Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 119 (2015) 360 – 369

Procedia Engineering

www.elsevier.com/locate/procedia

13th Computer Control for Water Industry Conference, CCWI 2015

A pre-screening procedure for pollution source identification in sewer systems

B.K. Banik[a]*, C. Di Cristo[a], A. Leopardi[a]

[a]*Dipartimento di Ingegneria Civile e Meccanica, Università degli Studi di Cassino e del Lazio Meridionale, Cassino 03043, FR (Italy)*

Abstract

Illicit intrusions in Sewer Systems (SSs), modifying the wastewater characteristics, may create problems to the treatment plant and/or to the final recipient water body. For this reason, the source identification (SI) problem is becoming an important issue also in SSs. For large real systems, the computational burden might make the SI methodologies impractical. In this paper a pre-screening procedure, based on the pollution matrix concept, is introduced and applied before the SI methodology. Selecting a group of possible candidate nodes and cutting consequently the scheme, a significant improvement both in terms of time and the accuracy is obtained.

Peer-review under responsibility of the Scientific Committee of CCWI 2015

Keywords: Source identification; Sewer systems; Pre-screening; Optimization problems.

1. Introduction

Nowadays, wastewater and drinking water systems are considered as the part of the critical infrastructure of a country. According to [1] improper management of wastewater facilities could result in loss of life; catastrophic damage to flora and fauna; and contamination of drinking water supplies. In addition, the consequences could result in long-term public health impacts, disruption to commerce and economy, which may lead to the overall disruption of the nation's way of life. Although in past the wastewater and storm water management have not got much attention to the researchers, a paradigm shift is evident evolving from simple sanitary and flood controls to overall environmental protection function. In many countries permit is mandatory for the operator to discharge their wastewater into the sewer systems (SSs). The security have been made a top priority in some country, like USA,

* Corresponding author. Tel.: +3907762994381; fax: +39 (0)776299 3939.
E-mail address: bk.banik@unicas.it

Peer-review under responsibility of the Scientific Committee of CCWI 2015
doi:10.1016/j.proeng.2015.08.896

where the SSs have been determined to be at risk to accidental and deliberate intrusion events. In this context, a very important aspect of the SSs management policy is to set up an early warning system for reliably identifying high-impact illicit intrusion events. With the new available methodologies, which permit on-line monitoring in SSs (e.g. [2], [3]), the development of these tools will provide the operational staffs to take corrective action to minimize the impact of such intrusions.

An important aspect of a good SSs management policy is represented by the identification of the input source and its characteristics. Source identification in water distribution system has got enormous attention to the researchers in the last decades. The same topic has not attained the same pace in the case of SSs mainly due to the lack of immediate and direct influence on public health. Few rare efforts have been reported in the literature. [4] and [5] have proposed methodologies to identify the illicit discharges in a separate storm drainage system, where the injection source is identified through sampling and analytical laboratory analysis. To the best of the authors' knowledge, the methodology proposed by [6] is the first effort made based on online sensor measurements to characterize illicit intrusion in a combined sewer system.

[6] have proposed a methodology for the identification of the source location, along with its characterization, represented through released input concentration, starting time and duration, in SSs. They have used the storm water management model (SWMM) to perform the hydraulics and water quality simulations and genetic algorithm for solving the source identification (SI) problem. The method ([6]) works fine with small networks as it takes very small amount of time to run a single simulation. But in real life the size of the network is much larger and so the needed computational effort will eventually make the method impractical. So, developing a methodology for solving the SI problem within a reasonable time is a challenging but important engineering problem.

This paper presents an original research of a newly developed pre-screening procedure to apply before the SI methodology. The proposed procedure uses the pollution matrix ([7]) concept to identify a set of 'Candidate' nodes. Once the possible candidates are identified the unnecessary elements of the network are removed by cutting some parts of the scheme. The flow coming from the cut parts are added to the next node of the remaining scheme, assigning a new average inflow and a new time pattern. The inflow is estimated from the last link of the cut portion, while the time pattern is estimated from the downstream link next to the node, where the value has to be assigned. In the successive SI optimization process only the 'Candidate' nodes are considered. So the computational time is reduced in two ways: i) reduced number of genetic parameters, being the search domain smaller; ii) less time require for an individual SWMM simulation, being the considered scheme smaller.

The paper is organized as follows. It starts with the formulation of the SI methodology, along with the presentation of the optimization procedure. The third section is devoted to the detailed description of the proposed pre-screening methodology. In the application and result section the methodology is tested on two sewer systems, with increasing complexity: a literature scheme from the SWMM manual and a real combined sewer. In the final section some concluding remarks and recommendations are drawn.

2. Source Identification (SI) methodology

The main goal of a SI methodology is to characterize the source of an illicit intrusion, when it is detected in a sanitary or combined sewer system. The SI methodology proposed by [6] is formulated as an optimization problem which uses sensor measurements through on-line monitoring. For its solution two main components are required: a model for hydraulic and quality simulations in sewer systems and an optimization problem solver. In particular, the USEPA's SWMM ([8]) is used to perform the hydraulic and water quality simulations, whereas the genetic algorithm library GALib ([9]) is used to solve the optimization problem.

In the present paper, a third component is introduced: a pre-screening procedure, in order to reduce the computational effort in case of complex schemes. The proposed SI methodology consists of two main parts, as depicted in Fig. 1:

- In the first part a group of candidate nodes, a subset of all nodes in the system, are selected through a ***pre-screening procedure***, which is explained in more details in section 3.
- In the second part, among the candidate nodes the approximate location of the intrusion point with its characteristics (input concentration, staring time and duration) is determined through a ***SWMM-GA optimization*** tool.

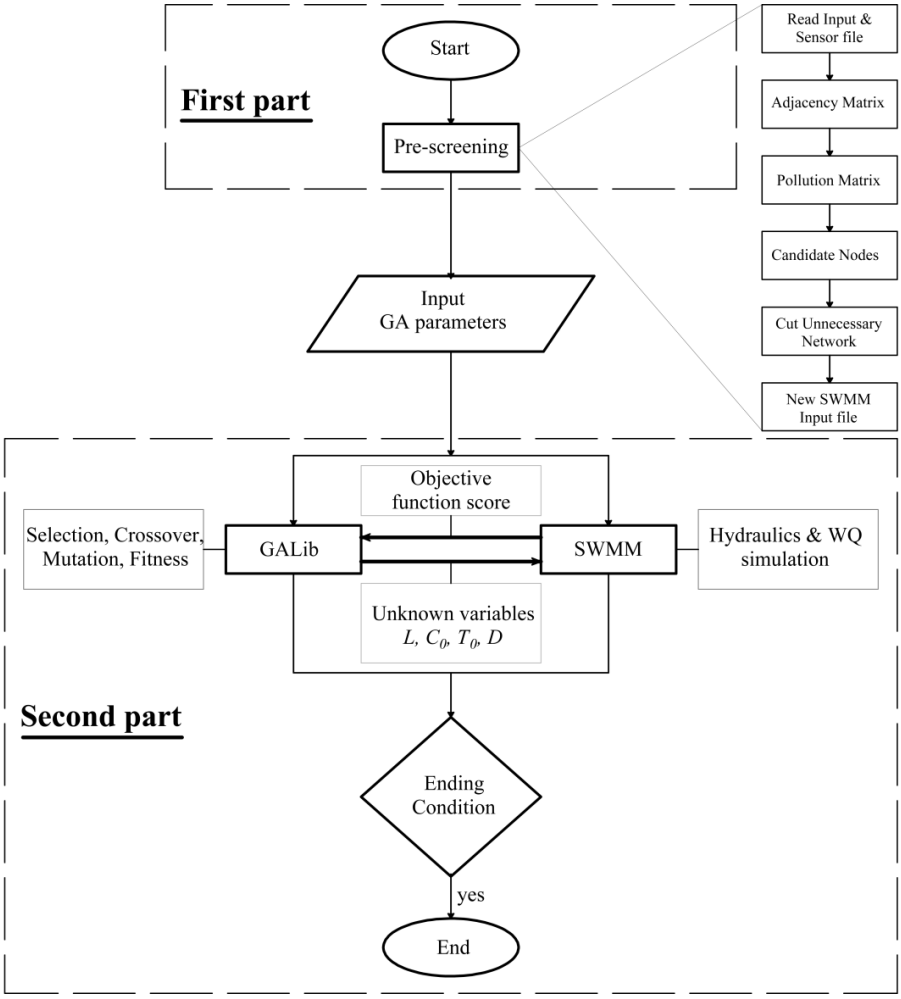

Fig. 1. Pre-screening procedure and SI methodology scheme

2.1. SI problem formulation

The SI problem is formulated minimizing a dimensionless objective function, F, defined as the normalized square difference between the simulated and the measured contaminant concentration values. The adopted mathematical form of F is:

$$F = \sum_{t=t_0}^{t_c} \sum_{i=1}^{N_s} \left(\frac{C_{it}^{obs} - C_{it}(L, C_0, T_0, D)}{(C_{it}^{obs} + C_{it}(L, C_0, T_0, D))/2} \right)^2 \quad (1)$$

where i = sensor index; t = time step; C_{it}^{obs} and C_{it} = measured and simulated concentration at sensor i at time step t, respectively; L = intrusion node index; C_0 = release concentration (mg/l); T_0 = release starting time; D = release duration (h); t_0 = time of the first detection of the contaminant at the sensor i; t_c = ending time of simulation; N_s = total number of sensors. The simulated concentrations are computed through a forward quality simulation, fixing the values of the unknown source characteristics, which are herein represented by the four decision variables: (1)

Pollution injection node, L; (2) Pollutant concentration at the source, C_0; (3) Injection starting time, T_0; and (4) Injection duration, D. Then, the optimal value of the objective function F (Eq.1) is obtained modifying the values of these variables. For solving the optimization problem the linked simulation-optimization procedure described in the next paragraph is used.

2.2. SWMM-GA simulation optimization procedure

As mentioned above, the USEPA's SWMM ([8]) is used to perform the hydraulic and water quality simulations, whereas the genetic algorithm library GALib ([9]) is used to solve the optimization problem. Genetic Algorithms (GAs) are already widely used in optimization problems related to water resources planning and management researches ([10]), thus many of their technical details are not reported herein. Briefly, GAs are a heuristic combinatorial search technique that mimics the natural evolution process of chromosomes. A typical GA, as the one incorporated in the methodology, starts with the generation of a random set of population, then the objective function score is evaluated for each individual in the population. Afterwards the algorithm generates a new set of population through selection, crossover and mutation. The above steps are repeated until a prescribed objective function threshold value is reached or if a predefined number of generations is attained.

In order to integrate the SWMM simulator with the proposed automated SI methodology an ad-hoc SWMM-TOOLKIT has been developed. As more detailed described in [6], the new SWMM.DLL used in this study has 22 additional functions for retrieving information about network nodes and time patterns, as well as for setting new values during the extended period simulation from a C++ platform.

3. The pre-screening procedure

The simulation-optimization procedure, based on a number of consecutive simulations, is computationally expensive and not easily applicable for large network. Using the measured concentrations, the pre-screening procedure selects a limited group of possible candidate nodes, excluding from the search domain the nodes that cannot be the location of the source. All the unnecessary nodes and their associated links are then cut from the scheme, in order to apply the SI methodology to a smaller system. In this way, due to the small number of candidate nodes, the simulation time is reduced because the GA parameters can be set on smaller values, which means a small number of runs. At the same time, being the ‘cut’ network smaller, the required time for each single simulation is less. The corresponding inflows coming from the cut portion of the network are successively integrated with the remaining network.

The proposed pre-screening procedure, based on the pollution matrix concept ([7], [11]), assumes that all wastewater passing through the intrusion point is considered contaminated. [11] have used this concept along with the water fraction matrix ([12]) one to select a group of candidate nodes before applying a pollution source location methodology in water distribution networks. The main difference of the presented pre-screening procedure with the one proposed by [11] are twofold. First of all in the current procedure the SWMM run is done only once, whereas for the previous procedure a total number of EPANET runs equals to the number of nodes in the network is required. Secondly, while in the previous method the network has the same size after pre-screening, in the current procedure the unnecessary part of the network is cut from the original scheme.

The pre-screening procedure, as schematically shown in Fig. 1 (top right corner), starts with reading the sensor input file, which contains the sensor node indices and the corresponding time series of concentration measurements. The following paragraphs describe the four main steps of the proposed pre-screening procedure.

3.1. Step1: Adjacency matrix

The first step of the pre-screening procedure consists of formulating the adjacency matrix, which is a $N \times N$ matrix having 0–1 coefficients, where N is the number of nodes, including tanks and reservoirs. The adjacency matrix (A) contains the information of whether the water directly travels from a particular node to another node (not via another node). If water travels directly from node i to node j then the coefficient of the element of the adjacency matrix is ‘1’, otherwise it is ‘0’. The pollution matrix is then formed from it.

3.2. Step2: Pollution matrix

Once the adjacency matrix (A) is formed, the pollution matrix can be constructed by using slightly modified Floyd-Warshall algorithm (Floyd 1962), which has been used extensively for finding the shortest path in a network. The pollution matrix is also a $N \times N$ matrix of 0–1 coefficients, where N is the number of nodes with '1' and '0' values, corresponding to polluted and unpolluted nodes, respectively. The i^{th} row of the pollution matrix represents the situation of the nodes in case of a pollution event in node i. The '1' values in the j^{th} column represents all the nodes that can pollute node j (domain of coverage of node j). The pollution matrix P is formed using the following algorithm:

```
P = A
for k = 1 to Number of nodes
       for i = 1 to Number of nodes
              for j = 1 to Number of nodes
                     {if ((Pik==1 and Pkj==1) or Pij==1)
                            Pij = 1}
```

Finally, all diagonal elements are assigned as '1'.

3.3. Step 3: Candidate nodes

Once a nonzero measurement is read from a monitoring station in a node, the nodes of its domain of coverage are probable candidates. The final group of candidate nodes will be formed from the intersection of the domain of coverage of all sensor nodes with a nonzero measurement. Conversely, all zero measurements in a sensor node means that the nodes of its domain of coverage have to be excluded from the group of candidates. Mathematically the array of the candidate nodes S can be expressed as

$$S_j = \prod_{k=1}^{M} P_{jk} \quad for\ j = 1\ to\ N \tag{2}$$

where N = number of nodes in the network; M = number of sensor nodes having at least one nonzero measurement; and k = node equipped with the k^{th} sensor. P_{jk} is the j^{th} row and k^{th} column element in the pollution matrix which is either '0' or '1'. If S is '1', the node j is assigned among the candidates. The number of candidate nodes will guide to fix the genetic algorithm parameters.

3.4. Step 4: Cut unnecessary network

Once the candidate nodes are known, the next task is to find which nodes are linked with them to all the ways to the outfall or treatment plant. They are connecting nodes and they cannot be cut. As shown in the example scheme reported in Fig. 2, the 'nodes to be cut' are then simply formulated by subtracting the connecting nodes from the total nodes. Those nodes and their associated links are cut from the original network, but the associated flow coming from that part needs to be integrated to the remaining network. Two other arrays namely 'last link of cut part' and 'next link' are created to compute the average inflow and the new hourly time pattern of the immediate next downstream node 'N' (Fig. 2), respectively.

The array 'last link of cut part' consists of all downstream links of each cut portion that were connected to the original network before the screening operation. The flow within those links is put into the corresponding next downstream node 'N'. At the same time the pattern factors at 'N' has to be changed. This new pattern factors are calculated based on the flow within the next link (Fig. 2) of node 'N'. Then the average inflow to add at node 'N' is calculated as follows:

$$DWF_{avg} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i^{last}}{\frac{1}{m}\sum_{j=1}^{m}(pf)_j^{next}} \tag{3}$$

where, n is the total number of flow data at last link; x_i^{last} is the i^{th} flow data in last link; m is the number of pattern factors during the simulation period; $(pf)_j^{next}$ is the new j^{th} pattern factor calculated from the next link.

Once the new inflows and the new time pattern are added to node '*N*' all the unnecessary part of the network is cut and a new input file is prepared for the next SWMM run.

3.5. Example of the pre-screening procedure

The entire pre-screening procedure is applied considering the simplified scheme in Fig. 2. The network is consists of eight nodes and seven links. Two sensors are placed at node 3 and seven 7 and both have at least one nonzero pollutant measurement. The adjacency matrix (A) and the corresponding pollution matrix (P) for this network are reported on the right side of Fig. 2.

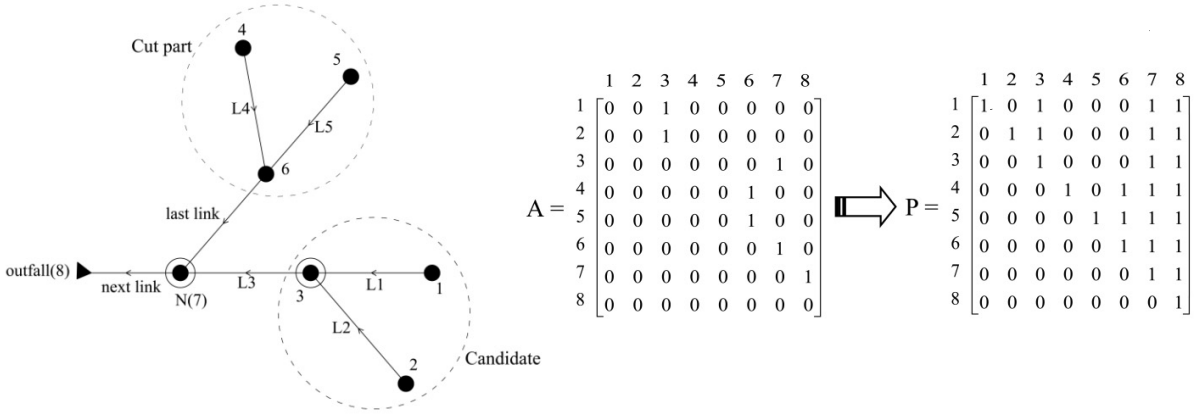

Fig. 2. Example scheme and application of the pre-screening procedure

The domains of coverage of two sensors are represented by the elements having value '1'of the corresponding columns in the pollution matrix. So, the domains of coverage of two sensors are:

3 → {1, 2, 3}

7 → {1, 2, 3, 4, 5, 6, 7}

The candidate nodes are furnished from the intersection of the domain of coverage among two sensors: {1, 2 and 3}.

Then, the connecting nodes are calculated by taking the union of the corresponding rows of the candidate nodes (i.e. rows 1, 2 and 3) in the pollution matrix. So, the connecting nodes are {1, 2, 3, 7 and 8} and the nodes to be cut are {4, 5 and 6}. The arrays of 'last link of cut part' and 'next link' are formed. For this particular case both array have only one element and they are {last link} and {next link} respectively. The average flow in 'last link', computed through Eq.(3), is added to the inflow of node N(7), while the flow in 'next link' is used to calculate the new time pattern factors of node N(7). Then all the unnecessary part (nodes in 'nodes to be cut' array and their corresponding links) is removed from the original network and a new input file will be ready for the next part of SI methodology.

4. Application and results

The applicability of the proposed methodology has been tested on two different schemes: the literature system Example 8 from the SWMM5 application manual ([13]) and the real sewer Massa Lubrense11, located near Naples, Italy. In both cases only dry weather flow (DWF) is considered and a conservative pollutant is assumed. A six hour simulation time starting from 8:00 AM is applied and the entire procedure is replicated 50 times to observe the success percentage in identifying the correct node. The GA parameters are chosen as: Population = 100, Generation = 40, Mutation Probability = 0.01 and Crossover Probability = 0.9. All the simulations have been performed on an Intel Core i7-4700MQ CPU @ 2.40 GHz processor, 20 GB RAM machine. It must be noted that five executables have been run simultaneously in the case of Massa Lubrense11 to save the overall simulation time. So, one replication will refer to a set of five parallel simulations for this case.

4.1. SWMM example 8 system

The SWMM Example 8 system is presented in Fig. 3. It is a combined sewer network, which serves an area of 0.12 km^2, and consists of 31 nodes (28 junctions, two outlets and a storage unit), 35 links (29 conduits, 1 pump, 1 orifice and 4 weirs). There are two outlets, the wastewater treatment plant (WWTP) and the outfall for discharging the combined sewer overflow, in the system. In Fig. 3(a) the lines with the C label, represents the natural stream, while the sewer pipes are reported with thick black lines, with the P label. The interceptor, indicated with thin black lines with the I label, are pipes designed to capture the sanitary flows during dry weather periods and convey them to the WWTP. The flow regulators (weirs and orifices) and the pump station are also depicted, along with the eight nodes in which the DWF are present.

Two monitoring stations are considered and arbitrarily placed at node JI3 and JI13 (Fig. 3). Synthetic measurement are generated through a SWMM quality simulation assuming a continuous three hours injection of a pollutant at node ‘J1’ with a concentration of 20 mg/l starting from 9 AM. The routing time step is 30 seconds and the reporting time step is 10 minutes.

4.2. Massa Lubrense11 system

The system Massa Lubrense11, presented in Fig. 4(a), is one of the 12 subcatchment of the real sewer system of Massa Lubrense, a town located near Naples, Italy. It is a combined sewer system, covering an area of 0.75 km^2. The scheme consists of 242 circular conduits connecting 241 junctions, one pump, two storage units and one outfall. The DWF distribution is also depicted in Fig. 4(a).

The ‘true’ pollution scenario is a three hours continuous injection of a pollutant at node ‘N16’ with a concentration of 20 mg/l starting from 9 AM. The system is analyzed with a routing time step of 10 seconds and a reporting time step of 10 minutes. Three monitoring stations are arbitrarily chosen at nodes N42, N241 and N219 (Fig. 4(a)), and the synthetic sensor measurements are generated through a SWMM quality simulation.

4.3. Results

In applying the pre-screening procedure two schemes are reduced as shown in Fig. 3(b) and 4(b) for the two considered test cases. While not a big difference is observed for the Example 8 scheme, (Fig.3(b)) the Massa Lubrense system is modified with the 69% reduction of nodes, after the application of pre-screening procedure (Fig. 4(b)). Table 1 reports the percentage of mean absolute error and the corresponding standard deviation in computing the flow in five links after the pre-screening procedure in the Massa Lubrense11 network. The reported values demonstrate that the errors introduced by cut are very small, without any significant influence on the simulation results. In fact, the highest mean absolute error in flow measurement is only 1.76% at link L5.

Table 2 reports a summary of the SI methodology results for the two considered test cases obtained with and without applying the pre-screening procedure. Column (2) and (6) report the success percentage in identifying the correct node, while column (3) and (7) show the estimated average input concentration. Column (4) and (8) present

the total simulation time consumed due to all SWMM runs associated in one replication, whereas column (5) and (9) indicate the time consumption due to the GA optimization in one replication. Using the pre-screening procedure described in this paper a significant reduction in computational effort, is observed, especially in the case of the large network. For the network Massa Lubrense11 a set of five complete SWMM-GA optimizations (one replication) has taken 2:34:31 hours, which compared with the 10:01:30 hours requirement without the pre-screening produces the 75% reduction in the overall computational effort.

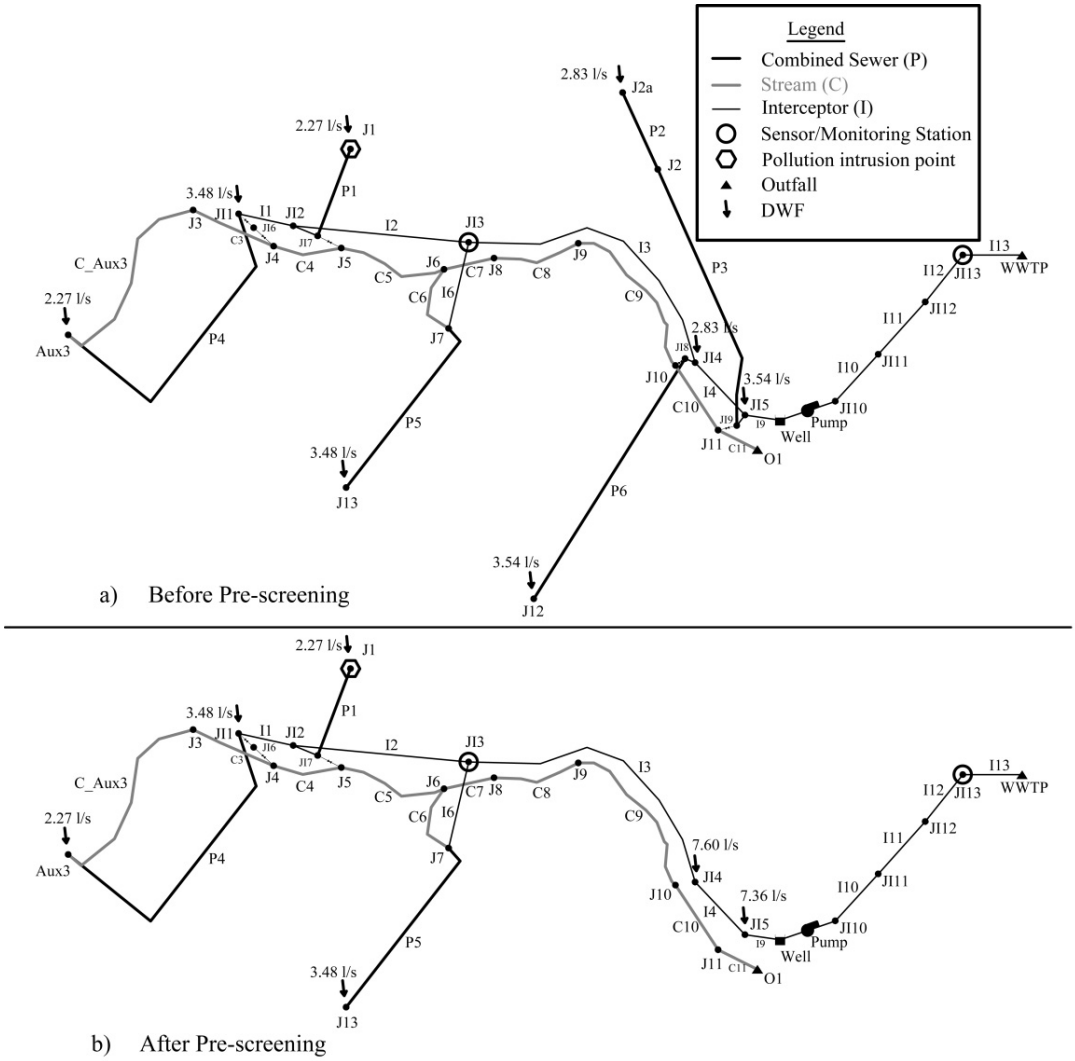

Fig. 3. Scheme of SWMM example 8 network (adopted from [14])

It is interesting to note, from Table 2, that the success percentage remains same after applying the pre-screening procedure for the SWMM example, whereas an improvement is observed for the system Massa Lubrense11. This change is due to the reduction of the search domain used for the optimization procedure. The improvement is also noticed in estimating the source concentration for the case of Massa Lubrense11. A more detailed study on this methodology along with sensitivity analysis on GA parameters as well as the uncertainty analysis on inflows and sensor measurement can be found in [14].

Table 1. Error associated with link flows in Massa Lubrense11 network due to the pre-screening procedure

Link	L1	L2	L3	L4	L5
Mean absolute error (%)	0.51	0.66	1.05	1.23	1.76
Standard deviation (%)	0.72	0.83	0.87	0.99	0.76

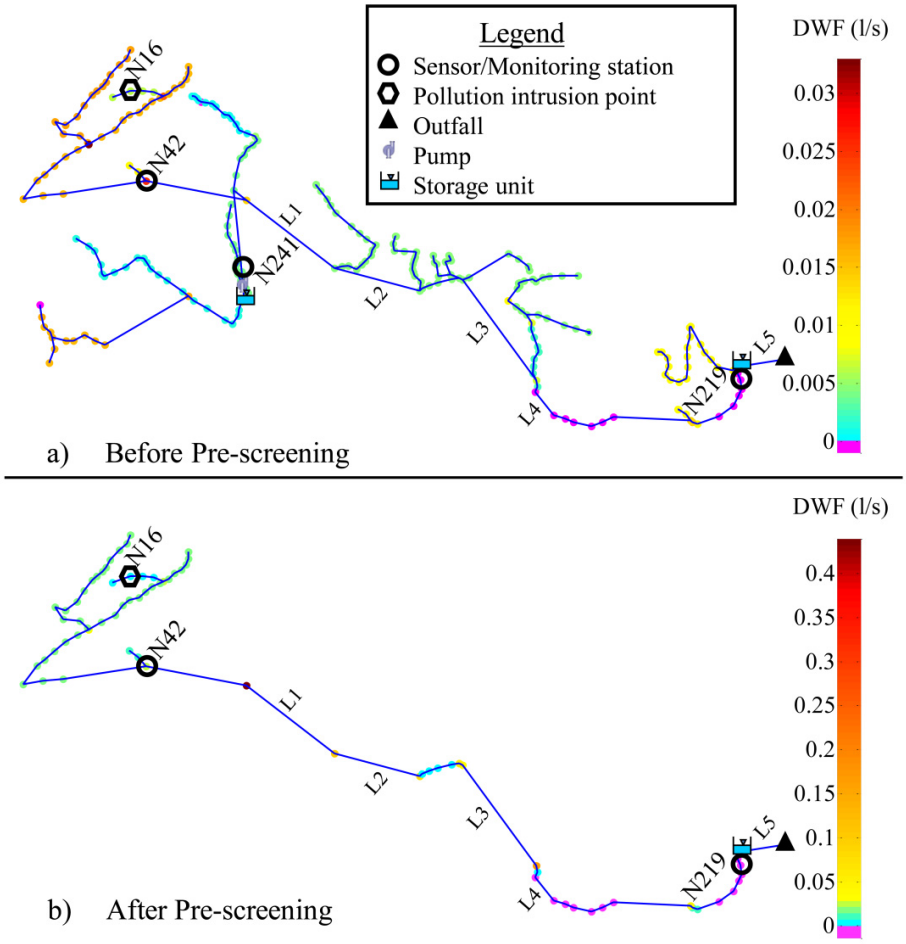

Fig. 4. Scheme of real sewer network Massa Lubrense11 (adopted from [14])

Table 2. SI result for the two test cases with and without considering pre-screening

Network	Without pre-screening				With pre-screening				Time reduction (%)		
	Success (%)	Conc. (mg/l)	Simulation Time SWMM	Simulation Time GA opti.	Success (%)	Conc. (mg/l)	Simulation Time SWMM	Simulation Time GA opti.	SWMM	GA opti.	Overall
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
SWMM Ex. 8	96	19.75	00:06:08	00:00:17	96	19.59	00:05:01	00:00:15	19	8	18
Massa Lubrense11	66	20.59	09:43:14	00:18:16	**86**	19.86	02:32:18	00:02:13	74	89	**75**

5. Conclusions

Computational burden is a big issue in applying a Source Identification (SI) methodology to sewer systems. A pre-screening procedure, based on pollution matrix concept, has been developed in order to apply it before a previously proposed SI methodology in which the SI is expressed as an optimization problem. Starting from sensor measurements, the pre-screening procedure selects a group of candidate nodes, which constitute the search domain in the optimization procedure, removing all the unnecessary parts of the scheme and producing a significantly small network. The flow associated with the removed part is integrated with the remaining one. The SI methodology performance with the pre-screening procedure is tested on two different schemes: a small literature system and a real large sewer. It is observed that the introduction of the pre-screening produces a significant improvement both in terms of time and the success percentage in identifying the correct location, especially for the large network.

The position of the sensors could be a vital issue for the effectiveness of the pre-screening procedure. In the present study the sensor placement in both networks is done randomly and no optimality test has been performed in terms of their locations. A recent study can be found in [15] for the optimal placement of the monitoring stations in a sewer network.

Acknowledgements

The first author would like to thank the EU for the financial support through the Erasmus Mundus Joint Doctorate Programme ETeCoS^3 (Environmental Technologies for Contaminated Solids, Soils and Sediments), grant agreement FPA n° 2010-0009.

The experimental data on the Massa Lubrense pilot site were collected from the SIMonA Project, financed by the Campania Region in the framework of the Campus Campania Programme (POR FESR Campania 2007-2013).

References

[1] R. Bahadur, W.B. Samuels, Wastewater Critical Infrastructure Security and Protection, in: R.M. Clark et al. (eds.), Handbook of Water and Wastewater Systems Protection, Springer, New York, 2011, pp. 87–102.

[2] A. Llopart-Mascaró, A. Gil, J. Cros, F. Alarcón, Guidelines for on-line monitoring of wastewater and stormwater quality, 11th international conference on urban drainage, 2008, Edinburgh, Scotland, UK [CD-ROM].

[3] X. Qin, F. Gao, G. Chen, Wastewater quality monitoring system using sensor fusion and machine learning techniques, Water research. 46 (2012) 1133–1144.

[4] R. Field, R. Pitt, M. Lalor, M. Brown, W. Vilkelis, E. Phackston, Investigation of dry-weather pollutant entries into storm-drainage systems, Journal of Environmental Engineering, ASCE. 120 (1994) 1044–1066.

[5] K. Irvine, M.C. Rossi, S. Vermette, J. Bakert, K. Kleinfelder, Illicit discharge detection and elimination: Low cost options for source identification and track down in stormwater systems, Urban Water J. 8 (2011) 379–395.

[6] B.K. Banik, C. Di Cristo, A. Leopardi, SWMM5 toolkit development for pollution source identification in sewer systems, Procedia Engineering. 89 (2014) 750–757.

[7] A. Kessler, A. Ostfeld, G. Sinai, Detecting accidental contaminations in municipal water networks, J. Water Resour Plan and Manage. 124 (1998) 192–198.

[8] L.A. Rossman, Storm water management model user's manual. Version 5.0. National Risk Management Research Laboratory, Office of Research and Development: US Environmental Protection Agency, 2010.

[9] M. Wall, GAlib: A C++ library of genetic algorithm components. Mechanical Engineering Department: Massachusetts Institute of Technology, 1996.

[10] J. Nicklow, et al., State of the art for genetic algorithms and beyond in water resources planning and management. Journal of Water Resources Planning and Management. 136 (2009) 412–432.

[11] C. Di Cristo, A. Leopardi, Pollution source identification of accidental contamination in water distribution networks. J Water Resour Plann Manage. 134 (2008) 197–202.

[12] A. Kumar, M.L. Kansal, G. Arora, Identification of monitoring stations in water distribution system, J. Environ. Eng. 123 (1997) 746–752.

[13] J. Gironás, L.A. Roesner, J. Davis, L.A. Rossman, W. Supply, Storm water management model applications manual. National Risk Management Research Laboratory, Office of Research and Development: US Environmental Protection Agency, 2009.

[14] B.K. Banik, C. Di Cristo, A. Leopardi, Illicit intrusion characterization in sewer systems, Urban water journal. 2015 (Submitted).

[15] B.K. Banik, L. Alfonso, A.S. Torres, A. Mynett, C. Di Cristo, A. Leopardi, Optimal placement of water quality monitoring stations in sewer systems: an information theory approach, 13th Computer Control for Water Industry (CCWI2015) conference, 2–4 September 2015, Leicester, UK (Submitted).