that this approach results in loss of information and accuracy.

The third alternative is to recalculate the hydraulic state of the system for the last few hours (reconstruction calculation). At this stage it cannot be definitely decided which of the described method is the most efficient.

With the existing software (SirOPC Drive) reconstruction calculations are already possible and can be used without any additional changes. It is also planned to implement the database version where the pipe velocity vectors of each time step are stored in a kind of FIFO queue see Figure 4. That means that an array of flow vectors over time (size Δt) is created. The array has fixed size and for every new flow vector that is calculated the flow vector that refers to t -Δt is removed from the vector queue.

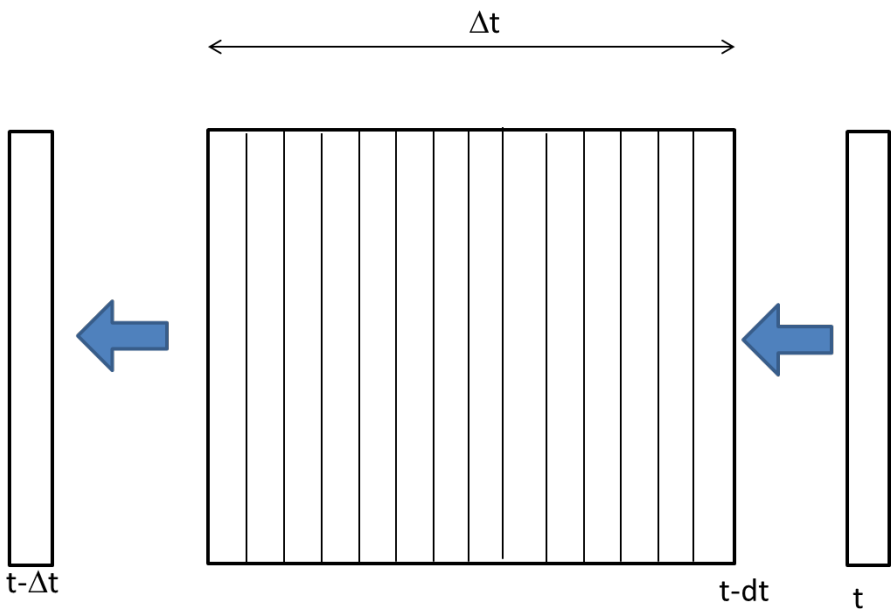

Figure 4: Sliding windows concept for pipe velocity storage in a FIFO queue

3 Graph simplification method

3.1 Temporal aggregation to the I/O matrix

To speed up the calculation before the optimisation method, it is possible to simplify the input/output matrix in time. Indeed the matrix constructed in the backtracking algorithm takes into account the sensor responses for each hydraulic time step, and the same applies to the potential sources. By gathering the different time steps together, by nodes, the matrix can be simplified considerably in size. Therefore the resolution of the optimization problem is quicker. In most cases this simplified calculation gives almost the same result as the complete one. The probability is averaged for each group of times and for each node. In a previous version, this method did not distinguish between positive sensor responses due by a same contamination which flows via different paths. To overcome the problem, a more complex form of simplification can be envisaged.

3.2 Appropriate graph simplification

From a topological point of view, the size of the problem can be reduced by subdividing the process into global and local steps. In the global step, source identification is carried out for the so called supergraph of the network. The supergraph is a topological minor of the original graph. Its nodes, the so called super nodes, are a real subset of the set of nodes of the original graph whereas the superlinks (being the links of the supergraph) consist of a series of pipes without bifurcation. In order to get the strongest simplification, the original network graph is first subdivided into forest and core and the topological minor is calculated for the core only. Figure 5 shows the two steps of forest-core decomposition and topological minor for an artificial example network graph. It is worth noting here that without the first step of forest-core-decomposition the example network has no topological minor and could not be simplified. In the following, the idea is described how the simplification can be used for solving the source identification problem.

Since there is no mixing within a superlink, the method can be subdivided into local and global parts. The global part consists of an implementation of the backtracking algorithm for the supergraph. Once the candidate region of the supergraph is known, the local solving process aims at improving the solution for the candidate region with consideration of all the details inside the superlinks (connected trees, demands, …). The local part is simple since there is no bifurcation and each point (or node) has a well-defined predecessor. In most cases it can be assumed that the sensors are located at supernodes only. As a consequence there is no additional information for the local part and the simplified global version of the algorithm delivers identical results as the full version.

However, it must be mentioned that there is no unique flow velocity within a superlink. For locally exact solutions, the different flow velocities have to be managed. After the implementation it will be proved whether the simplification including the overhead for flow velocity management in this case is more efficient than just applying the algorithm to the full system. As explained in the previous section, the dominated paths can be neglected and the PBA should be fast also for large systems.

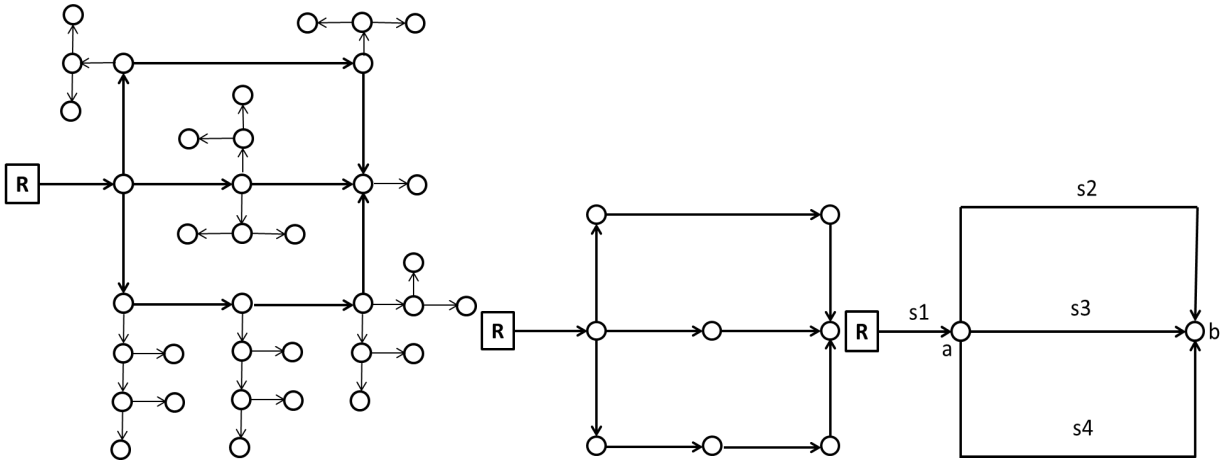

Figure 5. Graph simplification: a) original graph, b) core, c) topological minor

3.3 Consideration of shortest paths

For the creation of the I/O matrix it is not necessary to backtrack all particles until they reach their source. We are not interested in exploring all possible paths but in the path that has the shortest travel time. All the other paths are dominated by this path. That means that a sensor alarm, assuming perfect sensors could be released only by a particle that started from a source on that path and took the fasted way to reach the sensor.

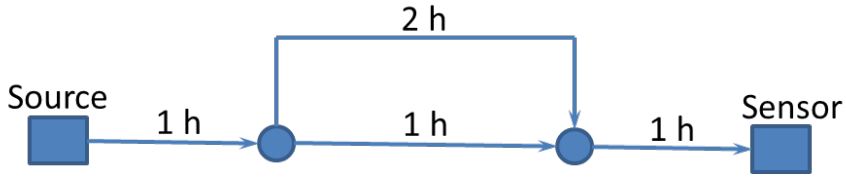

Figure 6: Dominated path

Figure 6 shows an example where two possible paths exist for the particle released by the source to reach the sensor. The first path takes 3 hours, in the alternative path the travel time is 4 hours. If we assume that a significant contaminant concentration passes the first bifurcation node then it follows that a part of the toxic matter takes the faster path and reaches the sensor after 3 hours. Consequently, it is sufficient for the Input-Output model to work with a simplified tree-link structure as shown in Figure 7.

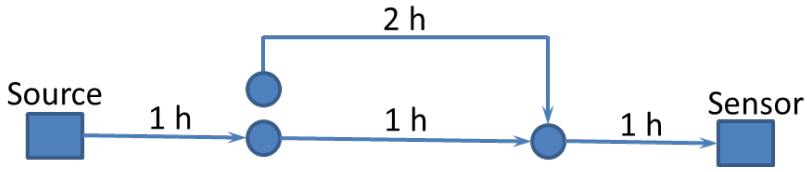

Figure 7: Modified network

Each sensor can be considered as root of a tree whose leaves and bifurcation nodes are the possible sources. If we have more than one sensor alarm the possible sources lay in the intersection of all trees in space and time. The same approach can be used for sensors with negative alarms. In this case the negative signal is traced back. All the node time pairs of the tree that are reached in this case are removed from the list of candidates.

4 Sensitivity analysis

The parameter uncertainty is a big issue and a source of error. In particular, the nodal demand uncertainty and the valve statuses may create large disparity in water distribution. Van Thienen *et al.* (2013) demonstrated that in some parts of an urban network, the stochastic nature of water demand may result in water taking different routes at the same hour of different days. Therefore, backtracking calculations should consider the parameter uncertainty. Here below, we propose an upgrade of the Monte Carlo method that Irstea proposed for the SecurEau project (2014).

4.1 Generation of the sensor responses:

Figure 8 hereafter explains the concept of the generation method. The aim is to generate different plausible sensor responses to different contamination events and different extended period simulation scenarios.

The initialisation step consists of reading the model data, the sensor locations and a contaminant event. The latter is entirely determined by the knowledge of the single or multi-source of contamination, and by the intensity, time and duration of the contaminations.

Then, a Monte Carlo process is executed for a sufficient number of demand scenarios (*e.g.* 100). At each iteration, a new hydraulic state for the entire period is determined by first drawing nodal demand time series, then running the hydraulic model and finally storing the velocity time series at consumer nodes. Secondly, the transport model for the contaminant event is solved. The concentration time series at the sensor set are stored for the benefit of the benchmark. This will stop when the maximum number of scenarios is reached.

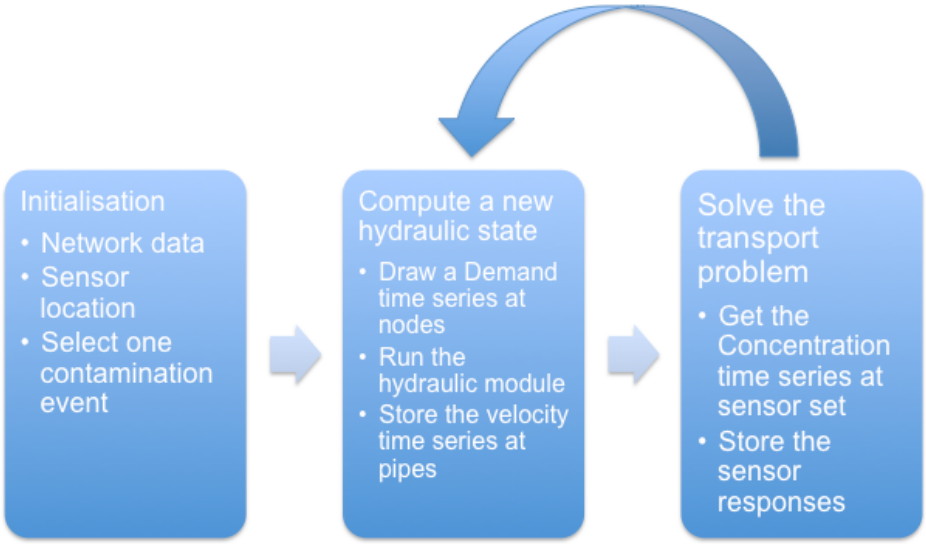

Figure 8: Concept for generating different sensor responses under demand uncertainty

At this step, several plausible sensor responses have been generated for the same contaminant event. Only, the demand variability impacts on the time and possibly on the location of the detection.
The method is explained in Figure 9. The outer loop corresponds to the repetitions (here around 100). The inner loop explores all the demand models defined for a network in the ideal data file.

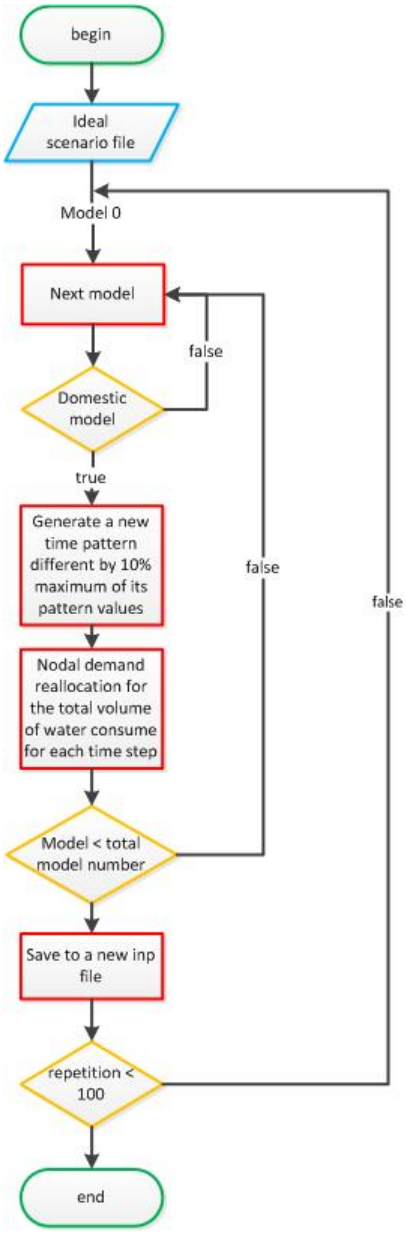

Figure 9: Flowchart of the demand scenario generation code.

A single iteration of the Monte Carlo process is described below:
We consider all the domestic demand models (classes) one after the other. Taking the ideal scenario file, the domestic demand patterns are modified. Each time coefficient of the pattern is drawn by the function rand() inside a 10% interval around its initial value. Then to keep the same volume of water consumed throughout the day, every value is multiplied by a coefficient that is the fraction between the initial and modified integral under the curve. The new domestic time pattern permits the determination of the total consumption for this model during each time step.

For a domestic pattern that concerns many nodes, a new demand repartition is made. Previously it was spread for each node considering its connections, but now each node has a random demand. Still these demands have to respect a criterion that at each time step the sum of the demands at the nodes has to be equal to the total consumption during this time step. To achieve this, at each time step a loop is made to distribute the total consumption. Iteratively, a small unitary demand (e.g., 0.005 l/s) is reallocated randomly to a node. This is repeated until the total

consumption is satisfied. At the end every node that is affected by this domestic model will have a new self time-pattern.

This operation is repeated for all the domestic demand models. Then, the nodal demand time series are stored in a new Porteau file or the Sir 3S database.

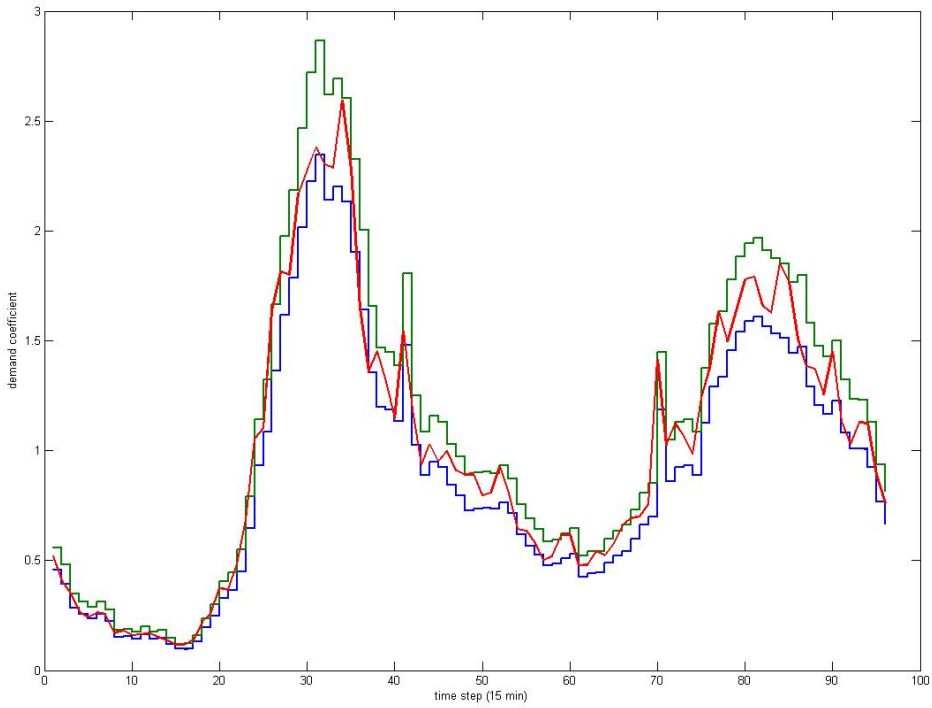

Figure 10: Determination of a time pattern realisation for a domestic demand model

Single or multiple contaminations are chosen to affect many nodes and hit several sensors. Then Figure 8, the water quality solver is used to build synthetic sensor responses for each of the contaminant scenarios.

4.2 Verification

Figure 11 below represents the variation of the demand at one node with 2 connections. The red square curve represents the maximum demand that is observed for each of the 100 simulations. The blue diamond one corresponds to the minimum values (here almost always zero). The 76th and the 100th simulation (repetition) are given as an example. It can be seen that both curves have the same general shape: consumption is high or low at the same hour. Nevertheless, there is a fluctuation for each time step and a zero consumption may be observed even at peak period.

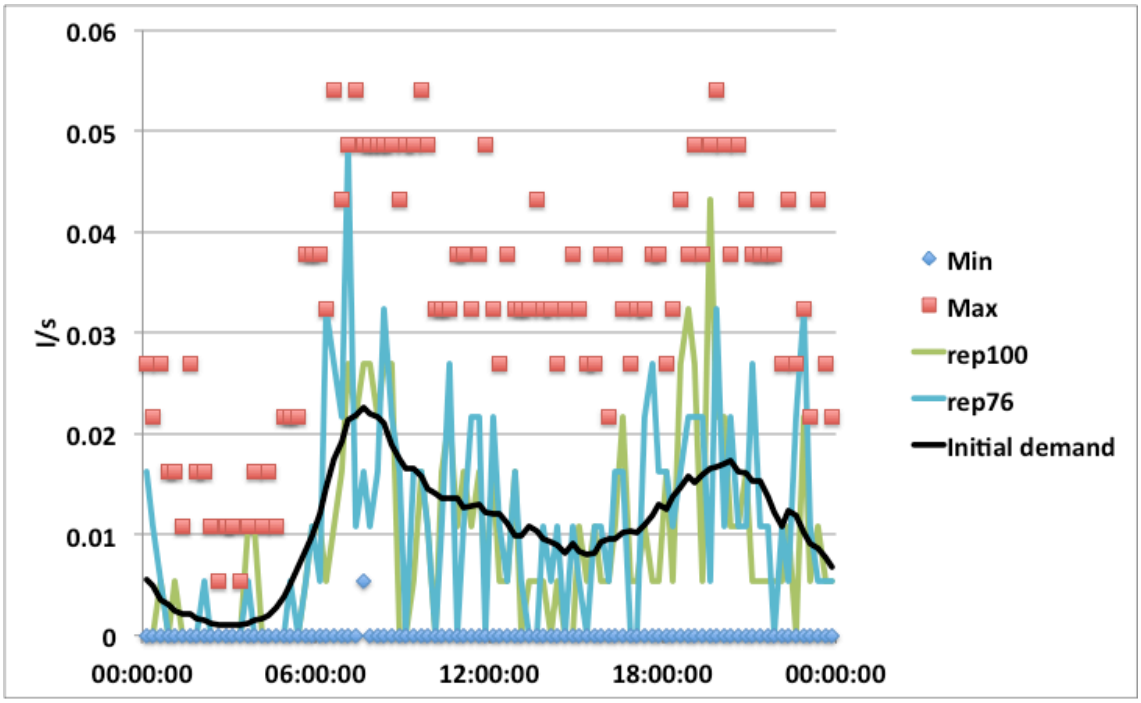

Figure 11: Example of variation for the nodal demand

4.3 Compliance analysis for the source identification given the demand uncertainty

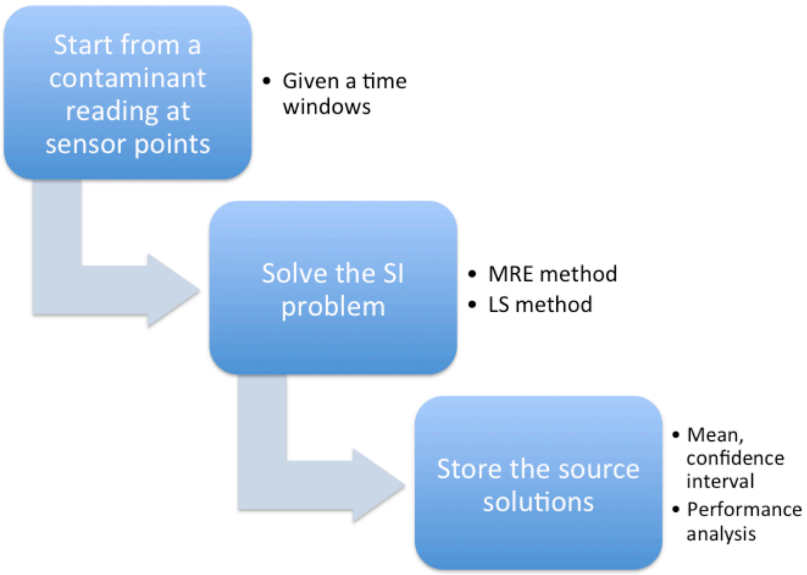

Figure 12: Source identification for the different sensor responses under demand uncertainty

For the same contaminant scenario and under demand uncertainty, we have potentially different response at the sensors. Here we simulate by a Monte Carlo process some of these responses, solve the SI problem (cf. Figure 12) and analyse the sampling distribution of the potential contaminant sources.

It is important to:

1. Verify if the method targets the real source.
2. Verify if there is consistency between the results achieved for the scenarios with and without uncertainties.
3. Quantify the number of possible contamination sources given by the method.
4. Quantify the time of computation necessary to achieve the results.

Several SI methods may be compared as well for their robustness and accuracy.

5 References

Neupauer, Roseanna (2011). "Adjoint Sensitivity Analysis of Contaminant Concentrations in Water Distribution Systems." *Journal of Engineering Mechanics*, 137(1), 31-39.

SecurEau (2014), http://www.secureau.eu/, last visited the 25 March 2014.

Van Thienen, Peter, Vries, Dirk, de Graaf, Bendert, van de Roer, Maurice, Schaap, Peter, and Zaadstra, Egbert (2013), “Probabilistic backtracking of drinking water contaminantion events in a stochastic word”, CCWI conference 2013, Perugia, 9 pages (to appeared in Proceedia Engineering).