Engineering Applications of Artificial Intelligence 25 (2012) 309–316

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Coupling of logistic regression analysis and local search methods for characterization of water distribution system contaminant source

Li Liu [a,*], Emily M. Zechman [b], G. Mahinthakumar [b], S. Ranji Ranjithan [b]

[a] School of Civil Engineering, Hefei University of Technology, 230009 Hefei, Anhui, China
[b] Department of Civil, Construction, and Environmental Engineering, North Carolina State University, 27695 Raleigh, NC, USA

ARTICLE INFO

Available online 9 November 2011

Keywords:
Logistic regression analysis
Local search
Contamination
Water distribution systems

ABSTRACT

Accidental or intentional drinking water contamination has long been and remains a major threat to water security throughout the world. An inverse problem can be constructed, given sensor measurements in a water distribution system (WDS), to identify the contaminant source characteristics by integrating a WDS simulation model with an optimization method. However, this approach requires numerous compute-intensive simulation runs to evaluate potential solutions; thus, determining the best source characteristic within a reasonable computational time is challenging. In this paper, we describe the development of a WDS contamination characterization algorithm by coupling a statistical model with a heuristic search method. The statistical model is used to identify potential source locations of contamination and a local search aims at further refining contaminant source characteristics. Application of the proposed approach to two illustrative example water distribution networks demonstrates its capability of adaptively discovering contaminant source characteristics as well as evaluating the degree of non-uniqueness of solutions. The results also showed that the local search as an optimizer has better performance than a standard evolutionary algorithm (EA).

1. Introduction

Water distribution systems (WDSs) are highly vulnerable to various threat attempts. Intentional biochemical contamination has become a major concern recently due to the potential hazard to human health and the inherent complexities of the contaminants as well as the system. The installation of well-designed monitoring stations is necessary to detect the contamination and enhance the response capability. Rapid and accurate characterization of contaminant sources, once detected, is critical for managing an accidental or intentional WDS contamination event. The process of contaminant source determination involves not only the rapid identification of the injection locations but also the characterization of start time, duration and magnitude of contaminants to effectively control the spread of contamination as well as remediating the contaminated area. However, insufficient data and countless possible contamination scenarios can pose challenge to the characterization process in terms of both accuracy and efficiency.

Various optimization based approaches for WDS source identification problems have been recommended by several researchers. The approaches reported previously can be classified into two categories. The first one employs direct sequential and simultaneous methods (e.g., van Bloemen Wannders et al., 2003; Laird et al., 2005, 2006). The second category couples a search procedure with a WDS hydraulic and water quality simulation package. One such approach, introduced by Guan et al. (2006), links the reduced gradient method with the WDS simulation to identify contaminant sources. Another simulation-optimization approach, proposed by Liu et al. (2006), uses a multiple population-based EA to search for a set of contaminant source characteristics that may result in similar sensor observations. Preis and Ostfeld (2007, 2008) described a straightforward approach for contaminant source identification by coupling EPANET simulator with a genetic algorithm. Zechman and Ranjithan (2009) investigated evolution strategy (ES)-based approach and suggested ways to best structure the algorithm to solve this class of problem. Overall, increasing attention has recently been attracted to heuristic search methods due to the complexity of such a problem, such as discreteness, nonlinearity, non-convexity as well as non-uniqueness. More recently, some new methods have been reported to enhance the solution efficiency and allow practical application, such as Perelman and Ostfeld (2010), Shen and McBean (2010), etc.

Heuristic search methods have their potential in tackling such complex optimization problems by incorporating simulation models. However, such a method results in increased computational burden because of a large number of time-consuming simulation runs needed to evaluate potential solutions. In particular, it poses

* Corresponding author.
E-mail address: lliuncsu@gmail.com (L. Liu).

0952-1976/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2011.10.009

challenges to a large network in terms of the identification time as well as the solution quality, even using parallel or distributed computing implementations. Computational requirements may be reduced by using a prescreening technique that eliminates infeasible solutions to reduce a priori the decision space in which the heuristic procedure must search. One such prescreening method is the back-tracking algorithm reported by De Sanctis et al. (2006, 2010), with the aim of identifying all possible locations and times that explain contamination incidents detected by water quality sensors. Di Cristo and Leopardi (2008) proposed an approach using the pollution matrix concept to determine a group of candidate nodes that could explain discrete solute concentration measurements. Another approach, proposed by Neupauer et al. (2010), identifies the probability density functions of possible prior times when the observed contamination was at any upgradient node. Logistic regression model (LRM) is another potential prescreening approach that shows promise, which has been investigated by Liu et al. (2011) to estimate the probability of each node being a candidate source node. The use of LR analysis would provide a more suitable tool due to its computational efficiency and the ability to describe source characteristics allowing for various uncertainties associated with the contamination event.

Although the knowledge of potential source locations from the LRMs assists decision makers in isolating the source hydraulically from larger network, estimation of release history is required to identify potential extent of contamination, thus reducing public exposure in the contaminated area (De Sanctis et al., 2006, 2010). Therefore, subsequent to the LR analysis, heuristic search methods (e.g., EAs) could be considered to further refine the identification solutions. EAs discover global optima independently of initial guesswork through a population-based scheme and particular operators (e.g., crossover, mutation); however, as the population gradually moves towards optimal solutions, the efficiency of EAs typically decreases due to the stochastic property of the search process (Gen and Chen, 1997; Xu et al., 2001). In contrast, local search (LS) approaches, such as the Nelder–Mead Simplex (NMS) and the Hooke–Jeeves pattern search method, focus mainly on locating locally better solutions by using a deterministic strategy, even though these methods are somewhat sensitive to the initial starting points. Hart (1994) and Land (1998) discuss the advantages of such LS procedures for determining the optima in a quick and computationally efficient manner when the search is focused on a local region.

In this study, nongradient-based LS algorithms are examined as the heuristic search method due to the computational efficiency. Considering that the quality of LS solutions depends on the quality of the starting solution to these iterative search procedures, it is desirable that the smaller set of candidate nodes identified by the LR approach will serve as a set of good starting solutions for the subsequent LS procedure. Thus, the overall procedure investigated in this paper consists of: (1) a LR analysis-based prescreening and (2) a LS technique-based optimization. These components are described in the following subsections.

2. Problem statement and solution techniques

2.1. LR analysis for estimating potential contaminant source locations

The WDS contaminant source identification is challenged by the high degree of uncertainties resulting from numerous possible injection scenarios, unknown water consumption at demand nodes, and errors inherent to measurements and models. To offer a fast probabilistic estimation of candidate source locations, a linear LRM-based approach has been reported in Liu et al. (2011) to model the likelihood that any given node is a source. The LRM, constructed as follows, describes the relationship between the probability of node i as a source and the observations at time t once the contamination is detected at one or more sensors

$$\log\left(\frac{p(A_i|C_1(t),\ldots,C_N(t))}{1-p(A_i|C_1(t),\ldots,C_N(t))}\right)=b_0(i,t)+b_1(i,t)C_1(t)+\cdots+b_j(i,t)C_j(t)+\cdots+b_N(i,t)C_N(t), \tag{1}$$

where $p(A_i|C_1(t),\ldots,C_N(t))$ denotes the likelihood of the contaminant introduced at node i given the observations at time t; A_i represents the contaminant entering through node i; $(C_1(t),\ldots, C_N(t))$ are the sensor observations at time t; and $(b_0(i,t),\ldots, b_N(i,t))$ are regression coefficients for node i at time t obtained by the maximum likelihood procedure. The LRMs are pre-established to describe the contaminant as a function of available measurements using a large number of hypothetical contamination simulations. Once established, using these LRMs can lead to fast estimation of candidate source nodes once the contamination is detected. The strength of the LRM is that it offers a simple and direct way to make a fast prediction with low computation costs. Additional procedural details and application results for example problems are described by Liu et al. (2011). In this study, the resulting candidate source locations from the LRMs are used to reduce the space of subsequent searches by eliminating unnecessary nodes that have estimated zero probabilities of being source locations.

2.2. LS approach for WDS contaminant source characterization

After identifying the set of candidate contaminant source locations, heuristic search methods are used to enhance the identification accuracy by determining the optimal characteristics (i.e., injection location, start time and loading history) of the contaminant sources. The objective is to minimize the difference (i.e., the error) between the simulated concentration values and the observed concentration values at the sensors. The following mathematical formulation describes a form of the error function that is minimized by the search method

$$\text{Find}\{L,M_{t_c},T_0\}$$

$$\text{Minimize } F=\sqrt{\frac{\sum_{t=t_0}^{t_c}\sum_{i=1}^{N_s}(C_{it}^{obs}-C_{it}(L,M_{t_c},T_0))^2}{N_s * t_c}}, \tag{2}$$

where F is the prediction error; L is the contamination source location; T_0 is the injection start time; t_0 is the initial detection time of contamination; t_c is the current time; $M_{t_c}=\{m_{T_0},m_{T_0+1},\ldots,m_{t_c}\}$, represented as a vector of mass injected at the source from time T_0 to t_c, denotes the contaminant mass loadings; C_{it}^{obs} is the observed concentration at sensor i at time t; $C_{it}(L,M_{t_c},T_0)$ is the model (i.e., EPANET) calculated concentration value at sensor i at time t; i is the sensor location; t is the observation time; and N_s is the total number of sensors.

The nongradient-based LS method used to solve this optimization problem in the present paper is the Nelder–Mead Simplex (NMS) search, introduced by Nelder and Mead (1965), with the aim of estimating the contaminant release history that corresponds to each candidate source node prescreened by the LRMs. The scheme of the NMS method is to exploit local information and direct the search towards the optimal or near-optimal solutions by replacing the worst vertex with the newly found better vertex in an adaptive manner through iterations. The algorithm terminates if the stopping criterion is reached. A detailed description of the NMS method can be found in Nelder and Mead (1965).