ABSTRACT

LIU, LI. Real-time Contaminant Source Characterization in Water Distribution Systems. (Under the direction of S. Ranji Ranjithan and G. Mahinthakumar.)

Accidental/intentional contamination continues to be a major concern for the security management in water distribution systems. Once a contaminant has been initially detected, an effective algorithm is required to recover the characteristics of the contaminant's source based on dynamically varying streams of sensor observations. This dissertation focuses on the development and demonstration of a new algorithm to characterize a contaminant source quickly, accurately, and robustly. An evolutionary algorithm (EA)-based adaptive dynamic optimization technique (ADOPT) is proposed, potentially providing a real-time response. In addition to offering adaptive capacity in a dynamic environment, this algorithm is able to assess the degree of non-uniqueness in the solution through multi-population scheme. This approach, however, requires a large number of time-consuming simulation runs to evaluate possible solutions, and it may be difficult to converge on the best solution or a set of alternative solutions within a reasonable computational time. For this reason, it is desirable to appropriately reduce the decision space over which the optimization procedure must search to reduce the computational burden and to produce faster convergence. A logistic regression-based prescreening technique is investigated in order to reduce the decision space by estimating the probability of a node being a contaminant source location. When a small set of potential source nodes are identified, applying the local search procedure to this set of locations is computationally efficient and potentially good at identifying the best solution. The EA-based ADOPT is then integrated with a logistic regression analysis and a local improvement method to expedite the convergence and to solve the problem potentially faster.

The effectiveness of the proposed methods is demonstrated for contamination source identification problems in two illustrative water distribution networks.

Real-time Contaminant Source Characterization in Water Distribution Systems

by
Li Liu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Civil Engineering

Raleigh, North Carolina
2009

APPROVED BY:

_____	_____
Dr. E. Downey Brill, Jr.	Dr. Emily M. Zechman
_____	_____
Dr. Sankar Arumugam	Dr. G. Mahinthakumar Co-Chair of Advisory Committee
_____	
Dr. S. Ranji Ranjithan Co-Chair of Advisory Committee	

BIOGRAPHY

Li Liu was born and grew up in Shouxian, a small city in China that boasts of a history of thousands of years. She is the youngest daughter of Zhitian Liu and Wenlan Wang, and has three older sisters, Chuanjun, Chuanxia and Chuanling. Li entered Hefei University of Technology in China for her undergraduate studies after graduating from high school. Upon graduation with a Bachelors degree in Civil Engineering, she was engaged as a lecturer by the Department of Civil Engineering at Hefei University of Technology. In 2001, she elected to remain at Hefei University of Technology for her Masters degree, specializing in Systems Engineering of Water Resources under the direction of Dr. Juliang Jin. In January 2005, Li enrolled in the doctoral program at North Carolina State University in Raleigh, North Carolina, where her research focused mainly on contaminant source characterization in water distribution systems.

ACKNOWLEDGMENTS

I would like to express my sincerest appreciation to my advisor, Dr. Ranji Ranjithan, for his continuous support, intelligent guidance, patience and encouragement during the past four years. This research could not been completed without his assistance and efforts. I am also indebted to the professors in my committee, Drs. Downey Brill, Sankar Arumugam, and G. Kumar Mahinthakumar for their invaluable suggestions and time.

I would like to thank my graduate and research peers, Emily Zechman, Sarat Sreepathi and Jitendra Kumar for their assistance and helpful discussions in conducting research. Thanks also to my officemates, Yong Jung, Xin Jin, Matthew Clayton, Eiman Abbas, Michael Tryby, Ozge Kaplan, and Pamela Schooler for their friendship and encouragement.

My special thanks go to my parents and sisters for their endless support and encouragement. And finally, my thanks are going to my husband, Xin Zhou, for fully supporting and understanding my decision to come to the United States and continue my studies.

TABLE OF CONTENTS

LIST OF TABLES vi
LIST OF FIGURES vii
CHAPTER 1: Introduction 1
CHAPTER 2: An Adaptive Optimization Technique for Dynamic Environments 7
2.1. Introduction 7
2.2 Solution Approach 10
2.2.1 Evolution Strategy (ES) for Dynamic Optimization Problems 10
2.2.2 Adaptive Dynamic Optimization Technique (ADOPT) 11
2.3. Case Studies 14
2.3.1 Moving Peaks Benchmark (MPB) Problem 15
2.3.2 Groundwater Contaminant Source Determination 21
2.4. Conclusions and Future Work 28
CHAPTER 3: 30
3.1. Introduction 31
3.2. Contaminant Source Identification Problem Description in a WDS 34
3.3. Solution Approach 35
3.3.1 Evolution Strategy for Contaminant Source Characterization 36
3.3.2 Conceptual Basis for ADOPT 37
3.3.3 Algorithmic Steps of ADOPT 39
3.4. Illustrative Case Studies 40
3.4.1 Small Example Network 42
3.4.2 Micropolis Example Network 56
3.5. Summary and Discussion 61
CHAPTER 4: Logistic Regression Analysis to Estimate Contaminant Sources in Water Distribution Systems 63
4.1. Introduction 64
4.2. Problem Description 66
4.3. Logistic Regression Analysis for the Rapid Determination of a Contaminant Source 67
4.3.1 Logistic Regression (LR) Analysis 67
4.3.2 Model Construction 68
4.3.3 Data Generation 70
4.3.4 Performance Evaluation 71
4.4. Applications and Results 71
4.4.1 Small Example Network 72
4.4.2 Micropolis Example Network 79
4.5 Final Remarks 84
CHAPTER 5. Contaminant Source Characterization using Logistic Regression Analysis and Local Search Methods 86
5.1. Introduction 86
5.2. Logistic Regression Model (LRM) 88

5.3. Local Search Approach 89
5.4. Coupling the LRM with the LS for Contaminant Source Characterization 93
5.5 Case Studies 95
5.5.1 Small Example Network 96
5.5.2 Micropolis Example Network 104
5.6 Summary 108
CHAPTER 6. A Hybrid Heuristic Search Approach for Contaminant Source Characterization 110
6.1. Introduction 111
6.2. Problem Statement 113
6.3. Solution Approach 114
6.3.1 ES-based ADOPT 114
6.3.2 Logistic Regression Model (LRM) 116
6.3.3 Heuristic Search Methods 117
6.3.4 Algorithm Framework 119
6.4. Applications 122
6.4.1 Performance Comparison among ADOPT, LRM-ADOPT, and LRM-ADOPT-LS 125
6.4.2 Local Search Selection 127
6.4.3 Effect of Mutation Sizes 131
6.5. Summary 133
CHAPTER 7. Summary and Final Remarks 135
REFERENCES 140

LIST OF TABLES

Table 2.1 Parameters Settings for MPB 17
Table 2.2 Comparison of Average and Standard Error of *Generation Error* for ADOPT and Multi-objective Optimization-based Methods by Bui et al. (2005) 17
Table 2.3 Sensitivity of Parameter Settings of ADOPT 18
Table 2.4 Description of Groundwater Simulation Parameters 24
Table 3.1 Allowable Range of Source Parameters for the Case Studies 42
Table 3.2 True Source Description for Four Contamination Event Scenarios 54
Table 4.1 Contaminant Source Parameters and Ranges for Generating Training Data Set ... 73
Table 4.2 Summary of Results under Various Uncertain Conditions 75
Table 4.3 Statistical Summary of the LRM Results that Correspond to the Injection Node .. 79
Table 5.1 Allowable Ranges of Contaminant Source Parameters 96
Table 5.2 True Source Description for Five Contamination Event Scenarios 97
Table 5.3 Number of Potential Nodes Obtained from the LRMs for Five Hypothetical Contamination Events 98
Table 5.4 Summary of the LRM-LS Results for Five Hypothetical Contamination Events .. 99
Table 5.5 Summary of Results for Scenario 1 under Different Sensor Conditions 103
Table 6.1. Allowable Ranges of Source Parameters 125
Table 6.2 Parameter Settings of the (μ+λ) ES-based ADOPT 125
Table 6.3 Allowable Ranges of Mutation Sizes 132

LIST OF FIGURES

Figure 2.1 Comparison of average *generation error* for ADOPT with different parameter settings 20
Figure 2.2 Variation of diversity with generation for ADOPT with different parameter settings 21
Figure 2.3 Comparison of ADOPT results at time steps 6, 8, and 10 for Scenario 1 and Scenario 2 25
Figure 2.4 Results of ADOPT at time step 20 for Scenario 1 26
Figure 2.5. Average number of the remaining subpopulations through time. 27
Figure 3.1 Layout of the small network. 43
Figure 3.2 Results for the base scenario using ES-based ADOPT 45
Figure 3.3 Results from different mutation strategies based on the base scenario 49
Figure 3.4 Effect of initial number of subpopulations on the ADOPT solutions 50
Figure 3.5 Effect of the number of generations on the ADOPT solutions. 52
Figure 3.6 Effect of the quantity and quality of monitoring data on the number of alternative solutions. 53
Figure 3.7 Four hypothetical contamination scenarios along with the base scenario: 54
Figure 3.9 Micropolis water distribution network schematic 56
Figure 3.10 Identified solutions at 6:50 p.m. for the hypothetical contamination event in the micropolis network. 58
Figure 3.11 Identified solutions at 9:00 p.m. for the hypothetical contamination event in the micropolis network. 59
Figure 3.12 Comparison of concentration profiles between the observed and calculated profiles at S3 for the hypothetical event in the Micropolis network 59
Figure 3.13 Comparison between the number of subpopulations and alternatives for a hypothetical event in the micropolis network 60
Figure 4.1 Water distribution network schematic (small network example) 72
Figure 4.2 Comparison of performance of LRMs between Scenario 1 and Scenario 2: 74
Figure 4.3 Comparison of results between perfect and binary sensor conditions 77
Figure 4.4 Layout of micropolis water distribution network 80
Figure 4.5 Comparison of results between two model-building strategies (micropolis network). 82
Figure 4.6 Locations of possible sources at 12:30 p.m. 83
Figure 4.7 Locations of the top 50 solutions at 1:40 p.m. 83
Figure 5.1 Flowchart of LRM-LS model for contaminant source determination in a WDS. . 94
Figure 5.2 Layout of small example network 96
Figure 5.3 Illustration of injection locations for five hypothetical events 97
Figure 5.4 Comparisons of results for Scenario 1 under different sensor conditions 102
Figure 5.5 Summary of results for 50 random contamination events 104
Figure 5.6 Schematic of micropolis water network 105

Figure 5.7 Location of possible source locations at 12:30 p.m. 106
Figure 5.8 Location of alternative solutions at 1:40 p.m.: 107
Figure 6.1 LRM-ADOPT-LS optimization framework 122
Figure 6.2 Layout of the example networks 124
Figure 6.3 Comparisons of the results from different approaches 127
Figure 6.4 Comparison of the results between LS selection strategies 130
Figure 6.5 Effects of mutation sizes 132

CHAPTER 1: Introduction

Municipal drinking water supply systems aim to provide safe drinking water to customers throughout an entire service area. These systems begin at water sources and convey drinking water to designated points after the water is treated through widely distributed municipal water networks that involve pipes, storage tanks, and pumps. Because the water network is wide and contains numerous possible access points, the system is vulnerable to possible threat, including physical attack, cyber-disruption, and biochemical contamination (Clark and Deininger, 2000). Biochemical contamination, either accidental or intentional, has become a major concern recently due to its inherent complications and the potential hazard to human health.

An effective source identification algorithm is needed to mitigate such possible threats by taking appropriate control actions in the event of contamination. Contaminant sources are typically characterized by their injection location, starting time, duration, and mass injection rates; these factors correspond to different time intervals that are based on sensor observations made shortly after the first detection. However, developing such an algorithm that can characterize contaminant sources in a timely and accurate manner is a challenge due to numerous possible contamination scenarios and sparse observation data. Contamination in the form of any type of pollutant may occur at any time and at any point in a water supply system. To complicate matters, pollutants could possibly be injected simultaneously at multiple locations or with a mixture of various chemical contaminants.

The accuracy of the characterization largely depends on the number and quality of the sensor observations. Ideally, the sensors installed in the network would provide perfect and contaminant-specific observations at each node in the water network. In reality, however, the data may be imperfect or may merely indicate the presence of contamination. The cost of sensor installation and monitoring discourages an adequate number of observations to be made, which further complicates the characterization.

Contaminant source characterization can be viewed as an inverse problem, which is typically ill-posed as opposed to a forward problem in which model parameters are known. For example, a contaminant with diverse strengths injected at different times and locations may yield similar explanations among the sensor observations. That is, current observations are unable to distinguish the true injection node from other potential nodes due to insufficient data. This non-uniqueness of solutions may result from limited data, model error, or measurement error, all of which contribute to the complexity of such a problem.

In addition to non-uniqueness, uncertainties in the system contribute to the complexity. Although the Supervisory Control and Data Acquisition (SCADA) system allows water utilities operaors to access the real-time or near real-time status of a network, determining the actual water consumption at each node still poses a challenge. In the event of contamination, fluctuations in water demand could be dramatic. Hydraulic simulations based on normal conditions are obviously inappropriate, because water demands are estimated according to the population (population density and the service area) and a statistical summary of daily water consumption under normal conditions, not contamination conditions.

One way to address this problem is to formulate it as an optimization problem by making use of currently available observation data to inversely seek the best source characteristics. This formulation could yield a good explanation of the water quality samples that are obtained. Recently, various efforts have been made to develop approaches to address this problem. The optimization methods that have been applied include direct and simulation-optimization approaches. In van Bloemen Wannders et al. (2003), a standard successive quadratic programming tool is applied to solve a small-scale problem. Laird et al. (2005) present an origin tracking algorithm to address the inverse problem of contamination source identification based on a nonlinear programming framework. Taking advantage of a simulation-optimization approach, whereby a search procedure is coupled with a simulation model, Guan (2006) demonstrates its applicability to nonlinear contaminant source and release-history identification.

The limitations of the aforementioned work often include the inability to address non-uniqueness, deterministic characterization, small network size, high computational costs, etc. In the event of contamination, a well-developed algorithm must be able to: 1) dynamically update the determination of source characterization as the number of observations changes over time; 2) adaptively assess the degree of non-uniqueness and also identify non-unique solutions if available; and 3) quickly provide real-time solutions.

Heuristic search algorithms have received increasing attention due to their potential in tackling complex optimization problems. In addition to the effectiveness of such algorithms, the objective function evaluation is the only information required to direct the search during the iterative process. This advantage allows wider applications to a variety of complicated

problems by coupling the simulation models. Evolutionary algorithms (EAs) (Holland, 1975), as one class of heuristic methods, present a global search (GS) mechanism and are of great benefit for solving large nonlinear optimization problems. EAs are also flexible enough to be extended to dynamic and noisy conditions. They have been used in several water distribution network design problems (e.g., Dandy et al., 1996; Savic and Walters, 1997). Although EAs have been used effectively to solve inverse problems, such as water distribution network calibration (e.g., Vitkosvsky et al., 2000; Lingireddy and Ormsbee, 2002) and groundwater source contamination identification problems (e.g., Mahinthakumar and Sayeed, 2005; Mahar and Datta, 1997), little is known about their applicability to source characterization in a water distribution systems (WDS) under dynamic and noisy conditions.

The primary objective of this study is to design a contaminant source characterization procedure for a WDS by integrating simulation models with a heuristic search-based optimization framework, given observation data that stream in over time, and to adapt this algorithm or couple it with some other method(s) for realistic source identification in the WDS. Overall, this dissertation has five main objectives:

1. Develop a new methodology, the EA-based Adaptive Dynamic OPtimization Technique (ADOPT), that can solve an optimization problem in dynamic environments.

2. Investigate the applicability of ADOPT to characterize contaminant sources in real time.

3. Develop a fast estimation methodology to determine the likelihood that any given node is a potential source by processing the observation data obtained from monitoring stations in the WDS over time.

4. Investigate the applicability of coupling the proposed prescreening technique with heuristic search methods, such as ADOPT, local search methods or a combination of both, to improve the efficiency of the solutions.

5. Examine the effects of uncertainties on the resulting solutions of the proposed methods.

The organization of this dissertation is as follows. Chapter 2 presents the new proposed methodology, ADOPT, which is an EA-based search procedure that is designed to adapt the search for time-changing optimal solutions under changing environments. This chapter also demonstrates the applicability of this procedure to a test problem and a groundwater contaminant source identification problem. Chapter 3 describes the application of ADOPT to contaminant source characterization in a WDS. The effects of different parameter settings on the results are also investigated. Chapter 4 presents a new procedure based on logistic regression (LR) analysis to estimate the likelihood that any given node is a contamination injection location, given the observations obtained from the monitoring sensors. This chapter also reports the results of the proposed procedure for hypothetical contamination events via two water example networks. In Chapter 5, a coupling procedure is proposed that integrates a prescreening technique with local search (LS) methods. Chapter 6 describes an algorithmic framework for solving contaminant source characterization problems in a WDS by integrating EA-based ADOPT using one prescreening and one

postscreening technique. In this chapter, the impact of uncertain factors on the proposed approach is evaluated. Finally, the findings obtained from the current study, as well as future plans, are discussed in Chapter 7.

CHAPTER 2: An Adaptive Optimization Technique for Dynamic Environments

Abstract. The use of evolutionary algorithms (EAs) is beneficial for addressing optimization problems in dynamic environments. The objective function for such problems changes continually; thus, the optimal solutions likewise change. Such dynamic changes pose challenges to EAs due to the poor adaptability of EAs once they have converged. However, appropriate preservation of a sufficient level of individual diversity may help to increase the adaptive search capability of EAs. This chapter proposes an EA-based Adaptive Dynamic OPtimization Technique (ADOPT) for solving time-dependent optimization problems. The purpose of this approach is to identify the current optimal solution as well as a set of alternatives that is not only widespread in the decision space, but also performs well with respect to the objective function. The resultant solutions may then serve as a basis solution for the subsequent search while change is occurring. Thus, such an algorithm avoids the clustering of individuals in the same region as well as adapts to changing environments by exploiting diverse promising regions in the solution space. Application of the algorithm to a test problem and a groundwater contaminant source identification problem demonstrates the effectiveness of ADOPT to adaptively identify solutions in dynamic environments.

2.1. Introduction

Many optimization problems must be solved in the context of a dynamic environment to provide real-time solutions efficiently. In general, such optimization problems of special

interest involve time-dependent objective functions that gradually change with time. Examples of time-dependent dynamic optimization problems include dynamic vehicle routing, scheduling, and threat management problems. The commonality among these problems is that the environment, such as dynamic information and time-varying restrictions, is continually changing while decisions are being made. With the intention of effectively discovering optimal solutions in real time, an adaptive approach is required that not only determines the current optimum but also quickly adjusts the solutions to a new environment when change occurs.

A number of researchers have demonstrated that EAs, as a class of heuristic search methods, offer the potential to handle such optimization problems due to their population-based search properties (Holland, 1975). In the context of complex real-world optimization problems, the use of EAs is particularly beneficial because they can integrate complicated simulation models to evaluate objective functions effectively. However, as EA populations converge to the best solution for the current setting of the problem, they have difficulty adapting to the changing environment, limiting the application of traditional EAs to solve dynamic optimization problems. Nevertheless, various methods that extend the use of EAs for addressing such dynamic optimization problems have been developed in recent decades, including memory-based and diversity-based techniques.

The memory-based techniques are composed of implicit memory schemes (Goldberg and Smith, 1987) and explicit memory schemes (Branke, 2002; Yang, 2005). The main enhancement of the performance in these schemes is the ability to store useful information from the current environment, taking into account that this information could be retrieved

later when a new environment emerges. The memory-based scheme has proved highly useful for the resolution of cyclical dynamic environment problems.

Another approach to solve dynamic optimization problems is to sustain adequate diversity while the environment changes. Several diversity-based approaches have been developed, such as maintaining diversity via random immigrants (Grefenstette, 1992), modifying selection processes (Ghosh et al., 1998; Goldberg and Richardson, 1987), multi-objective optimization (Bui et al., 2005), and multi-population techniques, such as self-organizing scouts (Branke, 2002), shifting balance genetic algorithms (GAs) (Oppacher and Wineberg, 1999), and multinational GAs (Ursem, 2000).

Hybrid approaches are a combination of different concepts, such as memory and diversity (Branke, 2002). Conceptually, this kind of approach incorporates the advantages of both memory and diversity schemes. An example of such a hybrid approach is a memory-based immigrant scheme within a GA, developed by Yang (2005), where the best solution is stored to generate random immigrants that potentially introduce the diversity of the search process in dynamic environments.

This chapter introduces a new adaptive search method based on EAs that is structured to search continually for the moving optimum in a dynamic environment. To avoid premature convergence and strengthen adaptability, the proposed approach attempts to guide groups of individuals of the EA population to move concurrently towards diverse promising regions in the decision space. Accordingly, at any stage of the solution procedure, a set of solutions, including the optimal and near-optimal solutions (within certain accuracy), is determined and used as the set of solutions for subsequent searches as the new environment

emerges. To allow a reasonable evaluation of the proposed approach, the previously developed multi-objective optimization-based method (Bui et al., 2005) is used as a base, and the results are compared with those of ADOPT via the Moving Peaks Benchmark (MPB) problem using the same parameter settings. In addition to illustrating the ability to adaptively maintain diversity, ADOPT's ability to adapt to a changing environment is demonstrated through different algorithmic settings and problem scenarios. In this chapter, the framework is applied to a groundwater contaminant source characterization problem where the monitoring observations are dynamically updated. This adaptive capability results in an effective assessment and resolution of the degree of non-uniqueness of the solutions. The methodology is sufficiently general to be applicable to other time-dependent optimization problems in which the environment changes gradually between consecutive time steps.

2.2 Solution Approach

2.2.1 Evolution Strategy (ES) for Dynamic Optimization Problems

EA-based approaches have been proven to be suitable for dynamic environments due to the dynamic and stochastic manner in which the solution is evolved (Branke, 2002). In addition, the ability to couple the algorithm with a simulation model of the real system enables its applicability to solve real-world problems. The ability of evolution strategies (ES), a class of EAs, to adapt its step length during the search is specifically suitable for search under dynamic conditions. The scheme that is beneficial to ES is self-adaptive mutation in which each individual in the population represents the decision variables as well as their

mutation step lengths (Yang, 2007). Once mutated, the step length is used to create a random vector to mutate the decision variables accordingly. Thus, the mutation strengths progress along with the individuals instead of through predetermined values. The ES has been demonstrated to possess a self-learning property, even in the dynamic context of an optimization problem (Hoffmeister and Back, 1992).

2.2.2 Adaptive Dynamic Optimization Technique (ADOPT)

Efficient evolution of individuals to locate the current optimum as the problem condition changes is critical for adaptively handling the changes over time. The strategy proposed in this work uses an ES-based adaptive dynamic technique that continually identifies the optimal solution for the current conditions. The procedure starts with an ES population that is comprised of a number of randomly generated subpopulations. These subpopulations search for a set of good solutions to the current problem conditions. The new ES-based approach to identify a set of good solutions is based on the EAGA (evolutionary algorithms to generate alternatives) method developed by Zechman and Ranjithan (2004, 2007), in which each subpopulation is designed to converge towards one of the many basins of attraction that expectedly include the optimal solution and a set of near-optimal solutions. To make the search efficient and effective, the EAGA approach implicitly maximizes diversity among the subpopulations. The solutions in these subpopulations represent the current attractive regions (with higher fitness values) in the decision space. These solutions at the current time are used as starting points for the subsequent search. As the fitness landscape of a dynamic problem gradually and continually changes as new problem conditions emerge,

the subpopulations continue to evolve and collectively track the migrating basins of attraction, consequently tracking the changing optimal solution to the changing environment.

In the proposed algorithm, one subpopulation is set up to search for the best solution (with the best fitness value) that forms a benchmark for assessing the near-optimality condition for solutions in the other subpopulations. A specified degree of relaxation from the optimal objective function value is then used as a basis to evaluate the feasibility of individuals to be accepted as near-optimal solutions. To enable the discovery of other near-optimal solutions that are maximally diverse and are similar in fitness, the remaining subpopulations evolve their individuals to perform well with respect to both the objective function (i.e., within an acceptable deviation from the best fitness value) and diversity (i.e., maximally distant from the other solutions). During the selection process, if the feasible individuals are dominant, then emphasis is placed on the distance evaluation. This process ensures that an individual with a larger distance value is more likely to survive to the next generation. If, instead, the unfeasible individuals prevail, then an individual with a higher fitness value is given a higher probability to be chosen. This procedure is iteratively performed until a stopping criterion is met or a new environment emerges whereby the objective function needs to be updated. At each stage, the resulting optimum and a set of widely distributed potential alternative solutions are anticipated to form the basis for the subsequent search, thus yielding faster convergence as slight change occurs. In the case of a dynamic environment, however, the global optimum may migrate from one location to another; that is, the global optimum does not always emerge from the same subpopulation.

After comparing the optima among all subpopulations at each generation, the one with the current global optimum then serves as the benchmark for the following search.

The proposed method attempts to achieve both convergence and diversification simultaneously. Distance, as an evaluation criterion, is expected to identify maximally different solutions; meanwhile, this algorithm concentrates each subpopulation on the basin of attraction of a single peak as quickly as possible. As a result, a set of potential solutions obtained at the current time could assist the following search, given that the resulting alternatives may perform differently for a new circumstance.

To summarize, the main steps of the algorithm are described as follows:

Step 1. Let time step t = 0. Create an initial population with N subpopulations. N depends on the complexity of the problem.

Step 2. Let time step t = t + 1. Construct the objective function for time step t. Set the generation as g = 0.

Step 2.1. Let g = g + 1. In each subpopulation, evaluate the fitness of each individual based on its objective function and its distance function that can be measured as the distance between the individual and the other subpopulations.

Step 2.2. Compare the objective values for the best individuals of all subpopulations and obtain the best solution and its objective value as the generation optimal value. Set the subpopulation in which the best solution is obtained as the first subpopulation. The generation optimal value is used

as a target to determine the feasibility of individuals in other subpopulations and to maintain good regions in the decision space.

Step 2.3. Apply selection and mutation operators to all subpopulations and create a new set of solutions. In the first subpopulation, selection is based on the objective function only, whereas in other subpopulations selection is based on both the objective and distance functions. And, in each subpopulation, the best solution in terms of the objective function needs to be carried to the next generation.

Step 2.4. If the criterion that g < max no. of generations is not met, then go to Step 2.1; otherwise, go to Step 3.

Step 3. Check for termination criteria. When t equals the maximal time step, stop the algorithm. Otherwise, go to Step 2 and use the current solutions as the starting points for the next step search.

2.3. Case Studies

The performance evaluation of the proposed algorithm is carried out through two representative problems in dynamic environments. The first one is the Moving Peaks Benchmark (MPB) problem, developed by Branke (1999), which serves as a basis for preliminary performance evaluation. The second case is a hypothetical groundwater contamination source characterization to investigate the applicability of ADOPT to a real-world problem.

2.3.1 Moving Peaks Benchmark (MPB) Problem

The moving peaks function as a benchmark problem has been applied by a number of researchers investigating dynamic optimization methods. The landscape of the moving peak function includes a number of peaks, with a changing height, width and location for each peak. The mathematical formulation of the time-dependent fitness of an n-dimensional test function with m peaks is described as

$$F(\vec{x},t) = \max(B(\vec{x}), \max_{i=1\ldots m} P(\vec{x}, h_i(t), w_i(t), \vec{p}_i(t))), \quad (2.1)$$

where $B(\vec{x})$ is a time-invariant "basis" landscape, and P is the function that defines a peak shape, where each of the m peaks has its own time-varying parameters: height (h), width (w), and location ($\vec{p}_i(t)$) . A detailed description of the MPB is available online (http://www.aifb.uni-karlsruhe.de/~jbr/MovPeaks). The key issue to solving the MPB problem is how to keep track of the highest location in a time-varying landscape. The potential of ADOPT to handle this problem is investigated as described below, beginning with a comparison between ADOPT and the multi-objective optimization-based methods developed by Bui et al. (2005).

To assess the performance differences between ADOPT and the method by Bui et al. (2005), ADOPT was applied to the MPB problem using the same parameter settings as those given by Bui et al. (2005), which are listed in Table 2.1. Four problem instances were generated by varying the values for the peak height and width parameters h_σ and w_σ, respectively (Table 2.2). To solve each scenario, ADOPT was set up with 20 subpopulations (n=20), each of which consists of two parents (μ=2) and three mutants (λ=3). The total

number of time steps was set to 40, with 25 generations for each time step. This implies a total 1,000 function evaluations for each trial, which is the same as that used for the results reported by Bui et al. (2005). In ADOPT, a 10% relaxation target was used to search for near optimal alternative solutions. At the end of each generation of ADOPT, the performance (indicated as *generation error* that represents a measure of accuracy of the solution) was evaluated as the difference between the objective value of the best individual at each generation and its corresponding true current global optimum. Considering the probabilistic nature of the ES, ADOPT was executed for 30 independent random runs, and the average and standard variation of *generation error* were obtained by summarizing the results of the 30 runs.

In Bui et al. (2005), a dynamic optimization with a single objective is converted to a multi-objective problem, and the methods are named as follows based on the way the distance is defined to maintain diversity: time-based, random, inverse, distance to the closet neighbor (DCN), average distance to all individuals (ADI), and the distance to the best individual of the population (DBI). This conversion is helpful for the population to remain diverse, thereby circumventing premature convergence during the search process. Table 2.2 summarizes the average *generation error* and its standard error for the solutions obtained using ADOPT and the multi-objective optimization-based methods.

Table 2.1 Parameters Settings for MPB

Parameter	Value	Parameter	Value
Number of peaks	50	Std width	0.0
Number of dimensions	5	Min coordinate	0
Min height	30	Max coordinate	100
Max height	70	Change every x evaluations	2500
Std height	50	Peak function	cone
Min width	1.0	Change step size	constant
Max width	12.0		

Table 2.2 Comparison of Average and Standard Error of *Generation Error* for ADOPT and Multi-objective Optimization-based Methods by Bui et al. (2005)

Method	**h_σ=7 w_σ=1**	**h_σ=7 w_σ=3**	**h_σ=15 w_σ=1**	**h_σ=15 w_σ=3**
Time-based	12.06±0.64	12.96±0.81	12.06±0.80	15.06±1.00
Random	11.29±0.55	12.30±0.96	14.79±0.66	14.20±0.83
Inverse	12.37±0.87	13.96±0.87	15.98±0.89	15.28±0.88
DCN	9.52±0.45	10.42±0.71	12.68±0.60	12.56±0.62
ADI	9.74±0.35	9.31±0.51	13.18±0.52	13.00±0.63
DBI	12.24±0.55	11.79±0.71	14.05±0.61	13.96±0.74
ADOPT	**7.72±0.29**	**8.85±0.26**	**8.24±0.29**	**9.09±0.21**

Of the results reported by Bui et al. (2005), the DCN, ADI and DBI methods yield relatively low average generation errors. The results generated by ADOPT indicate a better performance than all the multi-objective optimization-based methods for each of the four cases. These results imply that, even though both ADOPT and the multi-objective optimization-based methods attempt to achieve convergence and diversity simultaneously, ADOPT, with its multiple subpopulations, adapts better to changing environments than the methods reported in Bui et al. (2005).

Because the total number of objective evaluations within a certain period is limited, the appropriate settings for the size of each subpopulation and the total number of subpopulations are critical for yielding an effective performance by ADOPT. To evaluate the ways in which the performance of ADOPT varies with the number of the subpopulations and the size, a sensitivity assessment using six combinations of parameter settings were carried out, and their corresponding values are listed in Table 2.3.

Table 2.3 Sensitivity of Parameter Settings of ADOPT

Cases	**Number of Subpopulations (n)**	**Population Size**		**Generation Error**	
		μ	***λ***	***Avg+StdError***	***Lowest***
1	1	50	50	20.29±0.95	11.49
2	5	10	10	12.09±0.54	7.25
3	10	5	5	9.95±0.47	5.94
4	20	2	3	7.72±0.29	5.51
5	25	2	2	7.78±0.30	5.65
6	50	1	1	7.64±0.23	5.53

This sensitivity evaluation was conducted for the problem setting with the changing severity parameter values of 7 for h_σ and 1 for w_σ, and for the same number of evaluations. The results for ADOPT from 30 random trials are summarized in Table 2.3 and are graphically shown in Figure 2.1. Table 2.3 presents the lowest *generation error* and the average *generation error* over all time steps among all trials for each case. When each subpopulation contains at least four individuals, an increase in the number of subpopulations yields better solutions with lower *generation error* values. An increase in the number of subpopulations with fewer than four individuals does not yield, however, an improvement in

the ADOPT performance. This observation suggests that incorporating more subpopulations has a positive effect on the ADOPT performance, whereas smaller number of subpopulations with equivalently large population sizes do not perform as well. Figure 2.1 shows the variation of *generation error* with time for the six cases. After a sharp increase in the error at the beginning where the populations consist of random solutions, *generation error* gradually decreases as the solutions continue to converge to better objective function values. Comparing across all six cases, in general the cases with larger number of subpopulations perform better at all time steps although the population sizes are smaller. At the initial stage, however, the improvement resulting from a larger population size is evident. ADOPT, with its large number of subpopulations and small population size, is thus capable of swiftly identifying the optimum at later stages due to a high level of diversification it is able to maintain throughout the search.

To measure and monitor the degree of diversity, the distance of each individual from the individuals in other subpopulations is calculated. Figure 2.2 shows the normalized distance among all individuals for the six cases. Again, cases with larger number of subpopulations are able to maintain, in general, a higher degree of diversity at all generations. For multimodal problems, the ability of ADOPT to adaptively diversify and maintain solutions at different peaks in the solution space helps the algorithm quickly identify a good solution as the dynamically changing optimal peak migrates over time. It must be noted that in addition to the positive results of ADOPT for the cases presented in this chapter, one of the most valuable aspects of this approach is that ADOPT is able to assess the current status of the landscape. Accordingly, the number of subpopulations can be adjusted to

match a new situation once a change occurs, resulting in less computational costs and an improvement in the algorithm performance.

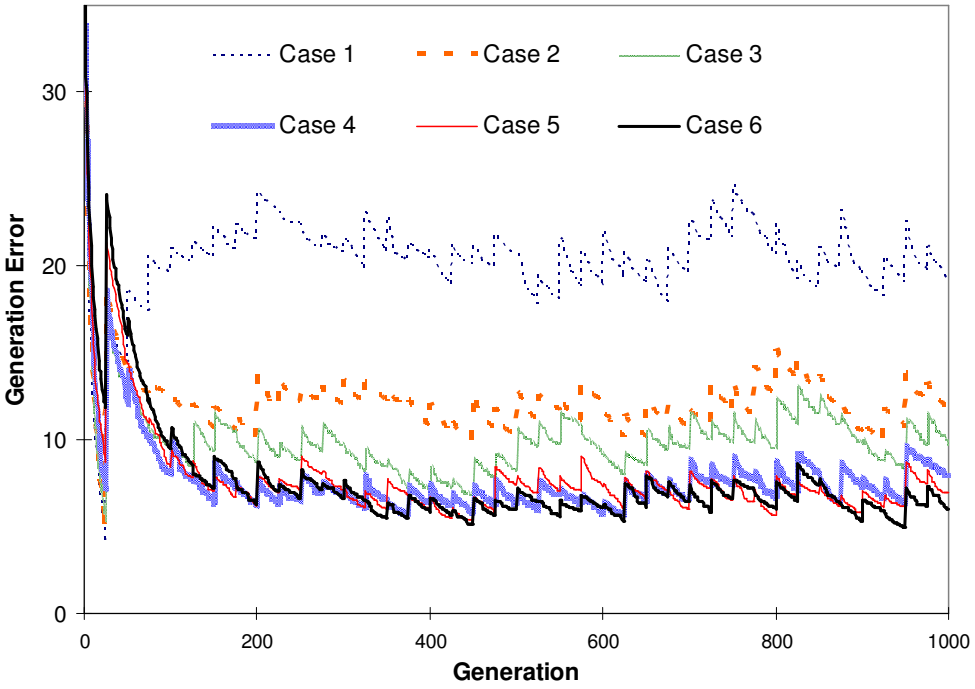


Figure 2.1 Comparison of average *generation error* for ADOPT with different parameter settings

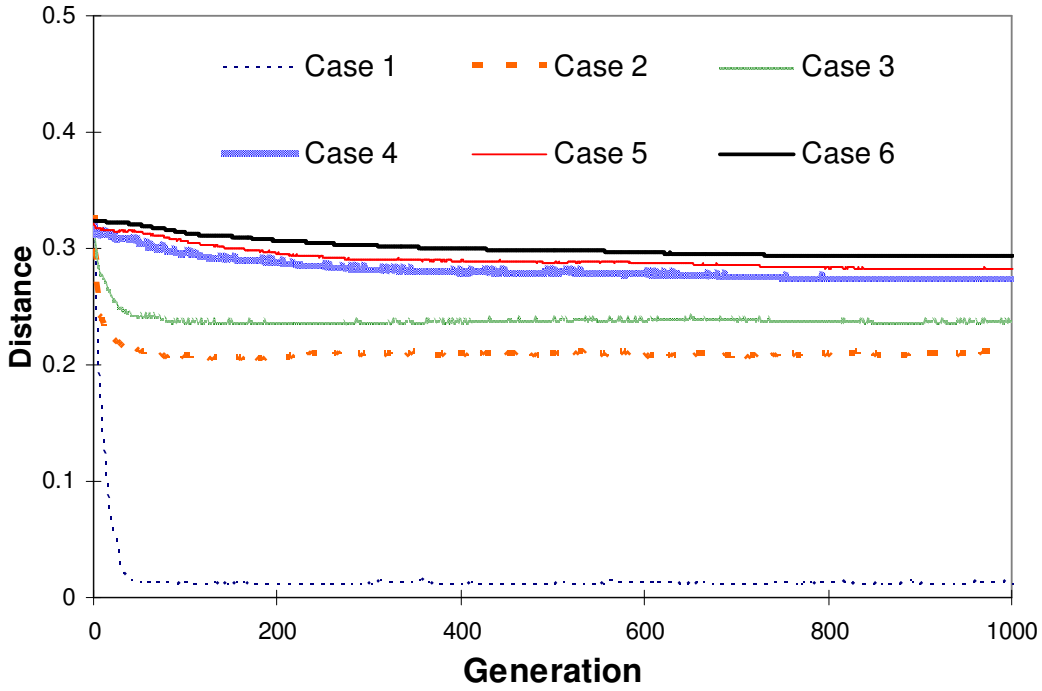


Figure 2.2 Variation of diversity with generation for ADOPT with different parameter settings

2.3.2 Groundwater Contaminant Source Determination

Groundwater contaminant source identification is an important and difficult step in groundwater remediation. This problem, which is posed as an inverse problem, involves the timely and accurate characterization of the contaminant source based on observations from a system of monitoring wells in an aquifer. For a given set of observations, source characterization is posed as an optimization problem. The source characterization is updated continually and dynamically as new measurements become available, and it turns into a dynamic optimization problem. The objective of such a problem is to find the optimal contaminant source characteristics, including location, size and release history, by minimizing the difference between the simulated concentration and the observed

concentration at monitoring wells over time. Therefore, this objective function needs to be updated when new information is available, and can be expressed as

$$\text{Minimize } f = \sum_{t=t_0}^{t_c} \sum_{i=1}^{N} (C_{i,t}^{sim} - C_{i,t}^{obs})^2 , \tag{2.2}$$

where $C_{i,t}^{sim}$ is the simulated concentration at the *i*th monitoring well at time step *t*; $C_{i,t}^{obs}$ is the observed concentration at the *i*th monitoring well at time step *t*; t_0 is the starting time for observation; t_c is the current time step; and *N* is the number of monitoring wells.

One issue in the groundwater contaminant source identification problem is the presence of non-uniqueness of solutions, i.e., more than one solution could explain the observations, especially when available monitoring information is insufficient. Thus, it is important to identify the set of non-unique solutions that fit the limited information. As additional measurements are incorporated, the set of non-unique solutions must be resolved to converge adaptively to the solution that describes the most likely source characteristics. Furthermore, if this were a real case, it would be necessary to assess whether the available measurements are sufficient, or if more information should be obtained by continuing measurements at existing monitoring wells or by adding new observation wells. Therefore, the source characterization must be conducted continually and adaptively until a unique solution is identified, or until the set of non-unique solutions that best fit the available measurements is identified.

In this chapter, ADOPT approach is applied to solve a hypothetical groundwater contaminant source identification problem. Initially, multiple populations are designed to maintain a set of alternative solutions that represents various non-unique solutions. As more

observations are added, the ADOPT solutions not only migrate to improved solution states, but also reduce the number of solutions as the degree of non-uniqueness diminishes, which accordingly decreases the number of populations. This step could be taken by comparing the similarity of solutions for different subpopulations. If two subpopulations converge to locations close to each other in the decision space, one subpopulation should be eliminated to avoid unnecessary computation. When observations are sufficient to identify the source within an acceptable accuracy threshold, the best solution is obtained from the existing population.

To begin to solve the groundwater contaminant source identification problem, a representative problem is identified such that it reflects a possible real-world scenario. The problem is kept simple enough for this preliminary study, yet it is able to generate results that illustrate the feasibility of this method. Here, the domain size is taken as 100 by 60 meters, and observations are simulated for 20 time steps. A single source problem is investigated; the temporal concentration release history and the shape (which is square with side lengths of 2 meters) of the source are assumed to be given, but the contaminant source location is treated as unknown. To identify the unknown, an optimization model is used, which minimizes the maximal absolute error between the calculated and observed concentrations based on all observation time steps and all monitoring wells. The search space for the source centroid, whose coordinate is represented as (x, y), is the whole domain ($0 \leq x \leq 100$, $0 \leq y \leq 60$). A detailed description of this scenario is shown in Table 2.4. For this problem, the ADOPT parameters used include an initial number of 20 subpopulations, each of which consists of 20 individuals, and the number of generations is 10 for each observation interval.

Table 2.4 Description of Groundwater Simulation Parameters

Parameter	Value
Field size	100 m × 60 m
Number of time steps	20
Time step size (Δt)	10 day
Grid spacing (Δx=Δy)	2 m
Dispersion parameters	αL=1 m; αT = 1 m; Dm = 0.01 m2/d
Flow field	Homogeneous
Velocity	1 m/day
True source description	Shape: square (with side length of 2 m) Centroid coordinate: (15, 29) Concentration: 70 mg/L

Preliminary results are shown for a two-dimensional homogeneous aquifer with the contaminant introduced at a single location. Plume is generated assuming advection-dispersion processes in the porous media. Synthetic observations at two monitoring wells through 20 time steps are generated assuming a pulse source. Source identification Scenario 1 represents a case wherein a new observation at only Well 1 is incrementally added at each time step. Scenario 2 represents a case wherein new observations at both Wells 1 and 2 are incrementally added at each time step.

The EA-based ADOPT was executed for 30 random trials for each scenario. The representative results shown here are based on a typical run. For each scenario, the progression of the search procedure is shown in terms of the non-unique solutions obtained for each scenario as more measurements are added over time that is represented by the number of time steps in Figure 2.4. The true source is also shown for comparative purposes.

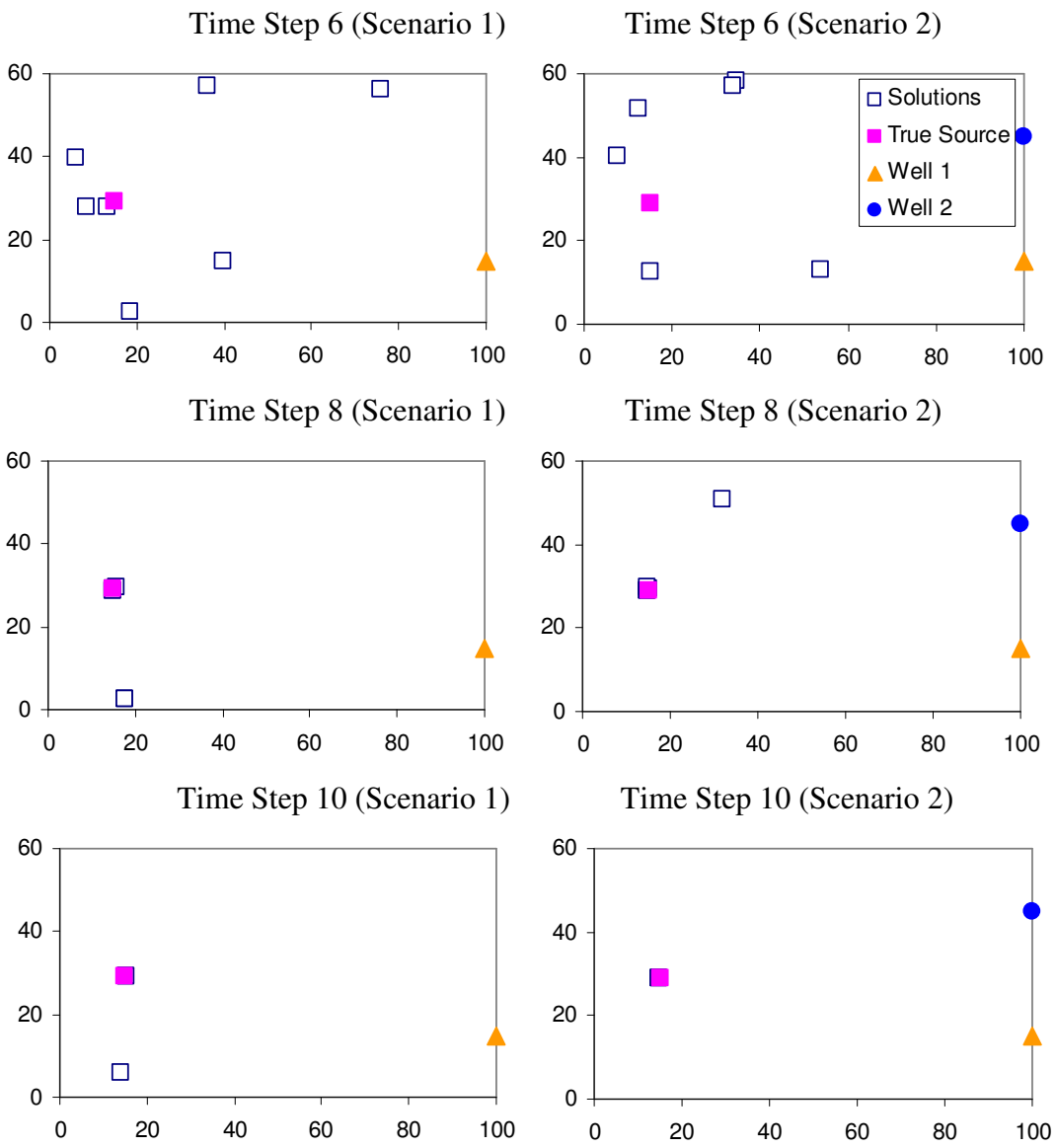
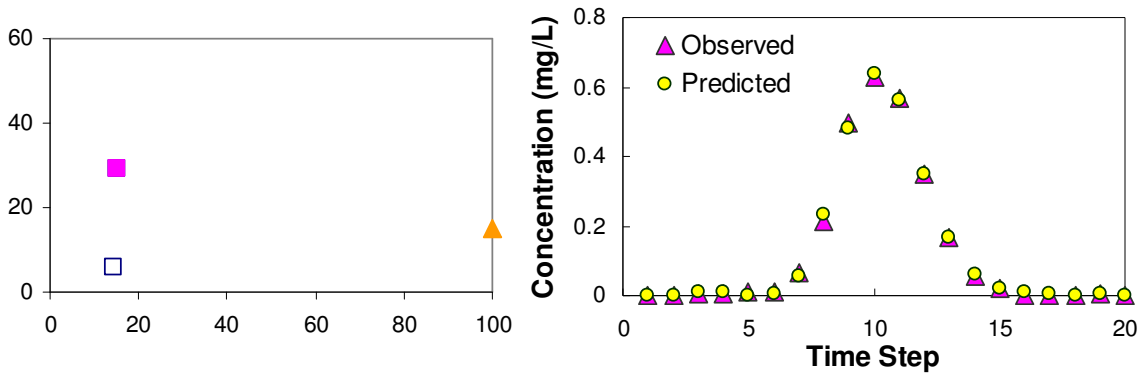


Figure 2.3 Comparison of ADOPT results at time steps 6, 8, and 10 for Scenario 1 and Scenario 2



(a) (b)

Figure 2.4 Results of ADOPT at time step 20 for Scenario 1:
(a) Location of the optimal solution; (b) Comparison of observed and predicted concentration profiles at Well 1.

Scenario 1 has several non-unique solutions that provide the same observations at Well 1 (see Figures 2.3 and 2.4). Out of the 30 trials, one of the non-unique solutions was identified 15 times. The rest of the trials converged to multiple solutions, indicating a high degree of non-uniqueness due to observations being limited to only one observation well. Overall, measurements over more time steps were required to resolve the non-uniqueness issue. Figure 2.4(a) shows two solutions identified by ADOPT at time step 20. These two solutions provide similar observations at the observation Well 1, one of which is close to and the other is distant from the true source location. Figure 2.4(b) shows a comparison of concentration profiles at Well 1 for the true source and the alternative source that is distant from the true source; as can be seen, there is good agreement between these two concentration profiles. Scenario 2 has, however, only one unique solution because all 30 trials converged to the correct solution prior to time step 20. Compared to Scenario 1, measurements over fewer time steps were sufficient in Scenario 2 to solve the problem as

additional observations from Well 2 help resolve the non-uniqueness and identify the source correctly.

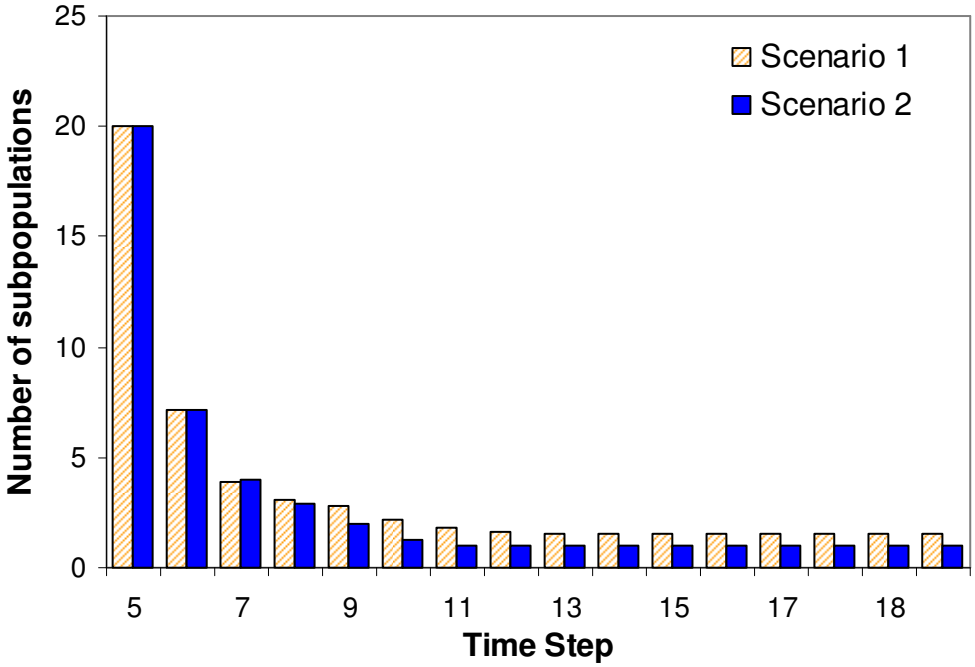


Figure 2.5. Average number of the remaining subpopulations through time.

The results shown in Figure 2.5 report the variation of the number of the remaining subpopulations over time within ADOPT, which is averaged from among the 30 random trials. Note that if an ADOPT run terminates prior to the final step, the number of subpopulations is recorded as 1 in the following steps. The graph in Figure 2.5 shows that the ADOPT procedure dynamically reduces the number of subpopulations as more available observations are become available. Compared to Scenario 1, Scenario 2 has fewer subpopulations although both start with the same number of subpopulations (20, in this study).

2.4. Conclusions and Future Work

This study introduces an adaptive dynamic optimization procedure based on EAs for solving dynamic optimization problems. Whereas the continually shifting optima challenge EAs, ADOPT is designed not only to capture the current optima, but also to simultaneously preserve an appropriate degree of diversity. To arrive at a balance between convergence and diversity, the emphasis is put on either the distance or the objective function in the selection process, which depends on the feasibility of the majority of the individuals in a subpopulation. Such balance ensures the necessary clustering of subpopulations around maximally diverse potential solutions depending on the current status of a problem, the number of subpopulations is dynamically adjusted. This ability could be helpful in improving the solution quality as well as avoiding unnecessary computational costs.

Application to a 5-dimensional moving peaks function, developed by Branke (1999), shows the effectiveness of ADOPT to adjust individual subpopulations to capture the moving optima. The solutions generated by ADOPT can be compared to the results of the multi-objective optimization-based methods developed previously (Bui et al., 2005). The effect of ADOPT parameter settings on the solution quality was further examined. Although a large population size can accelerate the convergence at the initial stages of the search, the number of subpopulations with a small population size is particularly beneficial at later stages of the ADOPT runs. Furthermore, this algorithm was extended to solve a groundwater contaminant source characterization problem with a fixed optimal solution to a dynamically varying objective function over time, which is slightly different from a general dynamic optimization problem. In addition, this dynamic optimization problem includes an added complexity in

that the degree of non-uniqueness in solutions varies over time. The results presented here demonstrates that ADOPT has the potential to dynamically identify a set of solutions that yields similar and good fit to the observed data. By varying the observations, either by utilizing longer observation periods or by incorporating more monitoring wells, ADOPT is demonstrated to be successful in adaptively reducing the number of subpopulations and assessing the degree of non-uniqueness.

Although the results have shown the success of ADOPT's ability to adjust the number of subpopulations when characterizing a groundwater contaminant source, appropriately regulating the number of subpopulations remains an issue. Further work is needed to examine ways to fine-tune the number of subpopulations in the context of various problem scenarios. The target for evaluating the solution feasibility is a vital parameter in the progression of solutions, as it controls the direction of each subpopulation. If the target is high, the selection pressure will be put on the objective function, thereby yielding rapid convergence but with a potential loss of diversity. Further exploration into the appropriate settings of these algorithmic parameter values is needed. Also, the applicability of ADOPT to more realistic dynamic optimization problems needs to be evaluated. As most real-world problems include model and input uncertainty issues, it is important to investigate the robustness of ADOPT when applied to dynamic optimization problems with uncertainties.

CHAPTER 3: Adaptive Contamination Source Identification in Water Distribution Systems Using an Evolutionary Algorithm-based Dynamic Optimization Procedure

Abstract. Accidental drinking water contamination has long been and remains a major threat to water security throughout the world. Consequently, contamination source identification is an important and difficult problem in managing safety in a water distribution system (WDS). This problem involves the characterization of the contaminant source based on observations that stream from a set of sensors in the distribution network. Because the spread of contamination in a WDS is often fast and unpredictable, rapid identification of the source location and related characteristics is critical in order to control the contaminant and take containment actions. The streaming data can be processed adaptively to provide an estimate of the source characteristics at any time once the contamination event is detected. That is, as the contaminantion event unfolds, this estimate is continually updated as new observations become available. In this chapter, an adaptive dynamic optimization technique (ADOPT) is proposed for providing a real-time response to a contamination event. Over time, additional data are observed at a set of sensors, thus changing the vector of observations that are to be predicted. The prediction error function is then updated dynamically, thus changing the objective function in the optimization model. A new multiple population-based search that uses an evolutionary algorithm (EA) is investigated. At any given time the EA represents the solution that best matches the available observations. The set of populations migrates to represent updated solutions as new observations are added over time. During the initial detection period, non-uniqueness is inherent due to inadequate

information; consequently, several solutions may predict the observations similarly well. To address non-uniqueness in the initial stages of the search and prevent premature convergence of the EA to an incorrect solution, the multiple populations in the proposed methodology are designed to maintain a set of alternative solutions that represent various non-unique solutions. As more observations are added, the EA solutions not only migrate to better solution states, but the number of solutions decreases as the degree of non-uniqueness diminishes. This new dynamic optimization algorithm adaptively converges to the solutions that best match the observations available at any time. The new method is demonstrated for contamination source identification problems in two example water distribution networks.

3.1. Introduction

Accidental and intentional contamination of a WDS is becoming an increasingly critical issue. For example, a pollutant source introduced into a WDS will spread through the system rapidly and expose the public to health risks. Detection of the contamination in the distribution system using a sensor network could yield useful observations to identify and manage such contamination threat events. Based on these observations, the location, strength, time and duration of the contaminant source needs to be determined to direct decision-makers toward containing and mitigating the event. Given a set of concentration observations at sensors in the network, an inverse problem can be constructed to identify the contaminant source characteristics (including location, strength and release history) by coupling a water distribution simulation model with an optimization method. Possible solutions to this inverse problem are determined by minimizing the error between predicted concentrations and actual

observations of the models at the sensor nodes in the network. In the context of a quickly evolving contamination event in a WDS, the correct source characterization must be resolved rapidly as the sensor observations, i.e., contaminant concentrations at the sensors in the network, stream in over time.

Although inverse modeling has been applied to a wide array of system identification problems in engineering, it has the potential for non-uniqueness in that different sources with significantly different pollutant release characteristics but with similar prediction errors may be identified. Because the non-uniqueness in a system is related to the amount of data available for identifying the source(s) of the contamination, more data, made available through either additional sensors or an extended monitoring time, may help reduce the degree of non-uniqueness in the system. If the available information is insufficient to determine whether an identified solution is unique or not, then it is important to determine if other possible solutions exist. Knowing that the identified solution is the only possible source characteristic that matches the observations is critical, because a non-unique but incorrect solution may yield potentially costly mitigation actions that may inconsequentially exacerbate the contamination.

The key challenges to solving this problem are: the determination of the source characteristics given the available measurement information at any given time, and the assessment of whether the solution that is identified is unique or not. This determination and assessment require a procedure that is able to: 1) adaptively search for the source characteristics as the observation data are dynamically updated over time; and 2) assess the

degree of non-uniqueness, i.e., whether more than one solution fits the available observations.

Recently, researchers have reported the development of several procedures to identify contaminant source characteristics by using information from sensor networks. A direct sequential technique, reported by van Bloemen Waanders et al. (2003), has been applied to solve a small-scale optimization problem using a standard successive quadratic programming tool. Laird et al. (2005) report a direct simultaneous approach. These methods attempt to identify a single solution using a fixed, i.e., not dynamically streaming, set of observations. New approaches are needed to solve the problem in an adaptive manner. At any given time, the procedure must be able to identify possible solutions that explain the observations available up to that time. Also, the solution procedure must identify not only the best estimate of the source characteristics that minimize the prediction error, but also identify a set of possible alternative solutions, if any, that similarly predicts the available observations.

This chapter reports a new adaptive search method that uses a simulation-optimization approach whereby the water distribution network model is coupled directly with a new dynamic optimization method to iteratively evaluate and identify solutions that minimize the prediction error. To assess the non-uniqueness in the solution, the procedure also incorporates a systematic method to identify alternative solutions that are as different as possible in the solution space. Thus, at any stage of the solution procedure, possible solutions that best describe the observations are determined and are used as starting solutions for subsequent searches as more information becomes available. The search method explored in

this chapter is based on EA that is coupled with an EPANET model of the water distribution network. The applicability of the method is illustrated using a hypothetical example.

3.2. Contaminant Source Identification Problem Description in a WDS

To capture the dynamic nature of the available observation data, the source identification problem is described in terms of the source characteristics, based on observations up to the current time. As the number of observations changes over time, the description of the problem is updated at some regular time interval, i.e., the observation frequency. At any instant, the problem can be solved to obtain an estimate of the source characteristics that best explain the currently available data. The following mathematical model is defined to determine, at any given time after the contamination is detected at one or more sensors, the contamination source location, the contamination event start time, and the corresponding contaminant mass loading history. Although the definition provided below assumes that the contamination is introduced at only one node in the network, the proposed approach can be updated to consider multiple contamination source locations.

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

$$\text{Minimize} \quad F = \sqrt{\frac{\sum_{t=t_0}^{t_c}\sum_{i=1}^{N_s}(C_{it}^{obs} - C_{it}(L, M_{t_c}, T_0))^2}{N_s * t_c}}, \tag{3.1}$$

where F = prediction error;

L = contaminant source location;

T_0 = contamination event starting time;

t_0 = time of first detection of contamination at sensors;

t_c = current time step;

M_{t_c} = contaminant mass loadings represented as a vector of mass injected at the source from time T_0 to t_c; $M_{t_c} = \{m_{T_0}, m_{T_0+1}, \cdots, m_{t_c}\}$;

C_{it}^{obs} = observed concentration at sensor i at time step t;

$C_{it}(L, M_{t_c}, T_0)$ = model estimated concentration at sensor i at time step t;

i = observation (sensor) location;

t = time step of observation; and

N_s = number of sensors.

3.3. Solution Approach

The search for the location and the time history of the contamination injection into the network is a nonlinear programming problem that poses computational challenges depending on the size of the problem. A few search methods to solve the source determination problem have been reported recently; these include direct sequential methods (van Bloemen Wannders et al., 2003) and particle-tracking methods (Laird et al., 2005). Another approach that can be used to solve the inverse problem is a simulation-optimization, or indirect, approach, in which a search procedure is coupled with a simulation model. EAs (Holland, 1975) are a class of heuristic methods that provide a global search mechanism to efficiently identify near-optimal solutions for large nonlinear optimization problems. EAs have been used in several water distribution network design problems (e.g., Dandy et al., 1996; Savic

and Walters, 1997). Although EAs can be used effectively to solve inverse problems, such as WDS calibration (e.g., Vitkosvsky et al., 2000; Lingireddy and Ormsbee, 2002) and groundwater source contamination identification problems (e.g., Mahinthakumar and Sayeed, 2005; Mahar and Datta, 1997), the applicability of EAs to dynamic source characterization in water distribution networks has not been fully investigated. Thus, in this chapter EAs are investigated as an approach to solve adaptively the source determination problem in water distribution networks that is posed as a dynamic optimization model (Eq 3.1). This new EA-based approach, called ADOPT – Adaptive Dynamic OPtimization Technique described in Chapter 2 – is a search procedure that is designed for adaptive optimization and considers dynamically varying streams of sensor observations and identifies alternative solutions, if any, to assess the degree of non-uniqueness in the system.

3.3.1 Evolution Strategy for Contaminant Source Characterization

In this chapter, ADOPT is implemented using evolution strategy (ES) (Schwefel, 1995) to continually search for optimal solutions while the observations are updated. Similar to other EAs, the ES uses a population of individuals simultaneously during the optimization process. Beginning with an initial population that typically is randomly generated, the ES explores new search space by mutation, and individuals with higher fitness values are more frequently selected to be transferred to the next generation. As individuals progress, their average performance is expected to advance gradually. The algorithm terminates when the specified stopping criteria, such as the maximum number of generations, no improvement in the optimum, etc., are met.

The ES presents an adaptive capability, particularly in dynamic circumstances, in that it typically adapts its step lengths during the optimization process. The benefit of the ES is its mutative self-adaptation, whereby each individual can be represented as a decision variable along with its mutation step lengths (Yang, 2007). During the course of mutation, the step length, once mutated, is used to create a random vector to mutate its corresponding decision variable. Thus, mutation rate changes based on the quality of individuals instead of using various predetermined values for parameters, such as mutation and crossover rates. The ES has been demonstrated to possess a self-learning capability, even in the dynamic context of an optimization problem (Hoffmeister and Back, 1992).

To identify a contaminant source in a WDS, a search for various characteristics of contaminants, such as location, starting time, duration, and injection rates at different time intervals is required. The injection profile is represented as an array of real variables, and the source location, starting time, and duration are encoded as integer values in this study. The mutation step size is encoded in a similar way to its corresponding decision variable. The fitness representing the prediction error of an individual is updated with increasingly available sensor observations. Because the mutation operator plays a key role in maintaining diversity within the ES, several mutation strategies to enhance performance were investigated in this study.

3.3.2 Conceptual Basis for ADOPT

The two key features of the new method, ADOPT, are: 1) optimization in dynamic environments, and 2) identification of alternative solutions. In the context of the water

distribution network problem, the number of sensor observations varies with time. That is, the objective function (i.e., the prediction error defined in Eq. 3.1) changes with time. The dynamic optimization procedure is structured to continually search for the best solution at each time step, t. Initially (i.e., at t_0 when the contaminant is first detected), the search uses a set of random solutions as the starting point for the search. The solutions that are found to best fit the observations up to the previous time step are used as the starting solutions for the subsequent instance of the problem that, in turn, is updated with the new observation data obtained during the next observation time step. This approach works well for a population-based search procedure, such as EAs, where the population of solutions continually explores the decision space and migrates towards the right solution as the objective function is adjusted dynamically based on updated observation information over time.

To address the issue of non-uniqueness, ADOPT is structured to search simultaneously for a set of alternative solutions. The EA-based search is designed, based on the method developed by Zechman and Ranjithan (2004, 2007), to consist of multiple subpopulations of solutions, each converging towards a different solution that best fits the current set of observations. For a systematic and efficient search to identify whether or not different solutions exist, each subpopulation of solutions is designed to migrate to a region in the decision space that is maximally different from that of the other subpopulations. If non-unique solutions exist, then more than one subpopulation will converge to a possible solution to indicate that the currently available observations are insufficient to resolve the non-uniqueness in the solution.

3.3.3 Algorithmic Steps of ADOPT

The EA-based procedure is structured to search for a set of possible solutions by exploring the decision space via multiple subpopulations. These subpopulations simultaneously search for solutions that are as different as possible from each other. To set a benchmark for the best possible solution, one of the subpopulations searches independently for the solution that best fits the observations. The remaining subpopulations use that benchmark to find other possible solutions that fit the observations equally or nearly as well as the best solution. To identify maximally different solutions, some measure of distance in the decision space between pairs of subpopulations is maximized. This procedure is executed for each observation time step. The number of observations available up to that time step is used to construct the objective function that represents a metric of prediction error. At any point in the search, each subpopulation represents the state of the best solution to fit the available observations. When new observations are added at the next time step, the objective function is appropriately updated, and the search continues from the current state of solutions represented in the subpopulations. By hot-starting the search at any time step based on the previous solutions, the search in the subsequent time step is expected to be conducted more efficiently, thus yielding better convergence. As a solution in one subpopulation becomes similar to one in another subpopulation, one of these similar subpopulations is removed. Eventually, when sufficient observations are available to identify a unique solution, only one subpopulation remains. These steps in the ADOPT procedure collectively identify at any observation time step the solution that best fits the currently available observations. They

also reveal other possible solutions, if any, to indicate the uniqueness of the solution. The steps of this procedure are listed below.

Step 1. Create an initial set of random solutions, equally divided among N subpopulations.

Step 2. Increment time step t ← t + 1. Set generation index as g = 0. Update the monitoring data with additional measurements and construct the prediction error function.

Step 2.1. Increment generation index g ← g + 1. In the first subpopulation (p = 1), evaluate the fitness based on the prediction error. In subpopulation p (= 2, 3,…N), evaluate the fitness based on the prediction error and its distance from all other subpopulations.

Step 2.2. In each subpopulation, apply selection, recombination and mutation operators, and create a new set of solutions.

Step 2.3. If the stopping criteria (e.g., g < max no. of generations) are not met, then go to Step 2.1; otherwise, go to Step 3.

Step 3. Eliminate subpopulations that represent duplicate solutions. If only one subpopulation remains, or the current set of solutions is acceptable, then stop.

Step 4. If no more observations are available, then stop; otherwise, go to Step 2.

3.4. Illustrative Case Studies

In this section, a number of hypothetical contamination events are studied via two water networks of different degrees of complexity. The purpose of the case studies is to

exhibit the ability of ADOPT to predict alternative source characteristics by coupling the EPANET model with the ADOPT search procedure. Specifically, the search is conducted to determine, at the end of each observation interval, a set of source characteristics that includes the location, start time, and mass loading profile of the contaminant as the contaminant is introduced into the network. Also, the sensitivity of the proposed approach to a range of parameter settings and contamination scenarios is investigated.

Upon confronting a contamination event, ES-based ADOPT starts to execute with a prespecified number of subpopulations. Over time these subpopulations are adaptively regulated, according to the sensor observations, to track of the best solution and a set of alternatives. The resultant alternative solutions must perform similarly well in terms of the prediction error, but are maximally far apart from each other. Given these sensor observations, alternatives can be determined according to the currently available data. For example, the target value, which is used to evaluate the feasibility of the solutions, is set to a relaxation value (120% in this study) of the root mean square of the observed data. Considering that the degree of non-uniqueness diminishes as more measurements are collected, the relaxation value is adjusted accordingly (by 0.7% in this study, based on a set of trials). To avoid unnecessary computational costs, a subpopulation is eliminated from the optimization process when it converges to the same location as another subpopulation. The algorithm terminates when the number of remaining subpopulations equals one or when no additional data are monitored.

In this study, it is assumed that a conservative contaminant is injected at a single location and that the hydraulic condition is deterministic and remains at a steady state during

each hour. These assumptions are made primarily to allow a convenient and viable investigation and exploration of the proposed approach; however, they are not expected to limit a broader applicability of the approach to problems whose conditions deviate from these assumptions. Table 3.1 describes the allowable ranges of source parameters for both case studies.

Table 3.1 Allowable Range of Source Parameters for the Case Studies

Source Parameter	**Small Example**	**Micropolis Example**
Location	Any node (1~97)	Any node (1~1575)
Starting Time (hr)	0~4	0~30
Duration (hr)	0~5	0~7
Mass Injection Rate (g/min)	0~30	0~30

3.4.1 Small Example Network

The first network examined is one of the examples available as a tutorial within EPANET (Rossman, 2000). The network is depicted in Figure 3.1, and further details can be found in the EPANET users' manual. This network consists of 97 nodes, including 2 sources, 3 tanks, and 117 pipes. Four sensors are distributed randomly in this network (denoted as S1, S2, S3, and S4 in Figure 3.1). The contaminant transport is simulated in 10-minute intervals, and the concentration values at the sensors are assumed to be observed in 10-minute increments.

To perform the sensitivity analysis, the investigation starts with a base scenario. A nonreactive contaminant is introduced into the network at node #205 (see Figure 3.1) at 3:00 a.m. for a duration of 2 hours and time-varying mass injection rates of 10, 20, 15, 20, 10, 15, 20, 20, 20, 10, 15, 10 g/min, each of which corresponds to a 10-minute interval. After several

trials under different contamination scenarios, this scenario was selected as a base run because the pollutant injected at this location seems to be relatively difficult to identify.

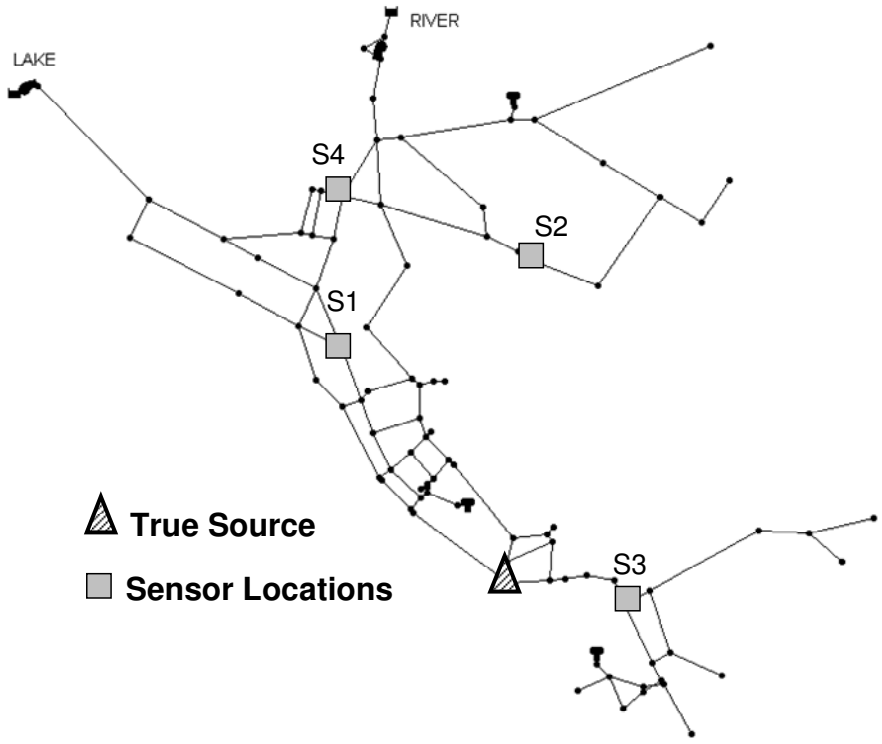


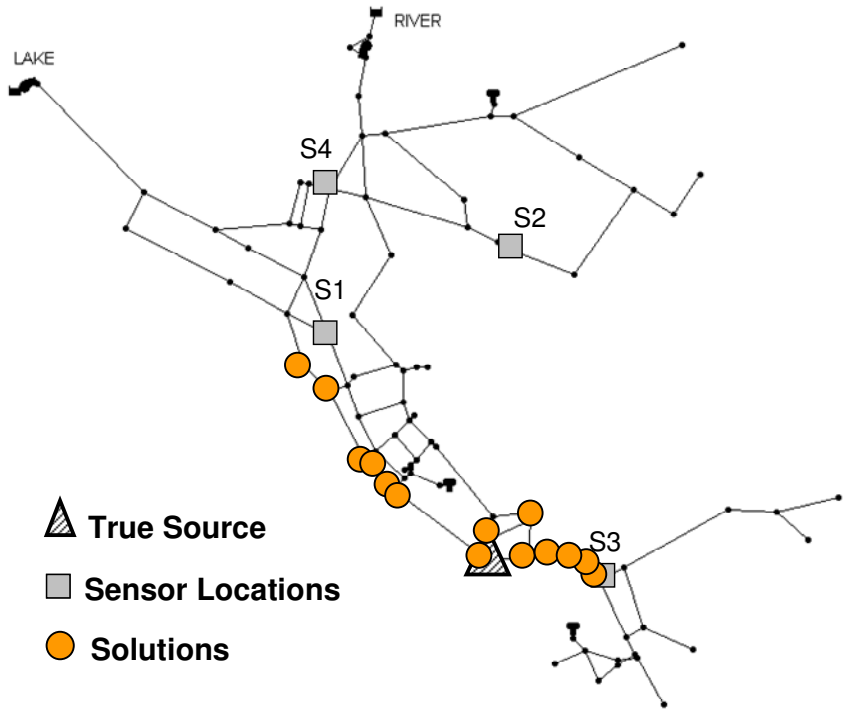
Figure 3.1 Layout of the small network. Contaminant source for the base scenario is indicated by the triangle. The sensor network is composed of S1, S2, S3 and S4, which are designated by squares.

ADOPT, based on (μ+λ)-ES, was applied to the hypothetical event described above. The parameters used include an initial number of 20 subpopulations, each of which consists of 200 parents (μ) and 300 mutants (λ), and the number of generations is 30 at each 10-minute interval. Unless otherwise noted, the same parameter settings were employed to other scenarios as well. To gain statistical significance, 30 random runs were carried out in each case. On average, each random trial using these parameter settings took approximately 5

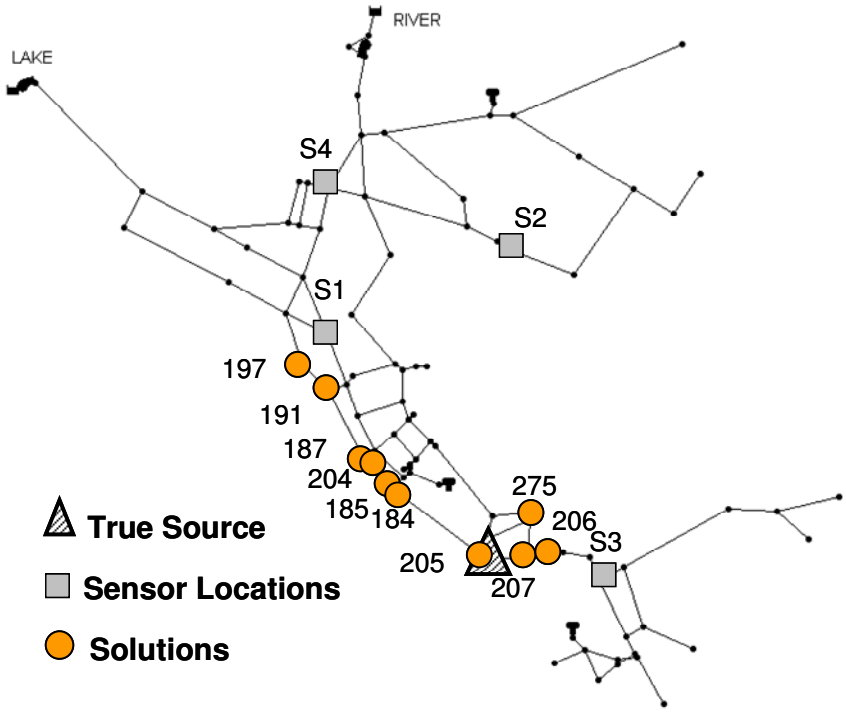
minutes on the Neptune system(an Opteron cluster) where EPANET simulations were executed in parallel on ten processors.

Figure 3.2 shows the ADOPT results of the base scenario from a typical run. The first detection occurred at 3:40 a.m., and ADOPT identified 14 alternatives for the first observations (see Figure 3.2 (a)). As the sensors collected measurements continually, the ADOPT solutions evolved accordingly. At 6:00 a.m., after the last measurements were taken, 10 acceptable solutions were identified and each of these solutions matched the observation within the specified error limit.. The locations of these solutions are shown in Figure 3.2 (b). In addition to the differences in injection locations, these solutions varied in the starting time, duration, and mass injection profile, as shown in Figure 3.2 (c). In spite of such differences, Figure 3.2 (d) indicates a good agreement between the observed and predicted concentration profiles for all 10 solutions at sensor S3. No abnormal measurements were obtained from the other three sensors.

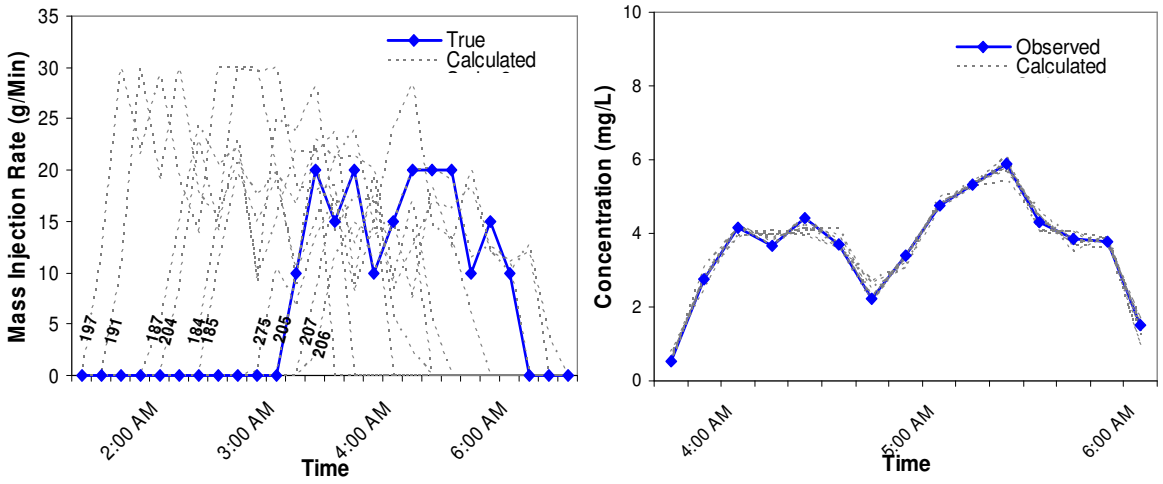
Figure 3.2 Results for the base scenario using ES-based ADOPT: (a) locations of alternatives at 3:40 a.m. (the number adjacent to a solution represents the node ID); (b) locations of alternatives at 6:00 a.m.; (c) comparison of mass injection rates between the true and calculated injection; (d) comparison of concentrations at the selected sensors.


RIVER
LAKE
S4
S2
S1
S3
True Source
Sensor Locations
Solutions

(a)


RIVER
LAKE
S4
S2
S1
197
191
187
204
185
184
275
206
S3
205
207
True Source
Sensor Locations
Solutions

(b)



(c) (d)

Parameter Settings. The proper setting of the algorithm parameters enhances the effectiveness of discovering solutions. In this section, the sensitivity of the results is evaluated in terms of the differences in the parameter settings for the mutation operators, number of subpopulations, and generation numbers. All experiments were performed for 30 random trials based on the base scenario described above. All the results are reported in terms of the average and one standard deviation (the error bar in the bar charts below) values computed from the solutions obtained for all random trials.

Mutation is primary key operator of in the ES procedure. Thus, determining an appropriate mutation strategy is one of the most important steps in the ES-based ADOPT procedure. A self-adaptive strategy was considered, whereby the mutation step size is subject to self-adaptation and evolves with the decision variables. In a static environment, a mutation step size typically decreases over iterations as the individuals gradually approach the optimum. This decrease in step size is especially favorable for fine-tuning at the converged

region in the case of a static situation. Although the ES has been shown to possess a self-learning property in a dynamic environment (Hoffmeister and Back, 1992), more time may be needed to self-adjust step sizes to proper magnitudes from small values than to reinitialize them once an environmental change occurs. As a result, the reconstruction of mutation step sizes may facilitate the adaptability in the dynamic context of the contaminant source characterization.

Several mutation strategies were considered. For each strategy investigated, mutation step sizes were set to the same number as the number of decision variables, one for each dimension. The first set of runs focused on a completely self-adaptive mutation, where mutation step sizes progressed with decision variables throughout the entire process, except for the random initialization at the beginning. For the second set of runs, mutation step sizes were randomized at the start of each generation; that is, the mutation step sizes between generations were unrelated to each other. The last set of experiments was conducted using the self-adaptive mutation strategy between changes, and the step lengths were re-initialized once new data came in. The results of these three strategies averaged over 30 random trials are illustrated in Figure 3.3. Figure 3.3 (a) plots the average minimal prediction error, which corresponds to the optimal solution among all subpopulations. The third strategy in which re-initialization in addition to self-adaptation was included, outperforms the other two schemes at this work. The differences increased with elapsed time. It is also worth noting that the self-adaptation strategy shows an improvement over the random mutation in the initial stages. However, this advantage gradually disappears over time, reflecting the progressive significance of the appropriate re-initialization as solutions move towards the optimum with

the changing environment. To better understand ways that ADOPT performs under different mutation strategies, the level of non-uniqueness represented by the number of alternatives (Figure 3.3(b)), and the number of remaining subpopulations (Figure 3.3(c)) that represents the convergence of individuals as well as the computational cost were compared. As shown in Figure 3.3 (b), the number of alternatives generated provides further evidence for the information presented in Figure 3.3 (a) that the third strategy surpasses the other two schemes. Figure 3.3 (c) reveals a diminishing rate of decline in the number of subpopulations as the sensor observations increase. This trend is more significant in the first experiment due to the lowest of the degrees of diversity introduced during the search process.

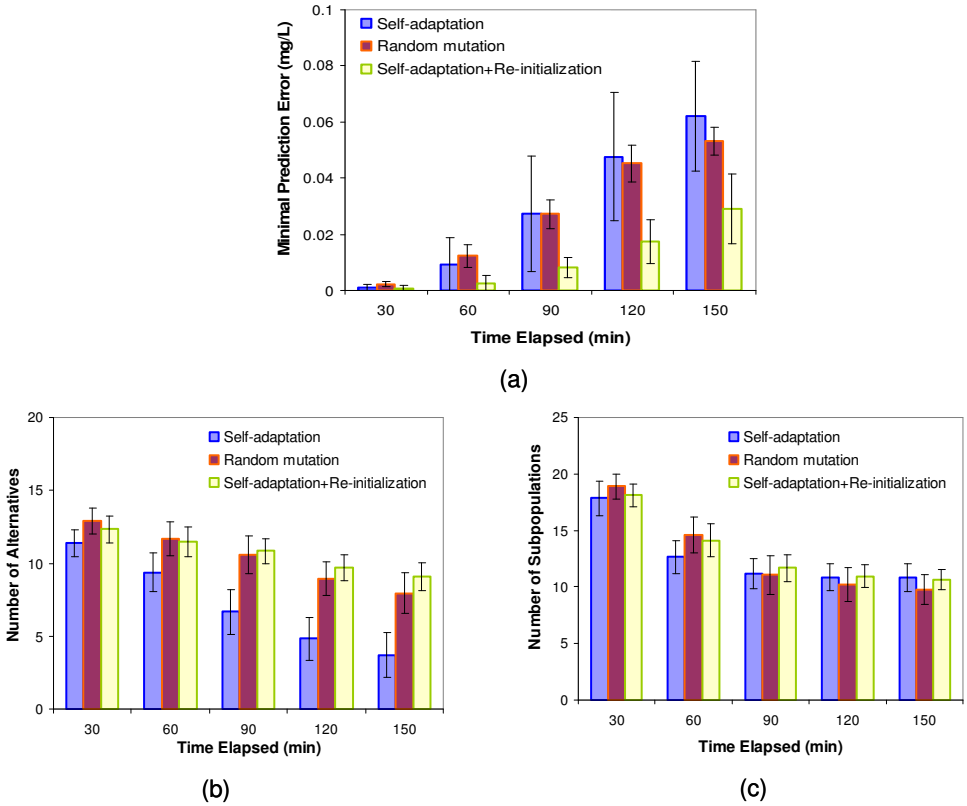
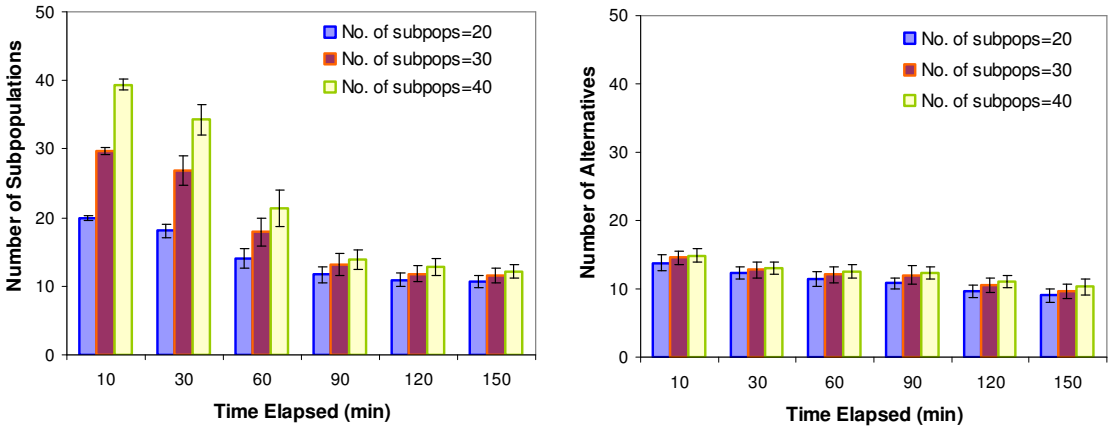


Figure 3.3 Results from different mutation strategies based on the base scenario: (a) minimal prediction error; (b) number of alternatives; (c) number of subpopulations.

When solving a contamination event using ADOPT, it is desirable for the number of subpopulations to be the same as the number of alternative solutions. It is difficult, however, to predetermine the exact number, and therefore an estimate must be specified to execute the procedure. Sensitivity of the performance of ADOPT to this parameter, the initial number of subpopulations, was conducted for 20, 30, and 40. Figure 3.4 (a) shows the number of subpopulations at different elapsed times for three scenarios. As expected, ADOPT, with more subpopulations converges much faster, but at the end of the observation period (150 mins), similar numbers of subpopulations were reached. Moreover, the initial settings of their numbers do not make much difference to the alternatives identified by ADOPT (see Figure 3.4 (b)).



(a) (b)

Figure 3.4 Effect of initial number of subpopulations on the ADOPT solutions: (a) number of subpopulations; (b) number of alternative solutions.

The sensitivity to the number of generations for each time interval was also investigated. From a practical viewpoint, the solution at any time step should provide the best estimates of the potential source. This requires ADOPT to maintain a balance between convergence and diversity at any time step. Obviously, more convergence yields better

solutions in terms of the objective value at any elapsed time, but compromises diversity or becomes mired in local optima. The over-convergence impacts current solutions as well as subsequent searches, especially at the initial stages when the observations are limited. Because of the loss of individual diversity, further effort is required to adapt solutions once new data become available.

Figure 3.5 shows the variation in the number of alternatives over elapsed time for cases, with 10, 30, and 50 generations per time step. More alternatives with low prediction error represent a better overall performance. As shown in Figure 3.5, the case with 10 generations resulted in fewer alternatives compared to the other cases, which was due, apparently, to an insufficient convergence. Overall, the cases with 30 and 50 generations produced similar numbers of non-unique solutions with slight differences between these two cases. For the 30 generations case, more alternatives were identified than the 50 generations case at the initial stages. This phenomenon may be due to premature over-convergence in the latter case. As more monitoring data were included, the case with 50 generations achieves a better performance than the 30 generations case.

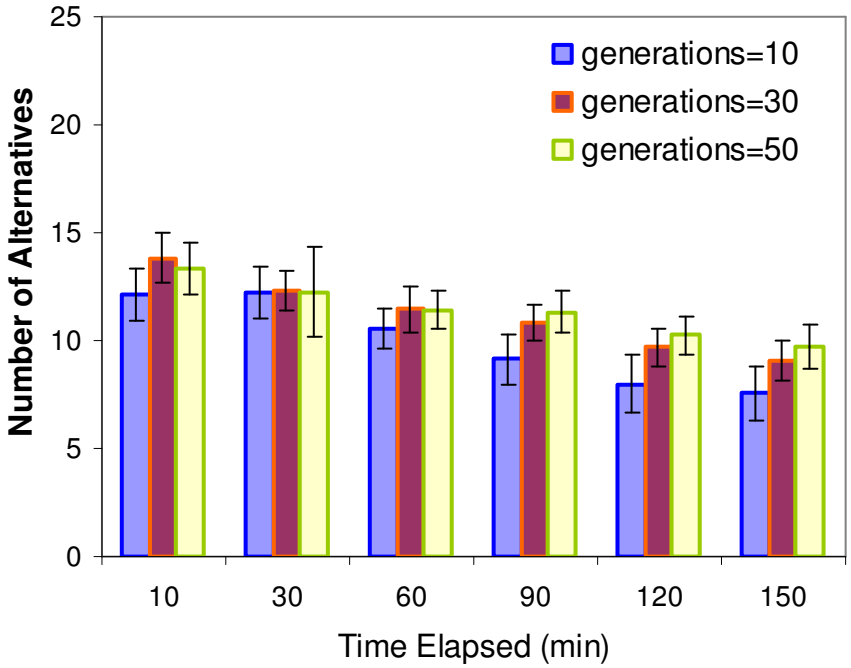


Figure 3.5 Effect of the number of generations on the ADOPT solutions.

Various Contamination Scenarios. While the results from the base scenario indicate the effectiveness of ADOPT in characterizing a WDS contaminant source, a range of contamination scenarios presented here allows the demonstration of the algorithm's robustness given the variations in the amount and quality of the monitoring data as well as the contaminant source characteristics.

An increase in the quantity or quality of observations reduces the degree of non-uniqueness of the solutions. To explore the ability of ADOPT to assess non-uniqueness, two scenarios were added to the investigation to characterize the same pollutant as the base scenario. Case 1 incorporates an additional sensor (denoted as 185 in Figure 3.2 (a)) for comparison to the base scenario, and Case 2 is assumed to have the same sensor distribution as the base scenario. Note, however, that these sensors were simulated to recognize the

existence of contamination rather than the concentration value. A fixed concentration value specifies the threshold or detection limit to trigger a contamination detection status. A reading that exceeds this threshold (a value of 0.1 mg/L is assumed in this study) represents the presence of contamination (converted to an output signal of 1); otherwise, it represents the absence of contamination (represented by an output signal of 0). The degree of non-uniqueness is measured in terms of the number of alternatives. Figure 3.6 shows comparisons of the number of non-unique solutions identified for the two cases and the base scenario. The degree of non-uniqueness diminishes as the amount of observations increases. In all three scenarios, a decrease in either the number or quality of the sensors leads to a higher degree of non-uniqueness in the solutions.

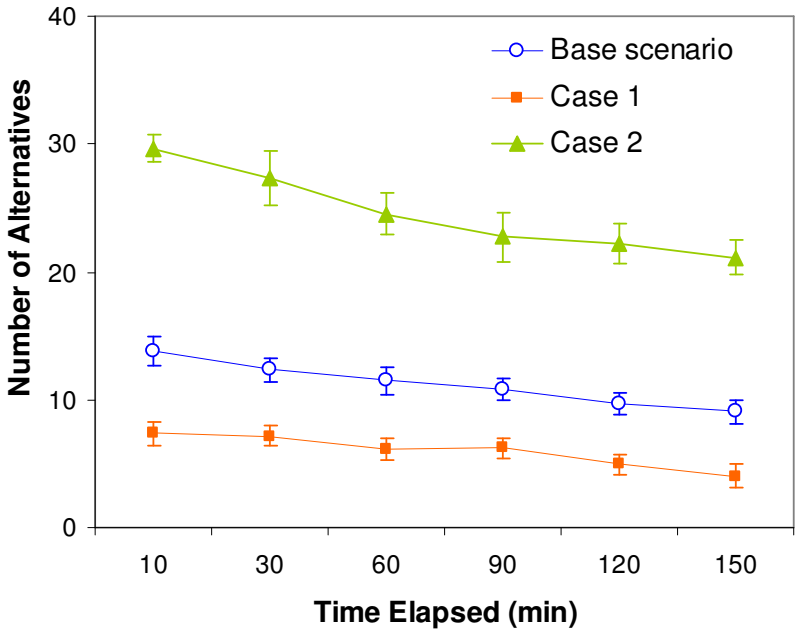


Figure 3.6 Effect of the quantity and quality of monitoring data on the number of alternative solutions.

The behavior of the ADOPT-based approach to different contamination events with various contaminant characteristics was then investigated. Four contamination events were simulated, differing in the injection location, starting time, duration, and/or mass injection rates. A description of the four scenarios is provided in Table 3.2. Figure 3.7 shows the injection locations of the four scenarios as well as the base scenario.

Table 3.2 True Source Description for Four Contamination Event Scenarios

True Source	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Location	Node 125	Node 105	Node 267	Node 171
Starting time	2:30 a.m.	3:00 a.m.	3:00 a.m.	2:30 a.m.
Mass Loading (g/min)	{15, 20, 25}	{10, 20, 25, 25, 15, 10}	{10, 20, 25, 25, 15, 10}	{5, 20, 25, 20, 15, 20, 15, 10, 10 }
Duration (min)	30	60	60	90

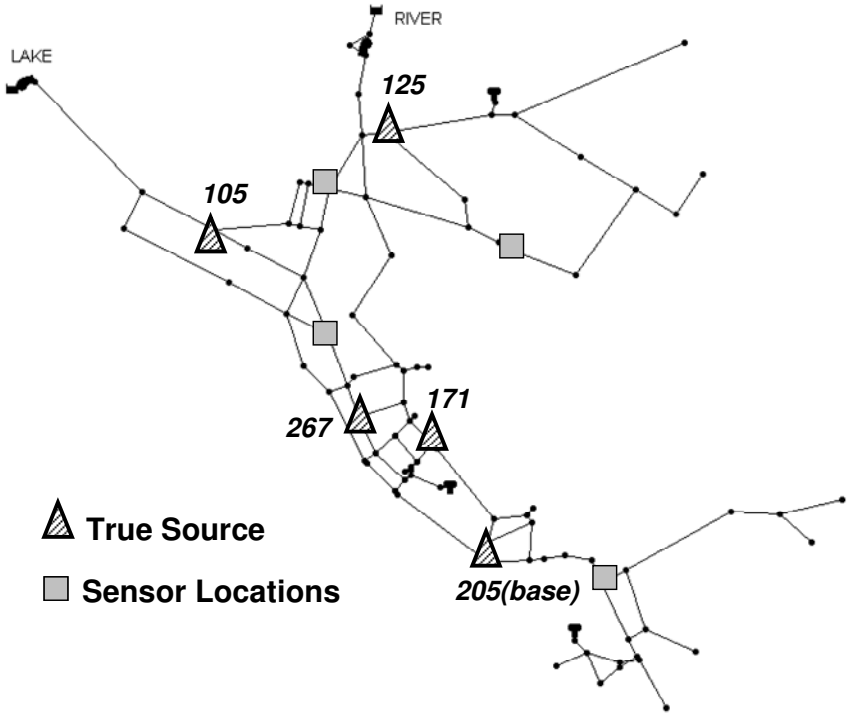


Figure 3.7 Four hypothetical contamination scenarios along with the base scenario: Contaminant source is indicated by the triangles; the number adjacent to the node represents the node ID; and the squares designate sensor locations.

The results of the four scenarios compared to the base run are shown in Figure 3.8. The observations were sufficient to eliminate the non-uniqueness for the scenarios with the contaminant source at nodes 125 and 105 (Scenarios 1 and 2, respectively). In contrast, a higher level of uncertainty was found for the other scenarios, which is likely due to the true source location in relation to the sensor locations in the network. It is also noted that for the contaminant introduced at nodes 125, 105, and 267 (Scenarios 1, 2, and 3, respectively), the true source node is always identified as one among the alternative set, whereas for the source at nodes 205 (scenario 2) and 171 (scenario 4), the true node is not always among the alternative solutions. With regard to the distribution of the resulting non-unique solutions for all scenarios, most of the solutions are concentrated in the vicinity of the true source, which may be due to the hydraulic similarities among their vicinities. These alternatives gradually move towards the true source node as more measurements become available.

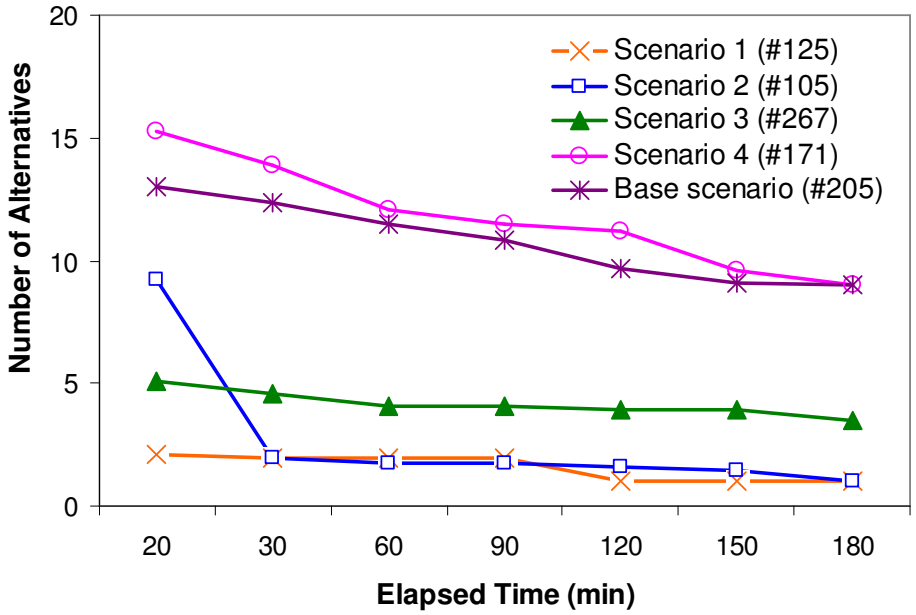


Figure 3.8 Comparison of ADOPT results among various hypothetical contamination events.

3.4.2 Micropolis Example Network

To understand the impact of the increasing level of problem complexities on the alternatives identified by ADOPT, a second network was studied. This relatively large network consists of 1574 junctions, 2 reservoirs, and 1 tank, as illustrated in Figure 3.9. This example was developed for a virtual city with 5,000 residents, further details of which can be found in Brumbelow et al. (2007). The locations of five ideal sensors are randomly selected within the entire network (see Figure 3.9). The contaminant transport is simulated at 10-minute intervals, and the concentration values at perfect sensors are assumed to be observed at each 10-minute increment.

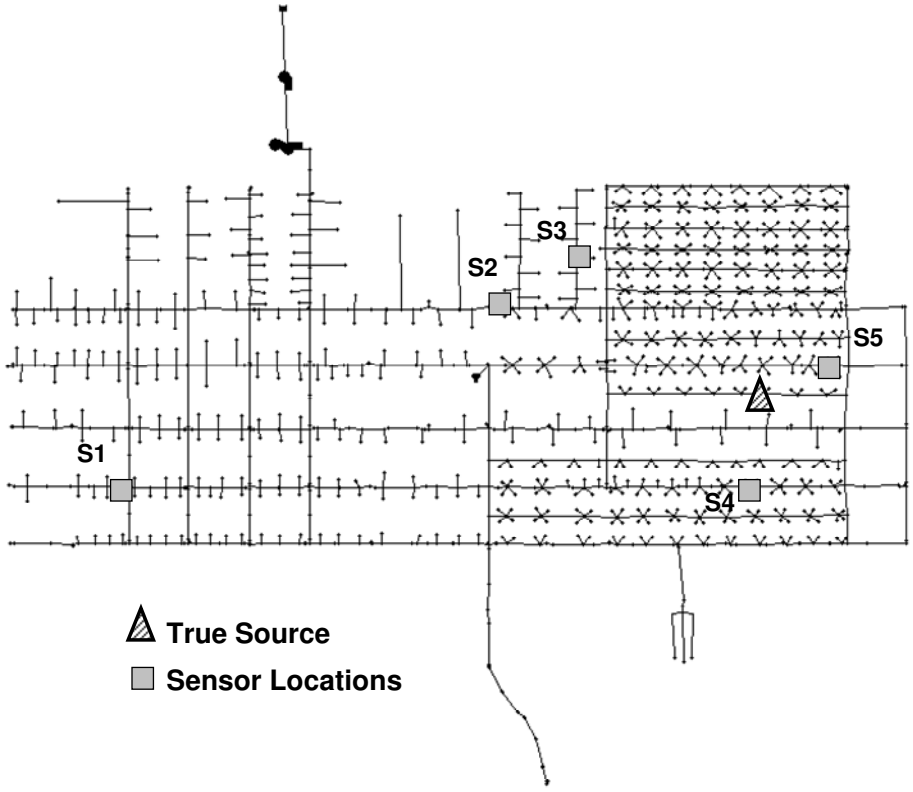


Figure 3.9 Micropolis water distribution network schematic. The triangle indicates the contaminant source, and the squares designate sensor locations.

A conservative contaminant is assumed to be introduced at an intermediate node (labeled as IN 1646) starting at 12:00 p.m. (after a 12-hour simulation) for 1 hour. The contaminant is injected at the rates of 10, 25, 10, 20, 28, 17 g/min, each of which corresponds to a 10-minute interval. The allowable ranges of source parameters are described in Table 3.1. After several experiments, the ADOPT parameter values were specified as follows: the number of subpopulations is 40, and each subpopulation contains 100 parents and 100 mutants. ADOPT is executed for 20 generations at each observation interval. For this event, the first detection occurred at 6:30 p.m., and the observation lasted until 9:00 p.m.. Applying the aforementioned parameter settings, the computation time for each trial was approximately 75 minutes on the Neptune system (an Opteron cluster) where EPANET simulations were executed in parallel on twenty processors. ADOPT was executed for 30 random trials in consideration of the randomness of the search process. The results from a typical run are summarized in Figures 3.10, 3.11 and 3.12.

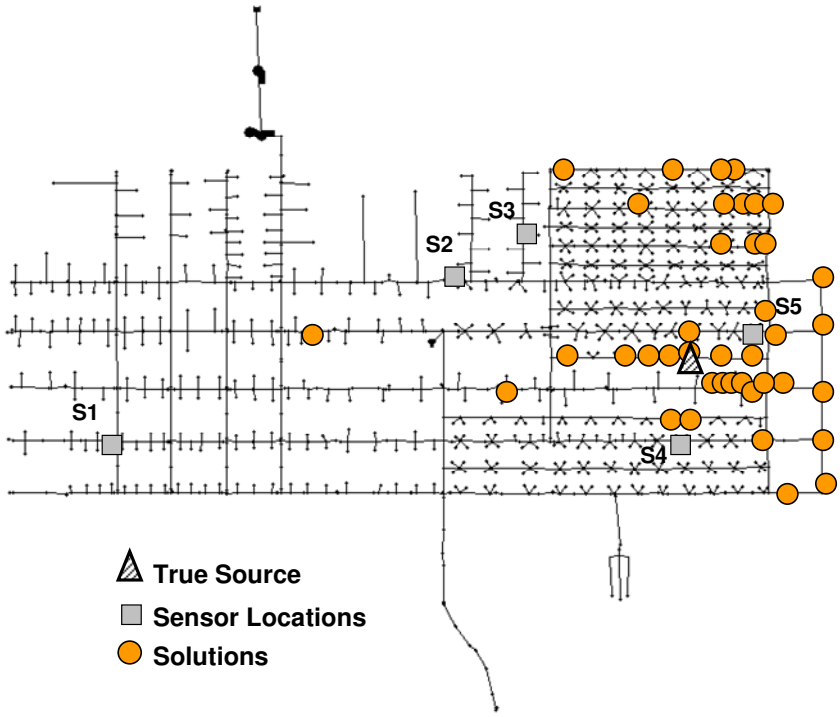


Figure 3.10 Identified solutions at 6:50 p.m. for the hypothetical contamination event in the micropolis network.

Figure 3.10 illustrates the locations of the 39 alternatives generated after taking 20-minute interval measurements. It can be observed that the 12 alternative solutions that remain after over two hours of elapsed time are very close to the true source location, as displayed in Figure 3.11. This observation reflects that the degree of non-uniqueness diminishes with additional observations. Figure 3.12 clearly shows that the calculated profiles of the 12 alternatives correspond well with the observed profile at sensor S3. No abnormal measurements were taken at the other sensors.

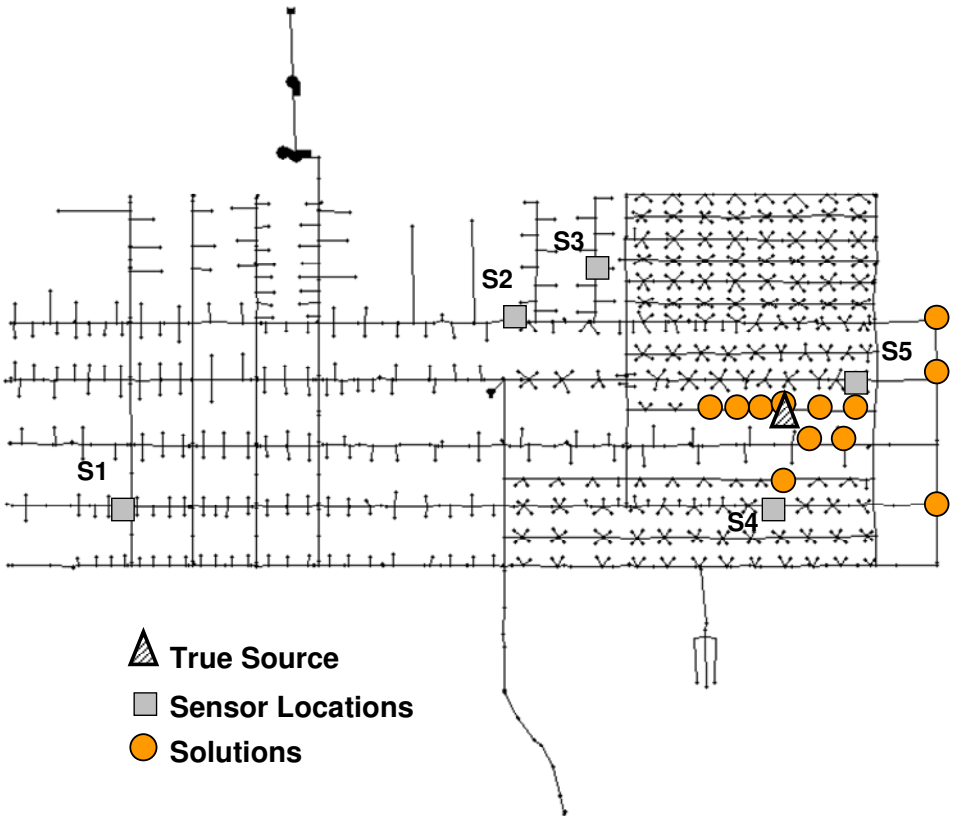


Figure 3.11 Identified solutions at 9:00 p.m. for the hypothetical contamination event in the micropolis network.

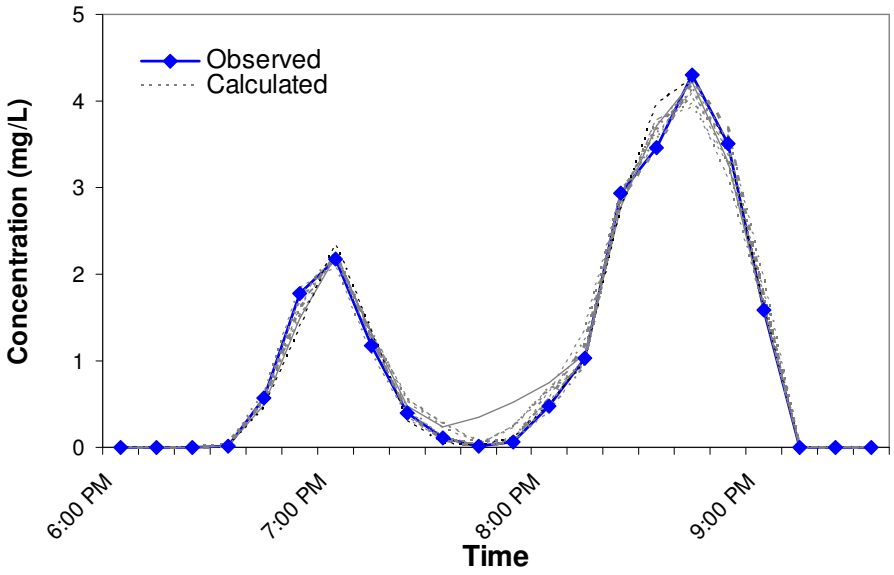


Figure 3.12 Comparison of concentration profiles between the observed and calculated profiles at S3 for the hypothetical event in the Micropolis network.

From the results of the base scenario in the small network and beginning with a given number of subpopulations, multiple subpopulations efficiently converge to the optimal and alternative solutions. Eventually, the number of remaining subpopulations tends be similar to that of the alternatives. For comparison, Figure 3.13 presents the number of subpopulations and alternatives that vary over time for this example and are averaged over 30 random trials. The results indicate that after the first set of measurements, the remaining subpopulations are much more numerous than the alternatives due to insufficient convergence in the search. This discrepancy in numbers also illustrates the necessity to enhance the algorithm's efficiency in the face of increasing complexities and limited computational resources.

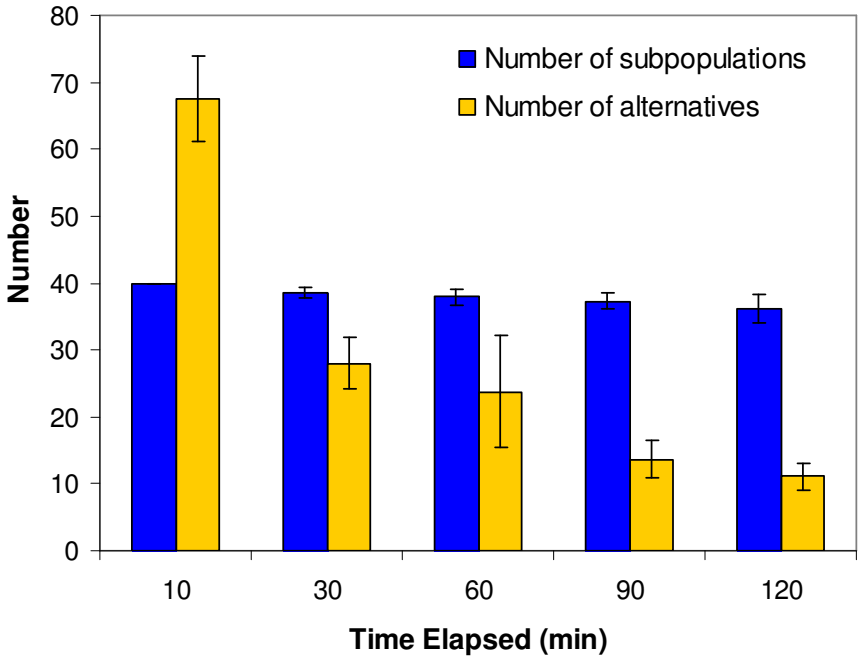


Figure 3.13 Comparison between the number of subpopulations and alternatives for a hypothetical event in the micropolis network

3.5. Summary and Discussion

In this work, an adaptive dynamic optimization technique for contaminant source identification has been presented and is structured to search for a set of possible solutions by exploring the decision space through multiple subpopulations. Several investigations address the efficacy of the proposed ADOPT. These explorations were designed to discover many aspects of the problem and solutions, including the degree of non-uniqueness, various problem complexities, and the range of the ADOPT algorithmic parameter settings.

This study suggests that ADOPT works adaptively to determine multiple alternatives that match the observations available at any time during the contamination event. Specifically, the level of non-uniqueness assessed by ADOPT diminishes as the number of sensors, measurement quality, or observation period length increase. The use of the time-varying relaxation target specifies alternatives according to sensor observations. The relaxation value is designed to be adaptive, and is determined through several experiments. Because of infinite possible contamination scenarios, the identification uncertainty of ADOPT is subject to the effect of the relaxation setting. Imposing a high relaxation value unavoidably slows down the search, thus causing unnecessary computational resources. Conversely, a low relaxation value may result in premature convergence, potentially skipping over the true source during the search. Future work is necessary to investigate the relaxation setting to eliminate the impact on solutions irrespective of contaminant characteristics.

In addition, the results show that varying the initial number of subpopulations produces similar results, suggesting that ADOPT is insensitive to the initial setting of the subpopulations, given that the number of subpopultions is greater than that of possible

solutions. It is also observed that more generations improve the solution quality in terms of the prediction error, but more generations may lead to over-convergence during the early stages. Indeed, the setting of the generation number relies on both the size of the water network and the allowable range of the decision variables.

In the context of a contamination event, contaminant source recovery involves the injection location as well as the release history. The characterization of the release history can be represented as a large array of decision variables that contribute to the difficulty inherent to solving such a problem. Effective fine-tuning would be helpful for the algorithm to exploit localized structures. Although the self-adaptive mutation strategy in the ES is a great benefit to the refining process, a changing environment necessitates the reconstruction of the mutation step sizes. The presented results suggest that the dynamic re-initialization strategy is increasingly advantageous to the search as time goes on.

Although ADOPT is shown here to be successful in a number of applications, the example problems solved are based on assumptions, such as the single injection location, known demands, conservative contaminants, etc., which are somewhat impractical in a realistic scenario. Future work will extend this algorithm to handle problems of increasing complexity and facilitate the application to actual WDS contamination events. Another issue associated with ADOPT is its degree of efficiency, especially for application to a large network where simulation models are time-consuming. The research team intends to investigate enhancements in the algorithm's efficiency. For example, to rule out unnecessary nodes and reduce the search space, a prescreening technique employed prior to the ADOPT runs may be valuable.

CHAPTER 4: Logistic Regression Analysis to Estimate Contaminant Sources in Water Distribution Systems

Abstract. Accidental or intentional contamination in a water distribution system (WDS) has recently attracted attention due to the potential hazard to public health and also due to the complexity of the characteristics and possible solutions of the contamination itself. The accurate and rapid identification and characterization of the contaminant sources are necessary to successfully mitigate the threat in the event of contamination. The uncertainty surrounding the contaminants, sensor measurements, and water consumption underscores the importance of a probabilistic description of possible contaminant sources. This chapter proposes a rapid estimation methodology based on logistic regression (LR) analysis to estimate the likelihood of any given node as a potential source of contamination. This methodology dynamically processes the concentration data gathered from monitoring stations within the system. Not only does this algorithm yield location-specific probability information, but it can also serve as a prescreening step for simulation-optimization methods by reducing the decision space and thus alleviating the computational burden. The applications of this approach to two example water networks show that this method can efficiently rule out a large region that does not result in contaminant observations. This elimination process narrows down the search space of the potential intrusion area. The results also indicate that the proposed method efficiently yields a good estimation even when some noise is incorporated into the measurements or demand values at the consumption nodes.

4.1. Introduction

The vulnerability of drinking water during the transport process within a WDS has received much attention in recent years. Contamination, either accidental or intentional, is a major issue associated with the security of drinking water quality. In order to detect contaminants, a WDS must have a set of sensors installed that can respond to a contamination event or to the threat of a contamination event. However, the installation and operational costs limit the large-scale use of monitoring sensors in a WDS. Thus, given the resultant sparse monitoring data, and that these data are either real-time or near real-time, the rapid identification of pollutant sources is necessary in order to process water quality samples efficiently and adequately. Solutions are needed to generate an effective threat management strategy that can mitigate the threat by taking appropriate actions, such as warning the impacted residents, isolating malicious contaminant sources, and flushing out the contaminant.

However, contaminant source characterization is complicated not only by the limited observation data, but also by the arbitrary nature of the contaminants that potentially can be injected from any point accessible to the public and with varying levels of strength. Based on sensor observations, this characterization problem can be categorized as an inverse problem. The complexity of real inverse problems, coupled with limited available data, typically yields ill-posed solutions, including *non-existence*, *non-uniqueness*, and *instability*. *Non-existence* refers to no solution, given the available observations. *Non-uniqueness*, caused by insufficient data, refers to different solutions that are identified in order to give similar explanations for the observations. *Instability* refers to inverse solutions that are sensitive to

the observations. Thus, in the context of a WDS contamination event, the dynamic nature and uncertainties of the system and the need for rapid characterization contribute to the complexities and difficulties inherent to the contaminants and their sources.

This contaminant source identification and characterization problem has captured the interest of researchers. Previous efforts have concentrated on characterizing the contaminant by constructing it as an optimization problem (e.g., Van Bloemen Wannders et al., 2003; Laird et al., 2005; Guan, 2006). These optimization approaches include direct methods and simulation-optimization approaches. In Van Bloemen Wannders et al. (2003), a standard successive quadratic programming tool is applied to solve a small-scale problem. Laird et al. (2005) present an origin tracking algorithm to address the inverse problem of contamination source identification based on a nonlinear programming framework. Taking advantage of a simulation-optimization approach, wherein a search procedure is coupled with a simulation model, Guan (2006) demonstrated its applicability to nonlinear contaminant sources and release-history identification. The limitations of the previous work often include the inability to address non-uniqueness, deterministic characterization, small network size, and high computational costs, etc. This study, then, recognizes the importance of the ability of decision makers to attain a probabilistic description of contaminant sources in the event of a contamination. In this chapter, a statistical model is used to predict the likelihood that a given node is the contaminant source. The estimated probability values attempt to present an overall explanation for water quality observations under various uncertain circumstances.

While the knowledge of an existing water network and its sensor placements allows simulations of various hypothetical contamination events, the relationship between

contaminant source characteristics and their resulting sensor observations may be pre-established through the simulation of a large set of potential contamination events. This chapter proposes a rapid contaminant source estimation algorithm by determining an estimation model that is built upon a large number of simulations and offers a probabilistic depiction of contaminant sources once abnormal detection occurs. This approach is expected to reduce on-line computational costs and statistically characterize contaminant sources based on the currently available concentration data collected from the sensors. The use of the developed method is demonstrated for both a small and a large network.

4.2. Problem Description

Numerous possible injection scenarios, unknown water consumption at any given node, and errors inherent to measurements and models contribute collectively to the high degree of uncertainty in a WDS contamination event. Because of these uncertainties, it is essential to provide a statistical depiction of the possible contaminant sources. Although a contaminant source is typically characterized by its location and corresponding mass loading history, the location is also a major concern in the context of real-time response. To enable a practical and viable investigation, this study concentrates on possible source locations in order to obtain a reasonable estimation quickly. The sensor observation data, from the first detection to the current time, are used as estimation model inputs to pinpoint the likelihood of any given node as a contaminant source. The contamination event is a dynamic process in which the set of observations changes. Thus, the estimation of the likelihood that any given

node is the contaminant source must be updated according to the varying number of sensor observations.

4.3. Logistic Regression Analysis for the Rapid Determination of a Contaminant Source

4.3.1 Logistic Regression (LR) Analysis

A logistic regression model (LRM) (Hosmer and Lemeshow, 1989) can be used to estimate the probability of the presence of an event, given information about predictors that can potentially influence the outcome. As a class of generalized linear models, LRMs are distinguished from ordinary linear regression models by the range of their predicted values, the assumption of the variance of the predicted response, and the distribution of the prediction errors. The general LRM formulation is

$$\log(\frac{p}{1-p}) = b_0 + bX \ , \tag{4.1}$$

where p represents the probability of a response of 1; $\{b_0, b\}$ are the regression coefficients; and X is a vector of the k explanatory variables. In the above formulation (Eq. 4.1), the term $\log(\frac{p}{1-p})$ is called a logit function, which is used to transform the predicted value between 0 and 1 to a response ranging from -∞ to +∞. This mathematical formulation assumes that a linear relationship exists between the logit function and the predictors.

LRMs have been used successfully in the field of water resources as predictive models to obtain categorical forecasts or estimates. The strength of a LRM lies in its ability

to directly provide a categorical forecast with low computational costs. The implementation of LRMs is simple and flexible in comparison to some other predictive methods. Regonda et al. (2007) obtained categorical probabilistic forecasts from a LRM using a large-scale climate predictor to estimate the probability of the leading mode of a basin stream flow above a given threshold. Also, Lu et al. (2006) investigated the use of a LRM in the relationship between the presence of dehalococcoides DNA in groundwater from monitoring wells and the values of selected biogeochemical parameters.

4.3.2 Model Construction

A linear LRM-based approach is employed to model the likelihood that any given node is the contaminant source injection location, and is driven by observed data obtained from sensors. The appropriate inclusion of the predictors is a major challenge, particularly in the model construction. With respect to model stability, the criterion of predictor selection can minimize the number of predictors, whereas incorporating more predictors into the model aids in an overall understanding of the problem. Unfortunately, a large number of predictors may result in the over-fitting of the model. Because a contaminant may be introduced arbitrarily into a network, the randomness of the contaminants and the resulting water quality data also pose challenges to the model construction. Given these considerations, in order to predict the possibility that any given node is the source at time t, a model is constructed using the observations at the current time as predictors. This model construction approach yields one model for one node at one measurement time step. Thus,

the total number of models for the whole network is the number of potential source nodes multiplied by the number of time steps for observation.

The following mathematical formulation is defined to determine, at time *t* after the contamination is detected at one or more sensors, the probability that node *i* is a contaminant source location based on the observation at time *t*. Here, it is assumed that *N* sensors are distributed throughout the entire system.

$$\pi_{it} = \log\left(\frac{p(A_i \mid C_{1t}, \cdots, C_{Nt})}{1 - p(A_i \mid C_{1t}, \cdots, C_{Nt})}\right) = b_{0t} + b_{1t}C_{1t} + \ldots + b_{jt}C_{jt} + \ldots b_{Nt}C_{Nt}, \tag{4.2}$$

where $p(A_i \mid C_{1t}, \cdots, C_{Nt})$ 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*; (b_{0t}, …, b_{Nt}) are regression coefficients obtained by the maximum likelihood procedure; and (C_{1t}, …, C_{Nt}) are the sensor observations at time *t*. From this formulation, the probability that node *i* is the source location can be calculated from the observed concentration at time *t* as

$$p(A_i \mid C_{1t}, \cdots, C_{Nt}) = \frac{\exp(\pi_{it})}{1 + \exp(\pi_{it})}, \tag{4.3}$$

The way in which the model is constructed may alleviate on-line computational burdens and produce a fast estimation while the contamination event is occurring, although the process of model-building unavoidably requires a large number of off-line simulation runs in which EPANET is used.

Ideally, it is expected that the LRM can identify the true source node with the greatest probability value compared to other nodes in the network. Several factors potentially impact the accuracy of the probability estimates, including the precision of the measurements,

hydraulic variability, the degree of non-uniqueness (as multiple locations could potentially yield similar observations at the sensors), and assumptions of linearity in the regression function form that may be resolved by dividing one model into several to fit the observation data at different levels. Nevertheless, the estimated likelihood values are expected to be favorable in creating an effective control strategy in the event of contamination. Additionally, this analysis can serve as a prescreening step for some other methods, such as heuristic searches, to discover the optimal mass loading profiles at potential nodes.

4.3.3 Data Generation

To develop the LRMs as described, first a large set of contamination realizations must be generated to represent the sensor observations at various intervals in response to possible contamination events. Each realization must include at least one non-zero sensor observation. Contaminants vary according to the injection location, starting time, duration, and mass injection rates, which are randomly selected from a uniform distribution and bound by specified values. Accordingly, a large set of sensor measurements is produced using EPANET simulations for the randomly generated events; these measurements are then used as inputs for developing the LRM. During the training of the LRM for each node, the probability value (i.e., the output of the LRM) is assigned a value of 1 (or 0) if the contamination occurs at this location.

4.3.4 Performance Evaluation

To assess the performance of the developed LRM, a validation dataset needs to be generated as well. Using the data generation approach described, a different set of realizations is created for validation purposes. For the initial investigations into the approach's applicability, the following three performance evaluation criteria are used: 1) the frequency of the true source location with a non-zero probability that it is a source location; 2) a cumulative distribution function of the number of possible nodes among a large set of realizations; and 3) the frequency with which the true source location is identified as the most likely source of contamination based on the LRM predictions.

4.4. Applications and Results

In this Section 4.4, two water networks with different levels of complexity help demonstrate the predictive potential of the LRM. In the small network, a large number of hypothetical contamination events are examined to reveal the identification ability of the LRM. This investigation is furthered by varying the number and quality of the measurements as well as giving consideration to water consumption uncertainty. In the micropolis network application, the development of the LRMs, considering their similarities between two consecutive time steps, is assessed.

The hydraulic and water quality simulations are executed by running EPANET during the generation of the dataset. The hydraulics remains at a steady state during hourly simulations. As a preliminary study, the conservative contaminant is assumed to be injected at a single location, where the hydraulic conditions are known. Although the varying

parameters that are used to create numerous scenarios serve as the characteristics of the contaminant sources in this study, the suggested approach can be extended further to incorporate system uncertainties when building the models.

4.4.1 Small Example Network

The first study uses a small network, which is one of the problem scenarios available as a tutorial within EPANET (Rossman, 2000). This network consists of 97 nodes, 2 sources, 3 tanks, and 117 pipes. The configuration of the network is depicted in Figure 4.1, and further details can be found in the EPANET user's manual. The contaminant transport is simulated in 10-minute intervals, and the concentration values at the sensors are observed at 10-minute increments.

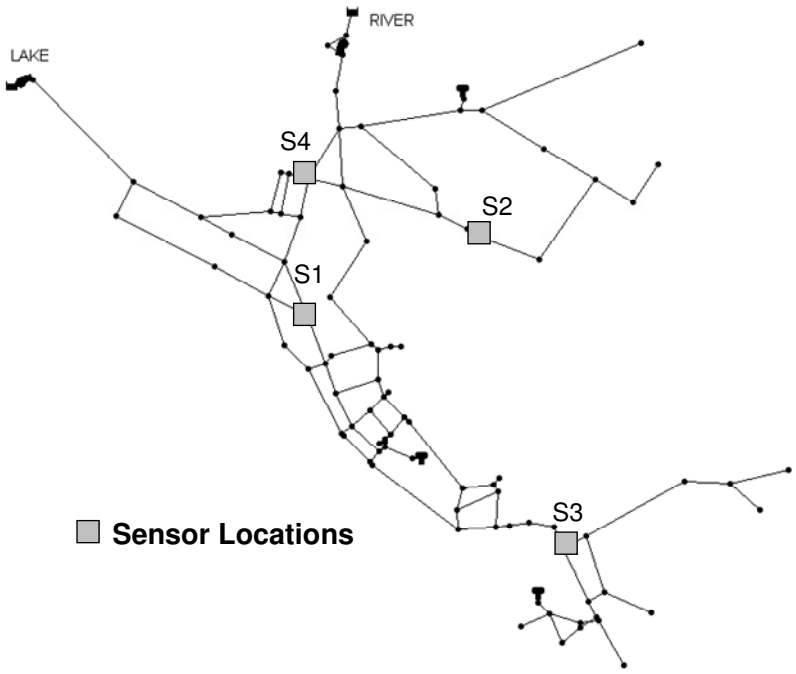


Figure 4.1 Water distribution network schematic (small network example). Squares designate sensor locations. For Scenario 1, the sensor network is composed of S1, S2, S3, and S4. Scenario 2 includes only sensor S3

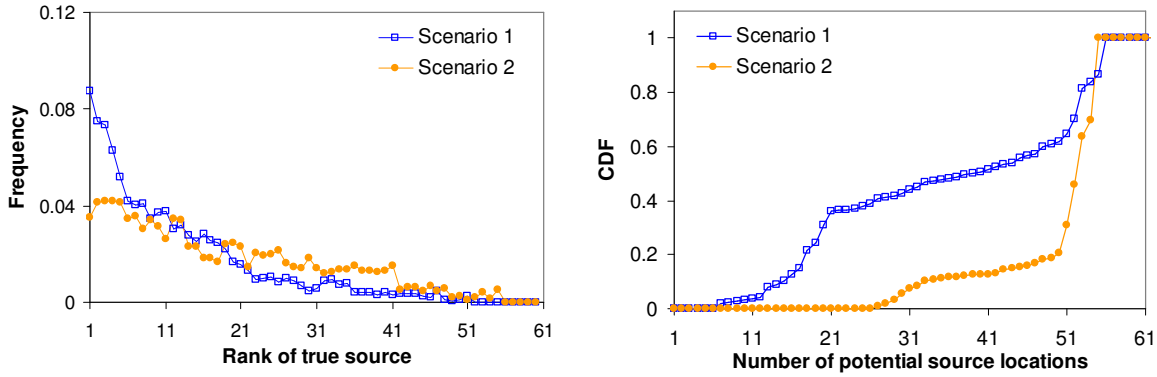
To demonstrate the algorithm's performance, a set of LRMs that corresponds to each node at each time interval is built upon the generated training data sets. Table 4.1 lists the parameters and their values that are used for the simulations of the hypothetical events. Here, LRMs that span 12 hours, which corresponds to 72 time steps (each represents a 10-minute interval), are chosen for the investigation. The maximum likelihood estimation method implemented within MATLAB is used to estimate the LRM coefficients.

Table 4.1 Contaminant Source Parameters and Ranges for Generating Training Data Set

Source Parameter	**Small Network**	**Micropolis Network**
Location	Any node (1~97)	Any node (1~1577)
Starting time	Within simulation 24 hours	Within simulation 48 hours
Duration (hrs)	0~6	0~6
Mass injection rate (g/min)	0~100	0~100

Number of monitoring sensors. In this section, the predictive capability of the LRMs is examined according to the varying number of observations. In addition to determining whether the true source is recovered as a potential solution, the relative rank of the true injection location compared to that of other nodes is evaluated, and the number of potential solutions is determined. The two scenarios employed here consist of four and one sensor(s), respectively. The locations of the selected sensors are shown in Figure 4.1. Scenario 1 is composed of sensors S1, S2, S3, and S4, and is also used for subsequent discussions, and Scenario 2 corresponds to the observations obtained from sensor S3. The same set of contamination realizations is used to achieve a meaningful comparison. Each realization contains at least one non-zero concentration data point, and the total number equals 1,000 at each time interval.

The generated results of both scenarios indicate that the established LRMs are capable of recovering the true source node as a potential solution, with an estimated non-zero probability. For each realization, the node series is ranked according to the calculated probabilities in descending order, whereby the node with the largest value is ranked first. Frequency, as a function of the rank of the true source location, is illustrated in Figure 4.2(a). Overall, a high frequency corresponds to a top rank for the true source node. Indeed, the frequency trend highly depends on the amplitudes of the parameter variations in the model-building. Compared to Scenario 2, the higher number of measurements in Scenario 1 improves the rank of the true source nodes. Further, the estimation uncertainty is measured as the number of potential solutions. The cumulative distribution function (CDF) of the number of potential solutions is shown in Figure 4.2(b). By incorporating the additional measurements of Scenario 1, the probability of identifying a smaller set of possible solutions increases, indicating that a large number of observations aids in reducing the uncertainty of the solutions.



(a) (b)

Figure 4.2 Comparison of performance of LRMs between Scenario 1 and Scenario 2: (a) rank of true source location vs. frequency; (b) CDF of number of potential source locations.

Impact of measurement errors and demand uncertainties. However, poor performance may occur as a result of the errors that are related to the contaminant source estimation, such as imperfect measurements or uncertain amounts of water consumption. To understand the effects of these uncertainties on LRM solutions, a normally distributed white noise was added to each factor. The mathematical formulation for modeling these factors with perturbation is expressed as

$$y_{it}^{err} = y_{it} + \alpha * y_{it} * N(0,1), \tag{4.4}$$

where y_{it}^{err} denotes the perturbed measurement of sensor i or the demand multiplier of node i at time t; y_{it} denotes the true measurement of sensor i or the demand multiplier of node i at time t; and α represents the error level added to the perturbed factor.

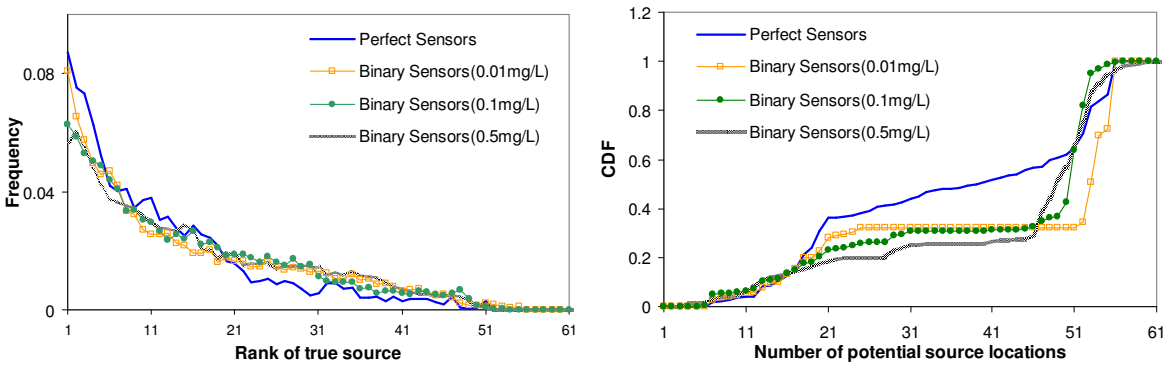
Table 4.2 Summary of Results under Various Uncertain Conditions

Scenario	**Frequency of true source with non-zero probability**
Ideal condition	100%
Measurement error (α = 10%)	100%
Measurement error (α = 50%)	99.9%
Uncertain demand (α = 10%)	99.7%
Uncertain demand (α = 50%)	97.2%

The results of incorporating different levels of either measurement errors or demand uncertainties are summarized in Table 4.2. Making use of the same set of hypothetical events as demonstrated above, the performance is evaluated in terms of the frequency that the LRM predicts the true source as a potential solution. As shown in Table 4.2, the 50% uncertainty level in the measurements causes a small number (1%) of cases in which the true source is identified as a non-solution. This result implies that the performance is not subject to the effects of measurement errors. An explanation for this behavior is that a conservative

contaminant leads to a linear relationship between the sensor observations and the magnitude of the contaminant at previous times. The same level of uncertainty associated with water consumption, however, yields a poor performance. In reality, if the demands are highly uncertain, changes in the flow direction in the network may occur, thereby biasing the prediction of the LRMs developed under normal conditions.

Binary sensor condition. In addition to the above analysis that considers the chemical-specific sensor network, the LRM performance is further investigated using imperfect binary decision sensors in the system. In reality, the deployment of binary sensors allows water utilities operators to access merely the status of the contamination, which may be specified by the level of the water quality indicators (e.g., PH, chlorine, conductivity). For simplicity, a concentration level is set as the detection threshold. The observation data are converted to 1 (presence of contamination) if the reading exceeds this threshold, and 0 (absence) otherwise. Thus, a set of LRMs is built under these given conditions. Although the same set of validation data is used as demonstrated above, the set must be converted to 0/1 according to the given sensor detection limit. The performance results using this detection threshold are assumed to be 0.01, 0.1 and 0.5 mg/L, respectively, and are compared in Figure 4.3.



(a) (b)

Figure 4.3 Comparison of results between perfect and binary sensor conditions: (a) rank of true source location; (b) CDF of number of potential source locations.

As a result of the level of uncertainty introduced by the sensors, a lower detection limit increases the likelihood that the true source node ranks first (see Figure 4.3 (a)). The CDF, as a function of the number of potential solutions, is shown in Figure 4.3 (b) and suggests that a large number of unlikely nodes is eliminated when the detection limit is low. The frequency of the LRMs that recognize true locations as potential solutions exceeds 99.9% among all realizations, even with the detection limit of 0.5 mg/L. These results indicate the effectiveness of the LRMs in ruling out unnecessary nodes for the sensors with extremely coarse data.

Real-time updates of probabilities. During a contamination event, sensor monitoring data are collected dynamically as time progresses. Whereas the LRMs offer the capability to predict the probability given the observations at the current time, the time series observed data, from the first detection to the current time, can be used collectively to recover the source of contamination. A joint probability that a node is not the source can be specified as a product of the likelihood that this node is a non-solution through a sequence of time intervals. Specifically, if available observations together reflect the non-occurrence of contamination at

a node, it is concluded that this node is not the injection location. Thus, the probability that a given node is a source can be updated in real time as follows:

$$P(A_i \mid C_0, \cdots, C_t) = 1 - (1 - p(A_i \mid C_0))(1 - p(A_i \mid C_{t_0+1})) \cdots (1 - p(A_i \mid C_t)), \quad (4.5)$$

where $P(A_i \mid C_0, \cdots, C_t)$ represents the updated probability of the contaminant injected at node i at time t given currently available observations $\{C_0, \cdots, C_t\}$; A_i represents that contamination occurs at node i; $p(A_i \mid C_t)$ denotes the predicted probability of node i as a source using the observation C_t at current time t, which is estimated by LRMs; and t_0 refers to the first detection time.

A total of 1,000 contamination events is hypothesized to achieve statistical significance. The 95% confidence interval of updated probabilities and the rank of true nodes are used to indicate the level of reliability of the results, as listed in Table 4.3. As is the case with increased measurements, a longer observation period yields a greater likelihood that a given node is the contaminant source. For measurements taken up to three hours, on average the true node is determined as a solution with over 50% likelihood that it is the contaminant source, and the confidence interval is small. However, this occurrence does not mean that true source nodes must be increasingly dominant over other nodes with more measurements. As shown in Table 4.3, the rank of the injection node shows a slight increase with time. An explanation for this behavior is that more nodes become incorporated into the candidate set due to increasingly available measurements. This observation also indicates the complexity of the estimation, which results from the high levels of uncertainty associated with such a problem.

Table 4.3 Statistical Summary of the LRM Results that Correspond to the Injection Node

Elapsed Time (hrs)	**Confidence Interval (95%)**	
	Probability (%)	***Rank***
1	[33, 36]	[6.83, 7.92]
3	[56, 59]	[7.53, 8.69]
6	[70, 73]	[8.24, 9.47]
12	[77, 80]	[8.54, 9.79]

4.4.2 Micropolis Example Network

To confirm the effectiveness of the LRMs, a relatively large micropolis network is examined. In addition to studying the effects of the increased problem complexities on performance, a strategy to reduce the number of LRMs is evaluated. The configuration of the water network is depicted in Figure 4.5, and is composed of 1574 junctions, 1415 pipes, 8 pumps, 2 reservoirs, and 1 tank. This example was developed for the micropolis virtual city with 5,000 residents, further details of which can be found in Brumbelow et al. (2007). The locations of five sensors are randomly selected within the entire network (see Figure 4.4).

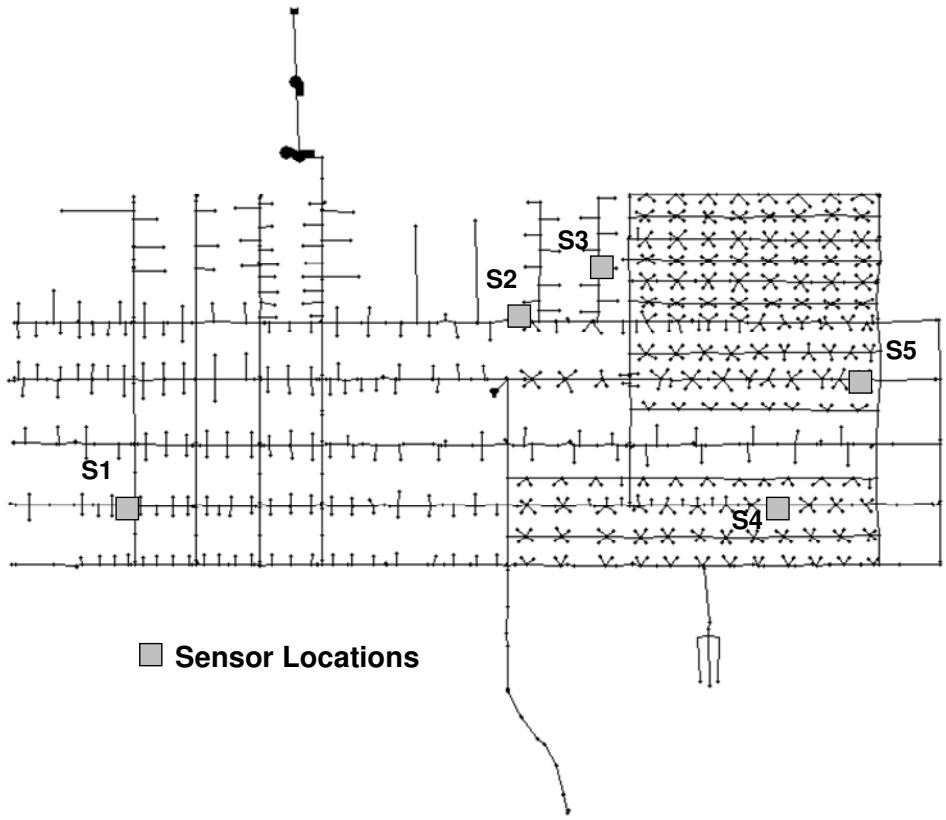


Figure 4.4 Layout of micropolis water distribution network. The sensor network is composed of S1, S2, S3, S4, and S5, denoted by squares

In a real network with a large number of nodes, building a large set of LRMs that corresponds to each node and spans a 1-day or 1-week period is too computationally intensive. Alternatively, some similarity may exist between a model at a current time step, say *t*, and a model at *t*-1. The need to regenerate a new model at *t* may require evaluation prior to model building. If a model at the previous time fits the current observation data, then it is used for the current time; if not, a new model that corresponds to the current time is created. This process is expected to reduce the computational costs during the model-

building process. The goodness of fit may be determined by the probability of the true source node over all realizations (e.g., ≥ a specified value).

The reuse of the LRMs is evaluated by comparing their performance against the independent model generation demonstrated above. Again, the LRMs that span 12 hours, thus representing 72 time intervals, are chosen for the investigation. The parameters and their values that are used for simulating hypothetical events are shown in Table 4.1. For evaluation purposes, a set of contamination realizations, containing 5,000 samples at each time interval, is generated. The first interval is selected as the time immediately after the one-day simulation period, which allows enough time for contaminants to reach the downstream nodes. A comparison of frequency as a function of the true source ranking between the reuse strategy and the independent model generation strategy is presented in Figure 4.5. Although the reuse strategy is slightly worse than the independent generation strategy in terms of the true source node ranking, both strategies capture the true source node as a potential solution among all realizations. It is important to note that the reuse strategy saves 92% of the computational costs during the model-building process.

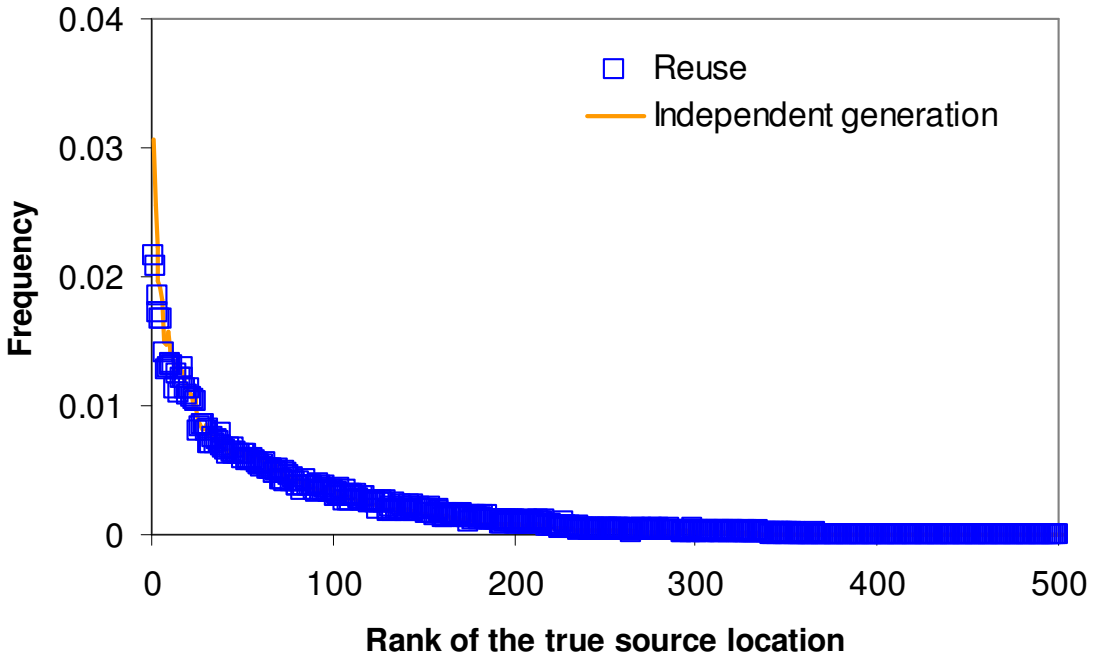


Figure 4.5 Comparison of results between two model-building strategies (micropolis network).

To examine the distribution of potential solutions obtained from the LRMs, one hypothetical contamination event is simulated. The contaminant, with a constant mass injection rate of 60 g/min, is introduced from an intermediate node (labeled as IN 1646; see Figure 4.6). The detection occurred initially at 12:30 p.m. and lasted until 1:40 p.m. at sensor S5. The locations of potential solutions determined by processing the first observation are illustrated in Figure 4.6. The LRMs identified 167 solutions out of 409 nodes that could contribute to observations at the given sensor locations. Figure 4.7 shows the top 50 solutions with respect to the estimated probabilities up to 1:40 p.m. It is worth noting that these possible sources are relatively close to the true source. For the given sensor network, however, a large set of unknown nodes exists, because such nodes are undetectable if the

contaminant is injected at these nodes. It is also noted that the true source location is one of the solutions across the observation period.

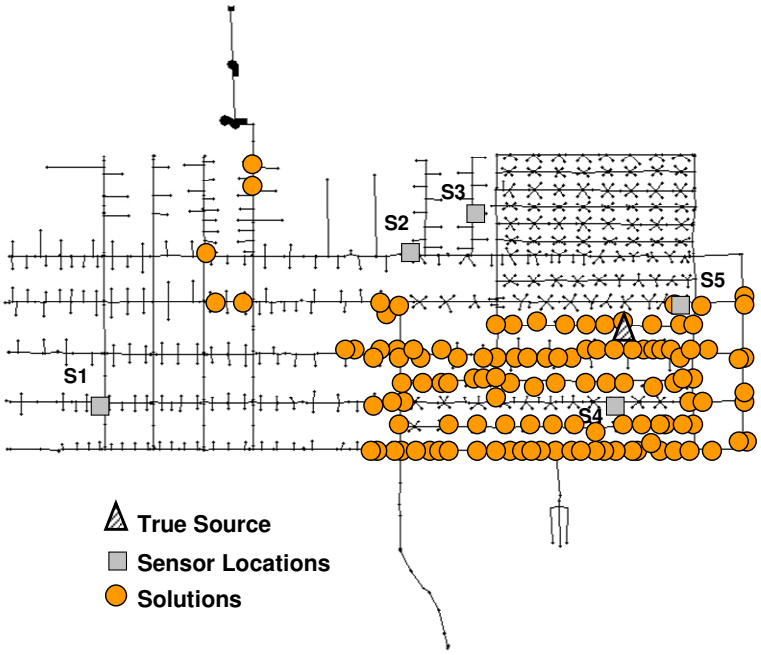


Figure 4.6 Locations of possible sources at 12:30 p.m.

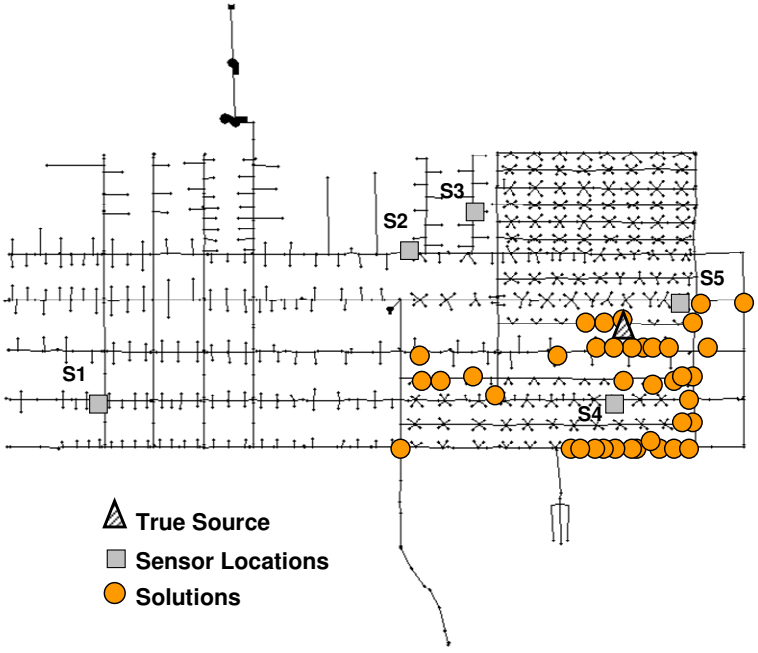


Figure 4.7 Locations of the top 50 solutions at 1:40 p.m.

4.5 Final Remarks

Identifying a contaminant source quickly is vital for creating an effective threat management strategy in the face of accidental or intentional contamination. A high level of uncertainty inherent to the contaminant and WDS complicates the characterization of the contaminant sources. The approach demonstrated here enables a probabilistic description of the source location that allows for various uncertainties associated with the contamination. While EPANET allows hydraulic and water quality simulations for a water network, a statistical model can be built to describe the contaminant as a function of available measurements using a large number of hypothetical contamination simulations. Together, rapid prediction and simple implementation support the use of LR analysis. Thus, in this work, the relationship between the likelihood that a given node is a source and the sensor observations is expressed by LRMs. A series of pre-established LRMs leads to a fast estimation once contamination is detected.

The developed LRMs are applied to two water networks through the simulation of numerous contamination events. The smaller network application focuses attention on examining the effects of the measured data as well as the demand variations on the solutions. A method to update the probabilities dynamically is proposed herein as well. The larger network application focuses on the development of LRMs, allowing for their similarities between two consecutive time steps.

From the results and analysis described, this approach is able to determine potential source locations. One of the resulting solutions is the true source node, even given coarse monitoring data. The obtained results indicate that demand uncertainty has a larger impact

than measurement errors due to the possibility of a change in flow direction. More measurements across space or time can improve the uncertainty of such a problem with respect to the rank of the true source as well as the number of potential solutions. The reuse of models can produce results that are comparable to those produced by the independent generation of models, with a significant reduction in computational costs.

Although the proposed approach facilitates a probabilistic characterization of each node in a contamination event, the other characteristics associated with the contaminant (e.g., injection starting time, duration, and mass flow rates) are underdetermined. Further work could consider the LRMs in combination with some other methods, such as heuristic search approaches, to enhance the characterization accuracy. Additionally, future work is required to extend this approach to a more realistic condition, such as the likelihood of simultaneous multiple injection locations, unknown hydraulic conditions, false positives and false negatives from sensor readings, etc.

CHAPTER 5. Contaminant Source Characterization using Logistic Regression Analysis and Local Search Methods

Abstract. Given a set of contaminant concentration observations obtained from sensors in a water distribution network, an inverse problem can be constructed to identify the contaminant source characteristics (including location, strength and release history) by coupling a water distribution simulation model with an optimization method. However, this approach requires numerous time-consuming simulation runs to evaluate potential solutions; so, determining the best solution or set of possible solutions within a reasonable computational time may be difficult. For this reason, it is desirable to reduce the decision space in which the optimization procedure must search in order to reduce the computational burden and potentially produce a rapid determination of the contaminant source characteristics. Previously, the authors proposed a method to reduce the decision space by efficiently identifying the probability of each point or demand node as a contaminant source location using mostly off-line computations. Then, the most likely source locations can be used as good starting points for local search (LS) methods to obtain the optimal injection profile(s) to match the observed concentration profile(s) over time. The proposed approach is demonstrated for a contamination source identification problem using two illustrative example water distribution networks.

5.1. Introduction

Rapid and accurate contaminant source characterization is critical for managing an accidental or intentional contamination event within a water distribution system (WDS). The

process of contaminant source determination involves the rapid characterization of the injection locations, starting time, duration and mass loading pattern based on limited data obtained from sensors over time. The basic purpose of this process is to quickly determine the contamination characteristics, but insufficient data and countless possible contamination scenarios challenge the characterization process in terms of both accuracy and efficiency.

Previous work on source identification problems in a WDS focuses on identifying optimal solutions by constructing optimization formulations that can be solved by several methods, including direct methods (e.g., Van Bloemen Wannders et al., 2003; Laird et al., 2005) and simulation-optimization approaches (Guan, 2006; Liu et al., 2006). Because of the discreteness, nonlinearity, and nonconvexity, as well as the limiting assumptions of such formulations, heuristic search methods have recently attracted attention. Nevertheless, computational efficiency remains of great concern because such methods require numerous time-consuming simulation runs to evaluate potential solutions. Also, it is especially difficult to obtain a good solution within a reasonable amount of computational time in a large network, even using parallel computing. Computational requirements may be reduced by using a prescreening technique that eliminates unfeasible solutions, thus reducing the decision space in which the heuristic procedure must search. One prescreening method is the back-tracking algorithm reported by De Sanctis et al. (2006). This method is able to identify only those source characteristics that explain the observations detected by the water quality sensors over time.

LR analysis is another potential prescreening approach that shows promise. This approach is investigated in Chapter 4 to identify possible source locations by estimating the

probability of each node (i.e., as the possible true source). The location-specific probability information is then used to limit the potential source nodes, thus reducing the search space for the heuristic search. In addition, the selection pressure in the subsequent heuristic search may be assigned to different regions of the water distribution network based on the probability that any given source location is the true source.

In this study, nongradient-based LS algorithms are considered as the heuristic search method. Although LS techniques are computationally efficient, the quality of the solution depends on the quality of the starting solution to these iterative search procedures. Although a good starting solution to the contamination source characterization problem is unknown in this study, it is expected that the smaller set (than all the other set of nodes in the network) of possible source locations identified by the LR approach will serve as a set of good starting solutions. Because only a small set of nodes is considered, applying the LS procedure to this set of locations is still computationally efficient and potentially good at identifying the best solution. Thus, the overall procedure investigated in this chapter consists of: 1) an LR analysis-based prescreening, and 2) an LS technique-based optimization. These components are described in the following subsections.

5.2. Logistic Regression Model (LRM)

Numerous potential contamination scenarios and system uncertainties contribute collectively to the complexity of source characterization in a contamination event. To offer a fast probabilistic estimation of potential source locations, a linear LRM-based approach has been reported in Chapter 4 to model the likelihood that any given node is a source, given the

sensor observations. The LRM, constructed as follows, describes the relationship between the probability of node *i* as a source and the observations at time *t* after the contamination is detected at one or more sensors.

$$\log\left(\frac{p(A_i \mid C_{1t},\cdots,C_{Nt})}{1-p(A_i \mid C_{1t},\cdots,C_{Nt})}\right) = b_{0t} + b_{1t}C_{1t} + \ldots + b_{jt}C_{jt} + \ldots b_{Nt}C_{Nt}\,, \tag{5.1}$$

where $p(A_i \mid C_{1t},\cdots,C_{Nt})$ denotes the likelihood that the contaminant is introduced at node *i*, given the observations at time *t*; A_i represents the contaminant entering through node *i*; (b_{0t}, …, b_{Nt}) are regression coefficients obtained by the maximum likelihood procedure; and (C_{1t}, …, C_{Nt}) denotes the sensor observations at time *t*.

While the knowledge of an existing WDS and its sensor placements allows simulations of contamination events, the LRMs can be pre-established through these realizations. Once the contamination is detected, the previously determined models can rapidly indicate the probability that the contaminant is introduced at a given node. Thus, the strength of the LRM is that it offers a simple and direct way to make a fast prediction with low computation costs. In this study, the resulting solutions from the LRMs are used to reduce the space of subsequent searches by eliminating unnecessary nodes that have estimated zero probabilities.

5.3. Local Search Approach

After identifying the set of possible contaminant source locations, heuristic search methods are used to determine the optimal characteristics of the contamination source. 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.

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

$$\text{Minimize } F = \sqrt{\frac{\sum_{t=t_0}^{t_c}\sum_{i=1}^{N_s}(C_{it}^{obs} - C_{it}(L, M_{t_c}, T_0))^2}{N_s * t_c}}, \tag{5.2}$$

where F = the prediction error;

L = the contamination source location;

T_0 = the injection starting time;

t_0 = the initial detection time of contamination;

t_c = the current time;

$M_{t_c} = \{m_{T_0}, m_{T_0+1}, \cdots, m_{t_c}\}$, 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} = the observed concentration at sensor i at time t;

$C_{it}(L, M_{t_c}, T_0)$ = the model (i.e., EPANET) calculated concentration value at sensor i at time t;

i = the sensor location;

t = the observation time; and

Ns = the total number of sensors.

This optimization problem is solved using a nongradient-based LS. Land (1998) and Hart (1994) 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.

Otherwise, due to the sensitivity to the initial starting points, various initial guesses may lead to diverse local optima. A detailed description of these search methods can be found in Belegundu and Chandrupatla (1999). The LRM-based prescreening for potential source locations is expected to provide a set of good starting points and reduce this sensitivity, such that the overall approach can robustly identify the best solution to this source characterization problem.

In this study, the Nelder-Mead Simplex (NMS) search method introduced by Nelder and Mead (1965) is employed to estimate the contaminant release history that corresponds to each possible node. Designed for unconstrained multidimensional optimization, the NMS method has been applied widely to various optimization problems due to its simple implementation in practice and the fact that no derivative information is required. 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). The six major steps of the NMS method are outlined as follows:

Step 1. Initialize the simplex constructed of the N+1 vertex, where N represents the number of decision variables. Then, evaluate each vertex in terms of the objective function.

Step 2. Obtain the indices of the worst, second worst, and best vertices. Calculate the centroid of the simplex by averaging the vertices, except the worst one.

Step 3. Generate a new vertex by reflecting the worst one that corresponds to the centroid. Accept the reflection if the newly generated vertex is worse than the optimum but surpasses the second worst one. Go to Step 4 and perform the expansion, if the new vertex outperforms the optimum; otherwise, go to Step 5 and perform contraction.

Step 4. Extend the search space when the reflection yields a new optimum by continuing in the same direction as the reflection. This process brings about a new vertex. Accept the expansion if the newly created vertex is better than the optimum; otherwise, accept the reflection by replacing the worst vertex with the new one generated by the reflection. Then return to Step 2 for the iteration.

Step 5. Perform an outside contraction if the reflected vertex is superior to the worst one but not as good as the second worst vertex. Accept this contraction if the contracted vertex performs better than the reflected vertex by replacing it with the worst one. Then return to Step 2 for the iteration; otherwise, go to Step 6 and perform a shrink transformation. Carry out an inside contraction when the reflected vertex is inferior to the worst vertex. Similarly, accept the contraction and go to Step 2 when the newly created vertex outperforms the worst one; otherwise, go to Step 6 and perform a shrink transformation.

Step 6. Attempt to move towards the optimum if the contraction described is unsuccessful. Then, go to Step 2 and continue with a new iteration.

5.4. Coupling the LRM with the LS for Contaminant Source Characterization

Figure 5.1 depicts the framework of the proposed LRM-LS method for WDS contaminant source characterization. The algorithm begins once contamination is initially detected at one or more monitoring sensors. Given the monitored data, the pre-established LRMs provide a quick estimation of the likelihood that any given node is the injection location. The estimated probabilities allow the set of potential nodes to be screened; the resultant set of possible nodes is then used as a basis for the subsequent NMS-based LS procedure. Characterization involves the search for source characteristics (i.e., contamination location, start time, duration, and mass loading profile). The LS operation for each candidate node at each time interval terminates when one of the following stopping criterion is met: 1) no improvement is made in the best solution in a few successive iterations, and the minimal prediction error is less than a specified value; and 2) the number of iterations reaches the maximal value allowed. At each time step the LS operation leads to a set of possible solutions that can match the sensor observations within a certain range of the prediction error. The process is repeated as sensor information becomes increasingly available. The set of possible solutions reduces in number as time progresses and eventually becomes a single solution, which is expected to be the true source when sufficient sensor information is available.

As a local optimizer, the NMS method focuses mainly on the exploitation rather than the exploration of the search space. This focus contributes to the sensitivity of its performance to initial starting points. It is imperative to introduce a certain level of diversity, particularly in the case of a changing environment. Thus, a re-initialization strategy is

incorporated into the NMS operation. When the search converges to the area with a high average prediction error or when new observed data emerge, the vertices in the simplex are randomly generated, except the optimum. This process, as expected, reduces the degree of sensitivity of the LS to the initial starting points and attempts to achieve both the exploitation and diversification during the search.

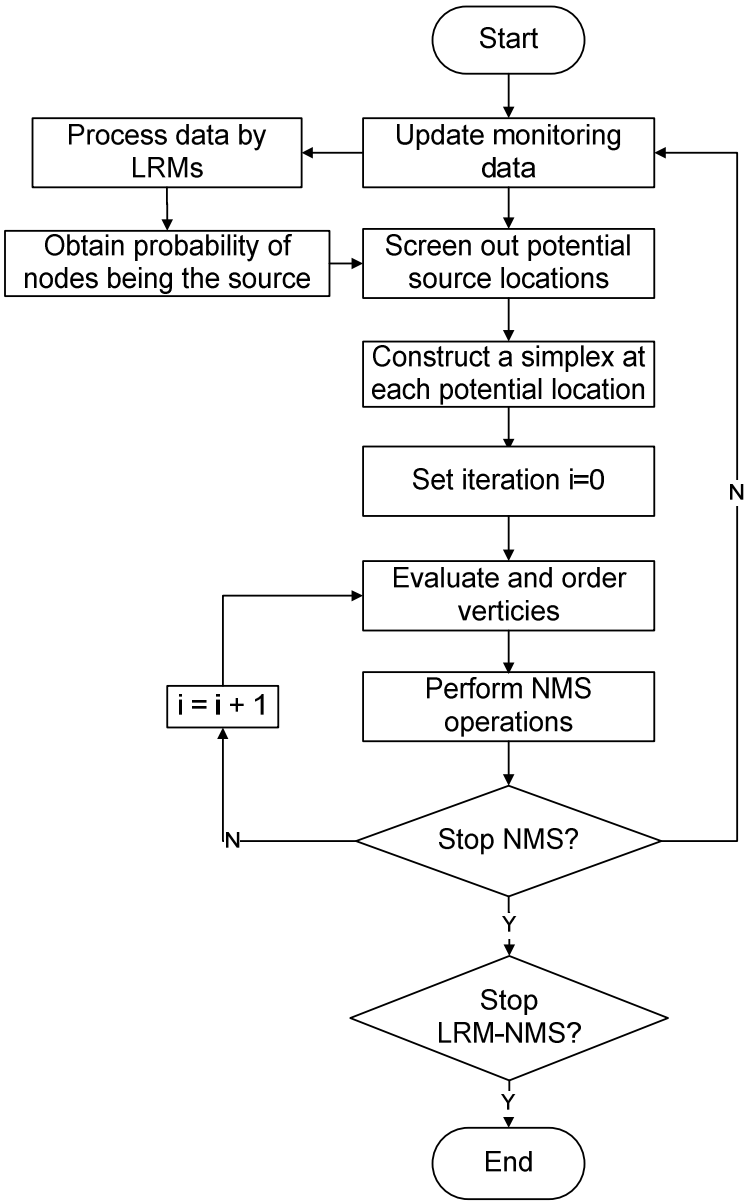


Figure 5.1 Flowchart of LRM-LS model for contaminant source determination in a WDS.

5.5 Case Studies

The application of the proposed LRM-LS model to two example networks is discussed in this section. The first example focuses on examining the performance and robustness of the LRM-LS method, given the variations in contaminant source characteristics as well as the amount and quality of the monitoring data. The second example, which is a relatively large network, illustrates the algorithm's applicability when facing greater difficulties.

Hydraulic and water quality simulations are generated by EPANET; it is assumed that the hydraulics are at a steady state and that the simulations are conducted in hourly intervals and the contaminant transport in 10-minute intervals. The synthetic sensor measurements are gathered at 10-minute increments. Contaminant characteristics are recovered on the assumption that a conservative contaminant is injected at one location under a known hydraulic condition. These assumptions are made primarily to enable a convenient and viable investigation into the proposed approach; however, they are not expected to limit the broader applicability of the approach to problems with conditions that deviate from these assumptions.

The effectiveness of the LRMs in identifying potential source locations has been demonstrated in Chapter 4. In this Chapter 5, the established LRMs are used directly to determine the candidate nodes, and the NMS method concentrates on the search for the optimal injection starting time, duration, and average mass injection rate at possible locations. Table 5.1 lists the allowable ranges of contaminant source parameters during the search.

Table 5.1 Allowable Ranges of Contaminant Source Parameters

Source Parameter	Small Network	Micropolis Network
Location	Any junction (1~97)	Any junction (1~1577)
Starting Time (hr)	0~15	0~30
Duration (hr)	0~5	0~5
Injection rate (g/min)	0~100	0~100

5.5.1 Small Example Network

The first network (Example 3 in EPANET), depicted in Figure 5.2, consists of 97 nodes, 117 pipes, 2 water sources (1 river and 1 lake), and 3 tanks. The base demands and the mean temporal variations of the 24-hour demand pattern assigned to all the nodes are assumed to be those described in the network input file for this network in EPANET. Four sensors are randomly located in the network.

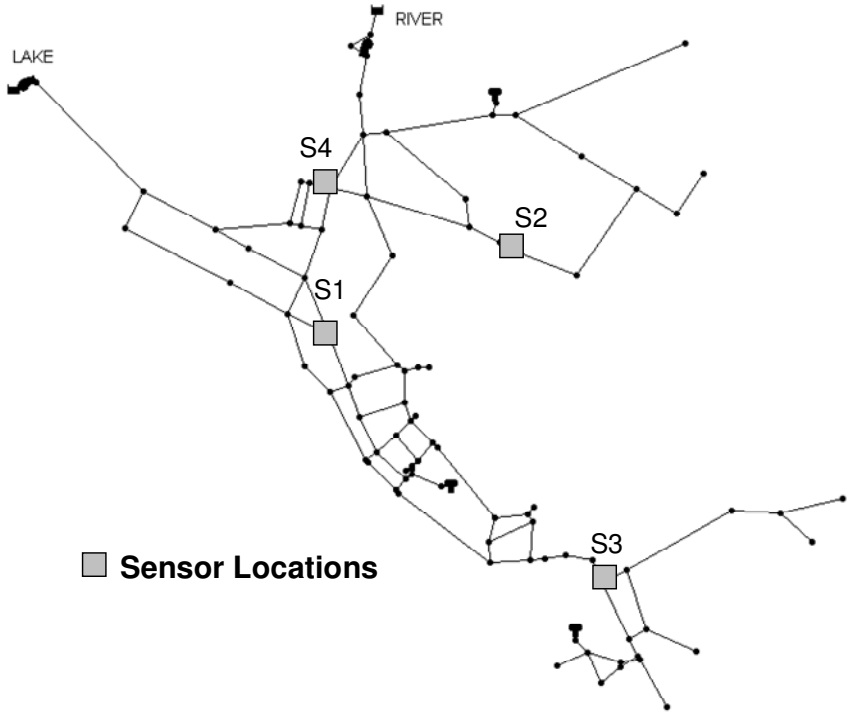


Figure 5.2 Layout of small example network. Squares designate sensor locations. The sensor network is composed of S1, S2, S3 and S4.

Various Contamination Events. To examine the performance of the LRM-LS method, it was first applied to five contamination events with various source characteristics. The description of these hypothetical events is provided in Table 5.2, and the injection locations of the contaminant are shown graphically in Figure 5.3.

Table 5.2 True Source Description for Five Contamination Event Scenarios

Characteristics	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**
Location	Node 205	Node 105	Node 125	Node 171	Node 267
Starting Time	12:00 p.m.	11:00 a.m.	10:00 a.m.	1:00 p.m.	2:00 p.m.
Mass Loading (g/min)	40	30	70	60	80
Duration (min)	120	60	90	10	30

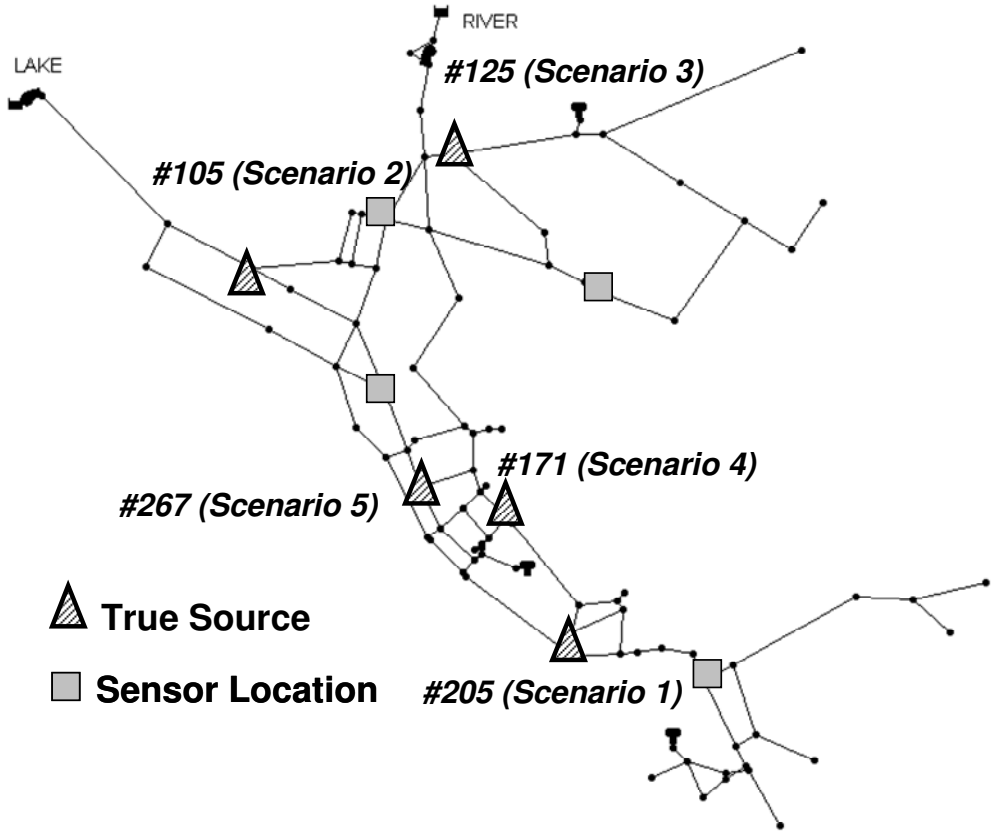


Figure 5.3 Illustration of injection locations for five hypothetical events. The triangles indicate the contaminant sources; the number adjacent to the node is the node ID, and the squares designate the sensor locations.

Whereas the LRMs can predict the probability given the observations at each time interval, the current set of potential nodes used to perform the LS is constructed from the intersection of the candidate sets from the first detection to the current time. Table 5.3 provides the number of possible injection nodes for the five events. Scenarios 2 and 3 are observed to have fewer candidates compared to the other scenarios. Following the LR analysis, the NMS method was employed at each likely node to seek the optimal injection profile. After several random trials using different parameter settings, the parameters associated with the NMS method were used, as follows. The simplex consists of four vertices in the case of three decision variables at each possible node; and the NMS method is repeated maximally for 30 iterations at 10-minute observation intervals. Due to the randomness resulting from the initialization, 30 random trials were performed in each scenario, and the results discussed below are summarized based on these random runs. The characterization results are summarized in Table 5.4.

Table 5.3 Number of Potential Nodes Obtained from the LRMs for Five Hypothetical Contamination Events

Time Elapsed (min)	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**
10	48	12	17	53	53
180	48	9	17	52	46

Table 5.4 Summary of the LRM-LS Results for Five Hypothetical Contamination Events

	Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**
Prediction error	0.017±0.048	0.01±0.028	0.025±0.104	0.001±0.000	0.085±0.052
Absolute error in starting time	0.00±0.00	0.00±0.00	0.00±0.00	0.18±0.03	0.51±0.50
Absolute error in duration	0.00±0.00	0.00±0.00	1.28±0.23	4.39±11.97	0.89±0.80
Relative error in mass injection rate	0.022±0.008	0.028±0.012	0.025±0.009	0.10±0.18	0.19±0.26
Rank	1.10±0.54	1.00±0.00	1.00±0.00	3.30±2.15	3.23±3.01
Number of alternatives	1.17±1.65	1.03±0.18	1.00±0.00	5.77±2.28	2.23±2.10

Table 5.4 provides comparisons of the obtained solutions to five scenarios using the average and standard deviation over 30 random runs. Reported results are related to the true source location after collecting 3-hour measurements. With regard to the accuracy of the source characteristics, the proposed approach results in good estimates for Scenarios 1, 2, and 3, whereas a relatively poor performance is observed in Scenarios 4 and 5. The discrepancies result mainly from the contaminant intrusion location in the network. It is interesting to note that Scenario 4 has the lowest prediction error, despite its higher inaccuracy in the prediction of source characteristics. This behavior reflects that the lowest prediction error does not always signify the best performance. In addition, the rank of the true source location was compared for the five cases, as listed in Table 5.4. Possible nodes were ranked according to the prediction errors. The node with the smallest prediction error was assigned to the first rank. The injection node was identified approximately in the first place when 3-hour measurements were taken since the first detections for Scenarios 1, 2 and 3, while Scenarios

4 and 5 were observed to have a worse performance in terms of the rank of the true injection node. The reason for this result might be that insufficient observation data did not allow the true source location to be distinguished from the other nodes in Scenarios 4 and 5. Additionally, the authors were interested mainly in the degree to which non-uniqueness varies with different contaminant characteristics. Non-uniqueness was quantified as the number of alternative solutions. A solution was determined as an alternative by a target value, which was calculated from the minimal prediction error multiplied by a relaxation value (1.5 in this study). Similar to the performance with respect to rank of the injection node, Scenarios 4 and 5 resulted in a large number of alternatives.

Sensitivity of Monitoring Data. Whereas the deployment of monitoring sensors enables water utilities operators to access real-time or near real-time data, the quality of the collected measurements is a major factor that impacts identification uncertainty. Note that it is common for sensor observations to incorporate a certain level of uncertainty. It is more likely that binary decision sensors are installed in the network, and these allow the status of the contamination to be known. To better understand the effect of monitoring data on identification uncertainty, two additional experiments were conducted based on Scenario 1. The first experiment simulates the scenario using uncertain measurements. A normally distributed white noise was added to the measured concentration data, and the mathematical formulation for modeling the factor with perturbation is expressed as follows:

$$y_{it}^{err} = y_{it} + \alpha * y_{it} * N(0,1), \tag{5.3}$$

where y_{it}^{err} denotes the perturbed measurement of sensor i at time t; y_{it} denotes the measurement of sensor i at time t; and α represents the error level added to the perturbed

factor (0.1 in this example). In the second test, the sensors were assumed to be able to indicate the contamination status. For simplicity, a fixed concentration level was specified as the sensor detection threshold (0.1 mg/L is assumed in this example). Then, the simulated concentration data were translated to 0/1 according to the specified threshold.

The results of Scenario 1 under the ideal, noisy, and binary sensor conditions were compared and are summarized in Figure 5.4 and Table 5.5. Figure 5.4 summarizes the mean and standard deviation of the number of alternatives against the elapsed time over 30 random trials. As can be seen, fewer alternatives were identified as a consequence of a decrease in measurement uncertainty. Also, it is noted that the true injection node does not always rank first with respect to the prediction errors. Therefore, one of the most valuable aspects of this proposed approach is that simultaneous searches at multiple locations yield reliable solutions and can resolve the problem of non-uniqueness of solutions.

Over time, however, the number of alternatives did not decrease consistently. This lack of consistency may be explained by the fact that the contaminant source characterization is a dynamic process whereby the observation data emerge dynamically. Each time a new observation streams in, the length of the search convergence depends on the severity of the changing environment.

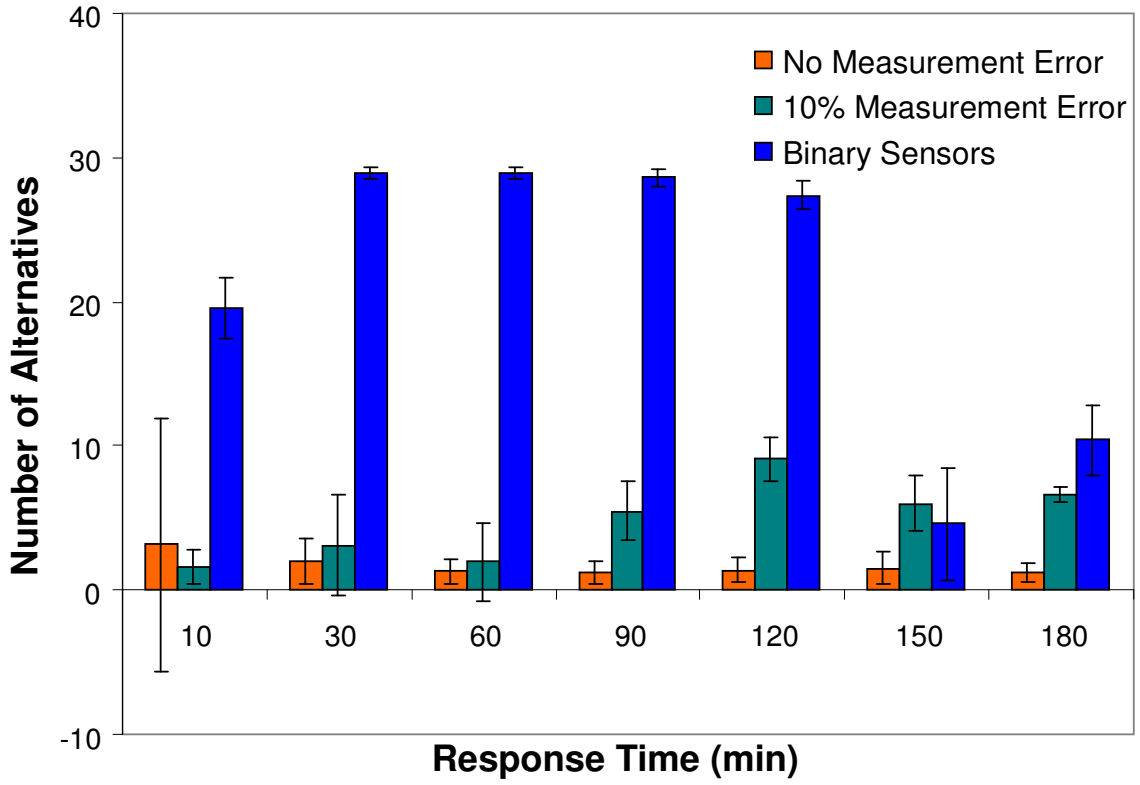


Figure 5.4 Comparisons of results for Scenario 1 under different sensor conditions

The results reported in Table 5.5 correspond to the injection node and number of alternatives when three hours have elapsed since the first detection. The contaminant source is characterized more accurately in the first two cases due to a lower level of uncertainty than in the other cases. Although the addition of a 10% measurement error results in slightly smaller errors in the prediction compared to the ideal condition, a larger number of alternative solutions is evident. In addition, it must be noted that the differences in the possible source locations obtained from the LRMs among these three conditions are very small.

Table 5.5 Summary of Results for Scenario 1 under Different Sensor Conditions

	No Measurement Error	**10% Measurement Error**	**Binary Sensors (0.1 mg/L detection limit)**
Absolute error of starting time	0.00±0.00	0.00±0.00	0.59±0.17
Absolute error of duration	0.00±0.00	0.00±0.00	3.71±1.17
Relative error of mass injection rate	0.022±0.008	0.006±0.015	0.31±0.40
Number of alternatives	1.17±0.65	6.57±0.57	10.40±2.39

Monte Carlo Simulation. In this section, the performance of the proposed approach is further evaluated using 50 hypothetical contamination events. These events were simulated by randomly varying the contaminant characteristics, including injection location, starting time, duration and strength. The varying source parameters and their specified ranges are listed in Table 5.1. For each arbitrarily simulated contamination event, the LRM-LS method was executed for one trial, and the summary of the performance measures across all the events at different elapsed times since the first detection are presented in Figure 5.5. It is observed that the average rank of the true injection location approaches one as time goes on. Similarly, the correlation coefficient of coordinates between the true source and the optimal solution along both the x and y axes becomes closer to one with a longer observation period. Overall, the results indicate the robustness of the LRM-LS method for a number of random contamination events.

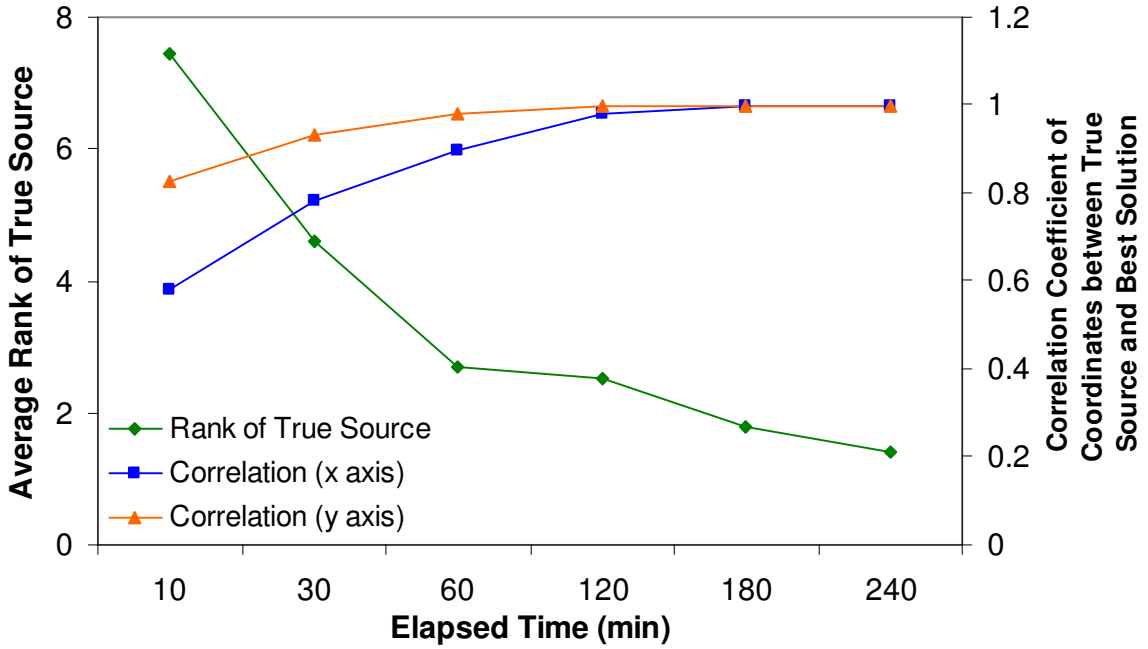


Figure 5.5 Summary of results for 50 random contamination events

5.5.2 Micropolis Example Network

To evaluate the impact of a large network on the LRM-LS solutions, a second network was studied, as shown in Figure 5.6. The entire system contains 1574 junctions, 1415 pipes, 8 pumps, 2 reservoirs, and 1 tank. This network was developed for a micropolis virtual city with 5,000 residents, further details of which can be found in Brumbelow et al. (2007). The locations of five ideal sensors were chosen at random within the entire network (see Figure 5.6).

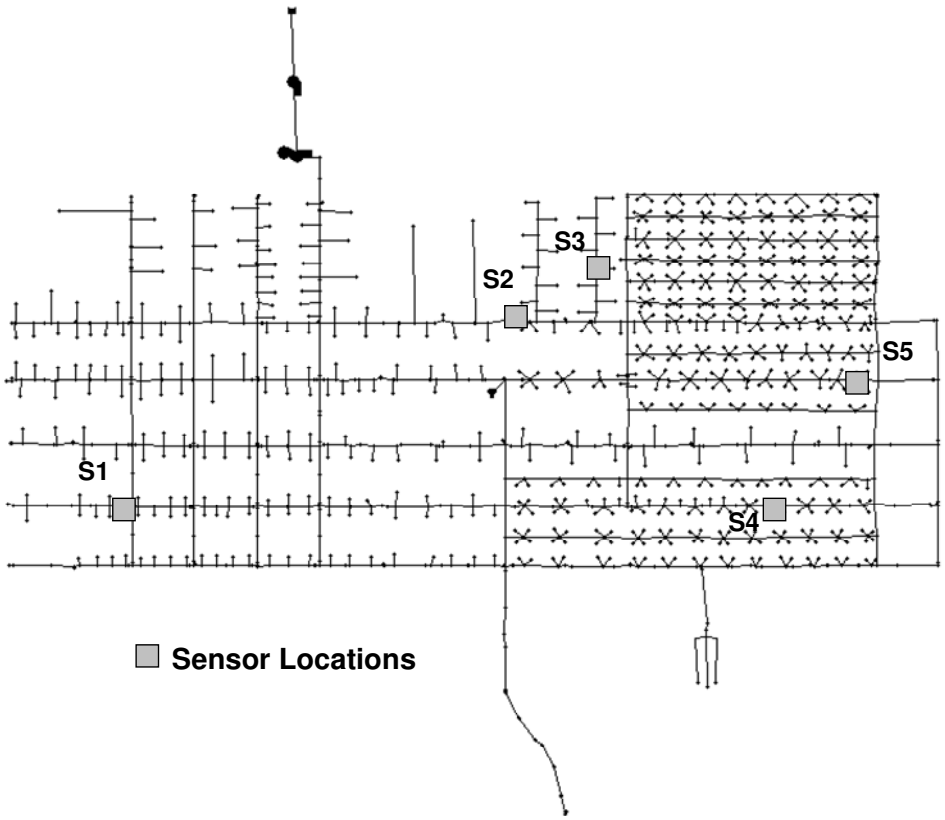


Figure 5.6 Schematic of micropolis water network. The sensor network is composed of S1, S2, S3, S4, and S5, designated by squares

For testing purposes, a contamination event was simulated, with the contaminant injection at node IN1646 (denoted as a triangle in Figure 5.7). At 10:00 a.m. the contaminant entered the network at the rate of 60 g/min, and the event lasted an hour. Given the five monitoring sensors, the detection occurred initially at sensor S5 at 12:30 p.m. and lasted until 1:40 p.m. The LRMs identified 167 solutions out of 409 nodes that could contribute to observations at the given sensor locations. Figure 5.7 illustrates potential source nodes with respect to the first detection. It is worth noting that these possible sources are relatively close to the true source, one of which is the true injection node. For the given sensor network,

however, a large set of unknown nodes existed due to the fact that these nodes were undetectable if the contaminant happened to be injected at those nodes. Subsequent to the LR analysis, multiple NMS-based LS operations were conducted to seek the optimal release history at each corresponding node.

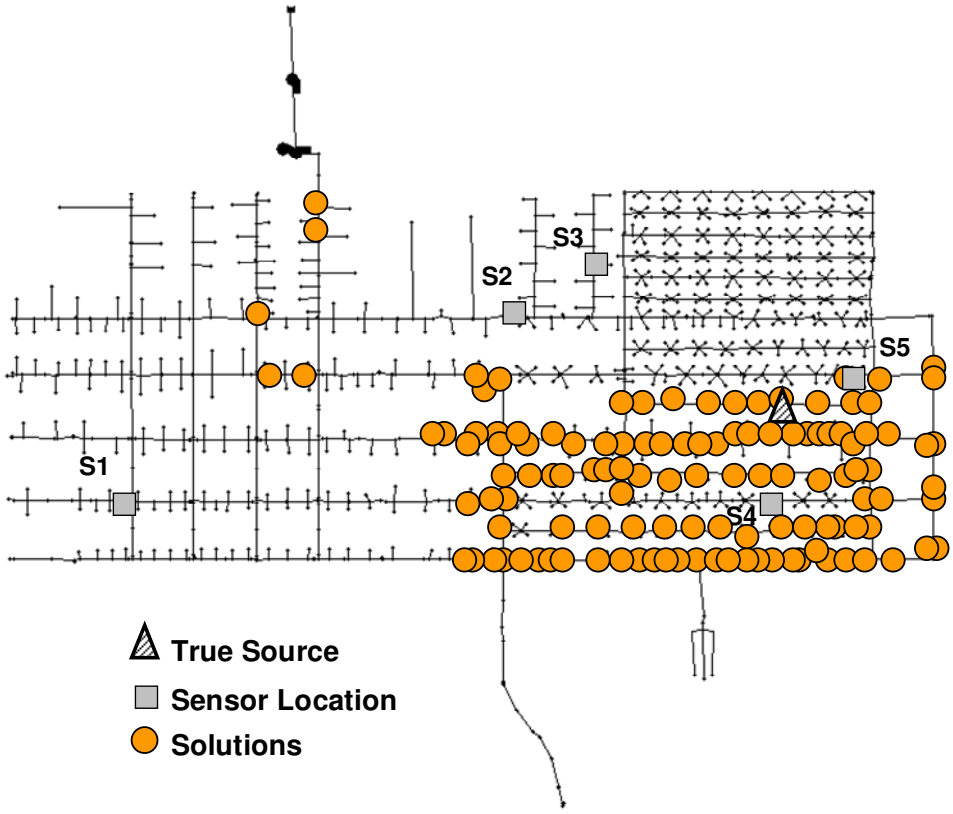
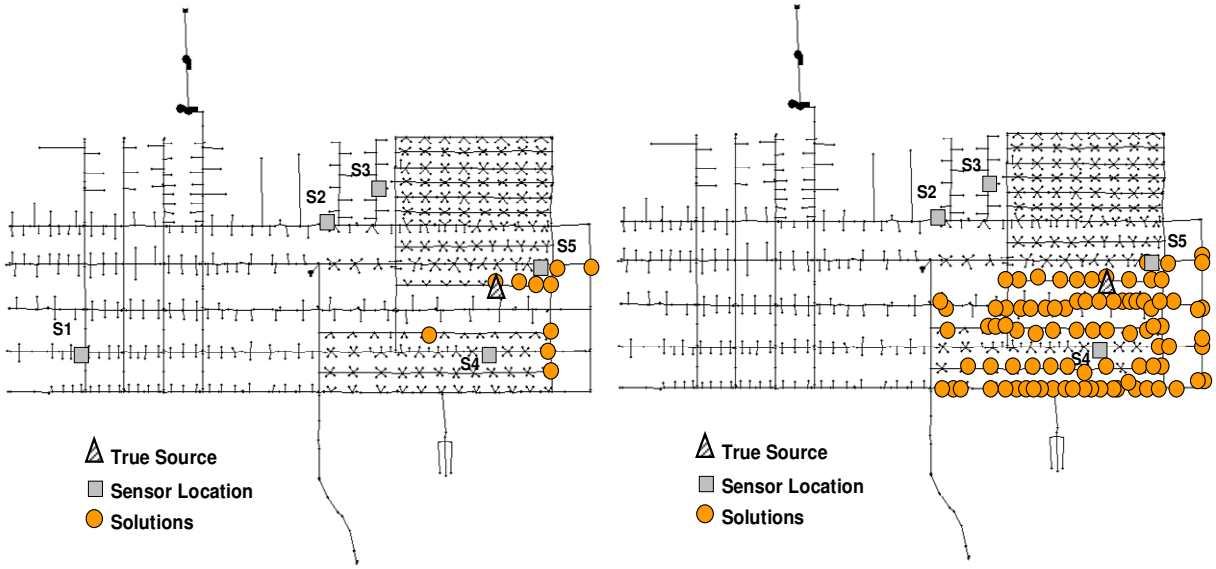


Figure 5.7 Location of possible source locations at 12:30 p.m.

In this study, one of the main interests was to discover the number of alternatives and their distribution with variations given the measurement quality. Similar to the previous case study, and in addition to the ideal sensor condition, a 10% level of uncertainty was added to the measured data, and the binary sensors condition (0.1 mg/L as the detection limit) was assumed. Again, the target for the alternatives was the minimal prediction error multiplied by

a relaxation value (1.5). Reported results correspond to the last measurement time step. Under the ideal sensor condition, a unique solution at the true injection node was determined. However, in the other cases, the resulting alternatives increased to 10 (for the 10% measurement error) and 111 (binary sensors), respectively. This finding implies that the uncertainties associated with measurements contribute to the complexity of source identification. Figure 5.8 presents the location of alternative solutions at 1:40 p.m. It is interesting to note that the obtained alternatives cluster around the true injection node, even in the case of the binary sensor condition. This information may be valuable for decision makers in locating the region of interest. On the other hand, the alternatives migrate towards the true injection node owing to the improved measurement quality.



(a) (b)

Figure 5.8 Location of alternative solutions at 1:40 p.m.: (a) 10% measurement error; (b) binary sensor condition.

5.6 Summary

Discovering a contaminant as it occurs requires an algorithm that can quickly locate the contaminant's injection node as well as its release history. Specifically, this study concentrates on the LRM-LS procedure to seek the contaminant characteristics. Whereas the LRMs facilitate the removal from the search of unlikely nodes as the sources, the subsequent LS procedure enables the efficient characterization of the contaminant in the localized region. Applications of the LRM-LS method to two example water networks include not only various contamination scenarios but also impose variations on the quality of the monitoring sensors.

The results suggest that the LRM-LS method is capable of adaptively discovering a set of possible contaminant locations along with corresponding release histories. Simultaneous searches at multiple locations assist in evaluating the degree of non-uniqueness and accordingly reduce the possibility of misidentification. The identification accuracy is impacted by the location of the contaminant, monitoring data, and the network size. It is worth mentioning that the true source node does not always rank first among candidate nodes. This phenomenon becomes more significant when the uncertainty is increased due to insufficient observations, coarse monitoring data, unknown water consumption, or a large network. However, this study shows that the identified solutions typically concentrate in the vicinity of the true source location. Future research could extend this approach to apply to reactive contaminants.

Nevertheless, the LRMs commonly yield a large set of possible nodes. To avoid unnecessary computational effort, future work could consider establishing the candidate set

by both the LRMs and the LS. Due to the difficulty in distinguishing adjacent nodes, given the sparse observed data, likely injection locations typically are close to the true injection node (Hill et al., 2006). Thus, a node close to the current optimum is worth further investigation, even with a high prediction error. Conversely, a node with a high prediction error that is far away from the current optimum may necessarily be eliminated from the candidate set if the contaminant is assumed to be introduced from a single location, which would further reduce the computational burden associated with the LS.

CHAPTER 6. A Hybrid Heuristic Search Approach for Contaminant Source Characterization

Abstract. The rapid discovery of the contaminant source and its mass loading characteristics in a water distribution system (WDS) is vital for generating an efficient control strategy during a contamination event. Previous work on the Adaptive Dynamic OPtimization Technique (ADOPT), which was developed as an Evolution Strategy (ES) based procedure, presents an approach to estimate the source characteristics adaptively, given dynamically updated observation data. Although this simulation-optimization approach is promising, it is computationally expensive, which poses challenges in the context of real-time solutions. This chapter reports the findings of an investigation that builds upon the prior work by introducing a hybrid heuristic search method for the real-time characterization of a contaminant source. This new method integrates the ES-based ADOPT with a logistic regression (LR) analysis and a local improvement method to expedite the convergence and possibly solve the problem quickly. As a prescreening technique, a LR analysis step is performed prior to ADOPT; this step reduces the search space by eliminating unnecessary source nodes as potential source locations. Then, a local search (LS) approach is embedded into some of the algorithmic steps in ADOPT to serve as a postscreening step that potentially speeds up the convergence in localized regions in the solution space. Numerical experiments for the proposed hybrid approach are performed on an example water distribution network, and the results are compared with those of the standard implementation of ADOPT.

6.1. Introduction

Water distribution systems (WDSs) are highly susceptible to various threat attempts, including physical attack, cyber-disruption, and biochemical contamination (Clark and Deininger, 2000). Intentional biochemical contamination has become a major concern recently due to its inherent complexities and the potential hazard to human health. To enhance the response capability in the case of a contamination event, one protective step that should be taken is the installation of monitoring stations within the WDS to measure abnormal water quality resulting from the pollutants. While real-time measurements are collected at these stations, the contaminant source must be determineded swiftly to facilitate appropriate decisions in response to the threat. Contaminant sources are typically characterized by the injection location, starting time, duration, and mass injection rates that correspond to different time intervals.

Contaminant source characterization can be deemed as an inverse problem, the solutions of which are typically ill-posed (e.g., *non-existence*, *instability*, or *non-uniqueness*). *Non-existence* occurs when no solution exists, given available observations. *Instability* is caused by the sensitivity of inverse solutions to the observations. *Non-uniqueness* can be caused by insufficient data when different solutions yield similar explanations for the observations. These ill-posed solutions create challenges, especially in the context of system uncertainties and the need for rapid identification.

One way to address these issues is to formulate the problem as an optimization problem by making use of currently available observation data to inversely seek the best source characteristics. This approach may yield a good explanation for the obtained water

quality samples. Most efforts that have been made to solve such a problem use optimization methods to discover the true source information by minimizing the deviation between the observed output from the sensors and the calculated concentration values. Such efforts include direct methods (e.g., Van Bloemen Wannders et al., 2003; Laird et al., 2005) and heuristic-based simulation-optimization approaches (Guan, 2006) that constitute the focus of this research due to their flexibility and robustness. In Liu et al. (2006), the ES-based Adaptive Dynamic OPtimization Technique (ADOPT) is employed to search for a set of contaminant source characteristics that may result in similar sensor observations. Although ADOPT has shown promise for its dynamically adaptive capability, the simulation-optimization approach suffers because it requires numerous computationally expensive simulations to adaptively recover the contaminant characteristics. The significant computational costs would unavoidably impact not only the identification time but the resultant solution quality. Thus, to minimize the number of simulations and maintain acceptable accuracy, ADOPT has been enhanced by incorporating techniques that can expedite the solution.

In this chapter, an algorithmic framework is introduced for solving a WDS contaminant source characterization problem by integrating ES-based ADOPT with one prescreening and one postscreening technique. The objective of a prescreening technique is to narrow the search space. While the search is gradually converging to local regions, a postscreening technique is embedded into ADOPT to promote its local convergence capability.

6.2. Problem Statement

Given increasingly available observations, a contaminant source characterization can be constructed as a dynamic optimization problem. The objective function is calculated by minimizing the difference between the estimated concentration values and the currently available data, which are time-dependent. The primary goal is to quickly estimate the source characteristics that best explain the observed data at hand. The following mathematical formulation is defined to determine the source location, starting time, and corresponding mass loading history at any time after the first detection of contamination. Although this definition assumes that the contamination is introduced at only one node in the network, the same model can be entended to consider multiple contamination source locations.

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

$$\text{Minimize } F = \sqrt{\frac{\sum_{t=t_0}^{t_c}\sum_{i=1}^{N_s}(C_{it}^{obs} - C_{it}(L, M_{t_c}, T_0))^2}{N_s * t_c}}, \tag{6.1}$$

where

F = prediction error;

L = contaminant source location;

T_0 = starting time of the contamination event;

t_0 = time of the first detection of the contamination at the sensors;

t_c = current time step;

$M_{t_c} = \{m_{T_0}, m_{T_0+1}, \cdots, m_{t_c}\}$, represented as a vector of mass injected at the source from time T_0 to t_c;

C_{it}^{obs} = the observed concentration at sensor *i* at time step *t*;

$C_{it}(L, M_{t_c}, T_0)$ = model (i.e., EPANET)-estimated concentration value at sensor *i* at time step *t*;

i = sensor location;

t = time step of observation; and

N_s = total number of sensors.

6.3. Solution Approach

6.3.1 ES-based ADOPT

Two major challenges that must be faced when addressing a contaminant source characterization problem are: 1) continuously tracking the optimal solutions in dynamic environments, and 2) capturing the potential non-unique solutions that could provide similar explanations to the observations over time. To tackle these challenges, ES-based ADOPT was developed to search for potential solutions that are as different as possible from each other by exploring the decision space through multiple subpopulations. To set a benchmark for the best possible solution, one of the subpopulations searches independently for the solution that best fits the observations. The remaining subpopulations use that benchmark to find other possible solutions that fit the observations equally or nearly as well as the best solution. This procedure is executed for each observation time step. When new information becomes available, the solutions from the previous time step can be adapted to the new environment if it is assumed that the observation data resulting from one contamination event.

Therefore, the number of remaining subpopulations can be adapted according to the dynamic environments. The algorithm terminates when no more available information or only one subpopulation remains. The key steps of this approach are outlined as follows (Liu et al., 2006):

Step 1. Create an initial set of random solutions, equally divided among N subpopulations.

Step 2. Increment time step as $t \leftarrow t + 1$. Set generation index as $g = 0$. Update monitoring data with additional measurements and construct the prediction error function.

Step 2.1. Increment generation index as $g \leftarrow g + 1$. In the first subpopulation ($p = 1$), evaluate the fitness based on the prediction error. In other subpopulations ($p= 2, 3,\ldots N$), evaluate the fitness based on the prediction error and its distance from all other subpopulations.

Step 2.2. In each subpopulation, apply selection and mutation operators, and create a new set of solutions.

Step 2.3. If stopping criterion (i.e., $g <$ max no. of generations) is not met, then go to Step 2.1; otherwise, go to Step 3.

Step 3. Eliminate subpopulations that represent duplicate solutions. If only one subpopulation remains or the current set of solutions is acceptable, then stop.

Step 4. If no more observations are available, then stop; otherwise, return to Step 2.

The efficiency of ES-based ADOPT in characterizing a contaminant source can be improved by eliminating unnecessary nodes as potential source locations. In Chapter 4, a predictive model based on logistic regression (LR) analysis was examined to determine the likelihood that any given node is a source. The resulting location-specific probability values could then be used to reduce the solution space for the optimization search, thus yielding a faster convergence. In addition, as a global optimizer, the ES typically becomes less efficient when the search concentrates on a local region. To further enhance the algorithm's efficiency, a local search (LS) operation is incorporated into ADOPT, which allows the advantage of local fine-tuning.

6.3.2 Logistic Regression Model (LRM)

The numerous potential contamination scenarios and system uncertainties collectively contribute to the complexity of source characterization in a contamination event. To offer a fast probabilistic estimation of potential source locations, a linear LRM-based approach was presented in the previous study to model the likelihood of a node as a source, given the sensor observations. The LRM, constructed as follows, describes the relationship between the probability that node *i* is a source and the observations at time *t* after the contamination is detected at one or more sensors.

$$\log\left(\frac{p(A_i \mid C_{1t},\cdots,C_{Nt})}{1-p(A_i \mid C_{1t},\cdots,C_{Nt})}\right) = b_{0t} + b_{1t}C_{1t} + \ldots + b_{jt}C_{jt} + \ldots b_{Nt}C_{Nt} \tag{6.2}$$

where $p(A_i \mid C_{1t},\cdots,C_{Nt})$ 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*; (b_{0t}, …, b_{Nt})

are the regression coefficients obtained by the maximum likelihood procedure; and (C_{1t}, …, C_{Nt}) denotes the sensor observations at time t.

While knowledge of an existing WDS and its sensor placement allows simulations of contamination events, the LRMs that correspond to each node at each time interval can be pre-identified through a large number of hypothetical contamination events. Once a contamination is detected, the models can rapidly return the probability estimation. Thus, the strength of a LRM is that it offers a simple and direct way to make a fast prediction of a small set of nodes that are likely candidates to be a source. In this study, the resulting solutions from the LRMs are used to reduce the search space in ADOPT by eliminating unnecessary nodes, which typically have zero probability as calculated by the LRMs.

6.3.3 Heuristic Search Methods

Heuristic search algorithms have received increasing attention due to their potential in tackling complex optimization problems. In addition to the effectiveness of such algorithms, the objective function evaluation is the only information required to direct the search during the iterative process. This advantage allows wider applications to a variety of complicated problems by coupling search algorithms to simulation models. Evolutionary algorithms (EAs) (Holland, 1975), as one class of heuristic methods, present a global search (GS) mechanism and are of great benefit to large nonlinear optimization problems. Making use of a population-based scheme and particular operators (e.g., crossover, mutation), EAs can discover global optima independently of initial guesswork as well as offer easy parallel implementation (Xu et al., 2001). However, as the population gradually moves towards

optimal solutions, the efficiency of EAs typically decreases due to the stochastic property of the search process. In the early stages, and because the large search space inherent to early stages contains potentially better solutions, the EAs are more likely to discover improved solutions. As the search approaches the vicinity of local optima, the chance of finding better solutions decreases, and the search yields a slow local convergence (Xu et al., 2001; Gen & Chen, 1997). In contrast, LS approaches, such as the Hooke-Jeeves pattern search and the Nelder-Mead Simplex (NMS) 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.

Recently, some researchers have investigated hybrid heuristic approaches (i.e., the hybridization of a LS and a GS procedure) that are more likely to be efficient because they take advantage of both capabilities. The methods reported previously can be classified into three categories. The first one employs a GS procedure followed by a LS process (Partheepan, 2003; Espinoza et al., 2005; Mahinthakumar & Sayeed, 2005; and Yeh et al., 2007), where the GS discovers globally promising regions, the subsequent LS procedure continually moves the resulting good solutions towards the optimum. An enhanced solution thus can be acquired through increased computational efforts associated with the LS procedure. The LS does not benefit the GS process in this case, however. The second category alternately applies GS and LS procedures (e.g., Espinoza et al., 2005; Hart, 1994; and Land, 1998) whereby the two search procedures can support each other. Although this category has the potential to achieve a higher solution quality and computational efficiency, the parallel implementation of such an approach is inefficient. The third category of methods

embeds a local optimizer into a GS procedure as a genetic operator. One such approach, introduced by Xu et al. (2001), couples a micro genetic algorithm (GA) with the LS to generate two new solutions in place of the two worst solutions using the pattern move operator. Another approach, proposed by Partheepan (2003), links the GA with the NMS method by performing the reflection step as a new operator during the GA evolution process. Overall, the embedment of the LS into the GS procedure does not yield additional fitness evaluations. This factor is valuable, particularly in the context of complex real-world optimization problems that require time-consuming simulation runs.

Most discussions associated with the hybrid heuristic search methods focus on ways to maximally improve the solution quality with low computational costs. This problem introduces several questions, such as which LS procedure to use, the frequency of the LS, the time that the LS is performed, the way of selecting individuals to conduct the LS, etc. Different LS procedures may bring about quite different solutions. If too many LS procedures are incorporated into a GS, local convergence may result. In addition, it is worth noting that selecting inappropriate solutions or conducting a LS at an improper time can bias the direction of the GS (Hart, 1994; Land, 1998). This chapter investigates a hybrid model by embedding a local optimizer into the ES-based ADOPT for characterizing a WDS contaminant source.

6.3.4 Algorithm Framework

Figure 6.1 shows the framework of the hybrid LRM-ADOPT-LS, starting with the LR analysis. At time step t, when the initial detection occurs, the LRMs input the concentration

data collected by the sensors and output the probability of each node as a possible source. This process screens the candidate set, which is composed of nodes with non-zero probability values, for potential source locations. Specifically, the candidate set is employed in the subsequent initialization and mutation operations.

The combination of the ES-based ADOPT with a LS procedure is then applied to determine the source characteristics (i.e., contamination location, start time, duration and mass loading profile). In this phase, the ES-based ADOPT focuses on exploring promising regions, while the LS exploits local characteristics to expedite the convergence in localized regions. In addition to achieving a faster convergence without the expense of additional fitness evaluations, this hybrid search is expected to avoid premature convergence by embedding the pattern move simultaneously within multiple subpopulations. The incorporation of a LS operator into ES-based ADOPT is described in two steps as follows:

Step 1. Assess the need for the LS operator in each subpopulation. Subsequent to the selection operation, assess the necessity of the local optimizer. This process relies on the extent of the convergence of individuals. LS operations that are too numerous or performed too early may lead to premature convergence. Therefore, the LS operator is applied in a subpopulation when the following two conditions are met: 1) The proportion of individuals located at the same node exceeds a specified threshold (denoted as α); 2) The best individual locates at a different node from those in other subpopulations or is better than those in other subpopulations if they are at the same location.

Step 2. Perform the heuristic pattern move. In a subpopulation, the heuristic pattern move is carried out when the current convergence of individuals suggests its necessity. Based on the approach introduced by Xu et al. (2001), the pattern move operates around the current optimum to search for improved solutions.

Step 2.1. Sort individuals according to the prediction error function.

Step 2.2 Record the indices of the best, second best, worst, second worst, and third worst individuals for the same location.

Step 2.3. Generate three new individuals through the linear combination of the best and second best individuals. If I_b represents the optimum, and I_{sb} corresponds to the second best individual with a slightly worse objective value, the new individual $I_{new} = I_b + \beta(I_b - I_{sb})$ is constructed. β is the pattern move step size. (0.2, 0.5, and 1.0 are used in this study.)

Step 2.4. Replace the three worst individuals with the newly generated individuals.

Finally, this hybridization process is repeated over time. At each time step, this procedure leads to a set of possible solutions that is able to match the sensor observations within a specified range of prediction error. With increasingly available measurements, the set of alternatives reduces in number, and eventually reaches a single solution when sufficient sensor information is available. The identified single solution or a set of alternatives is expected to contain the true source. The algorithm terminates when one subpopulation or no additional measurements remain.

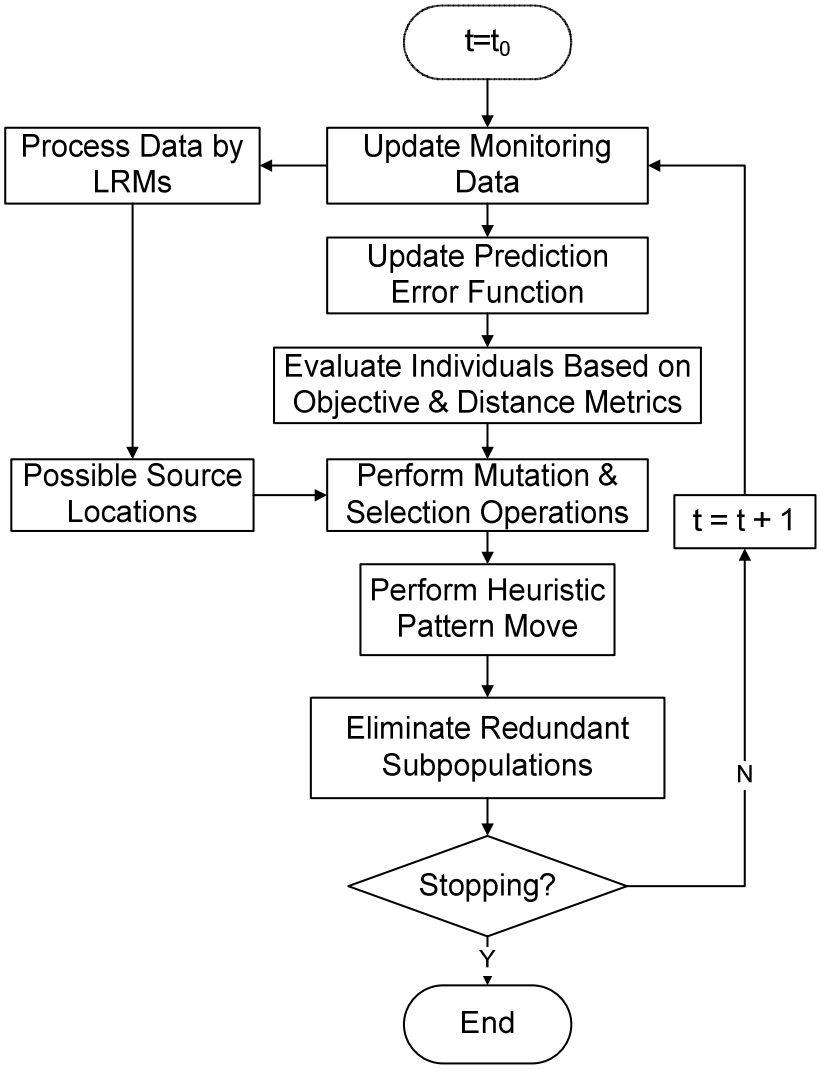


Figure 6.1 LRM-ADOPT-LS optimization framework

6.4. Applications

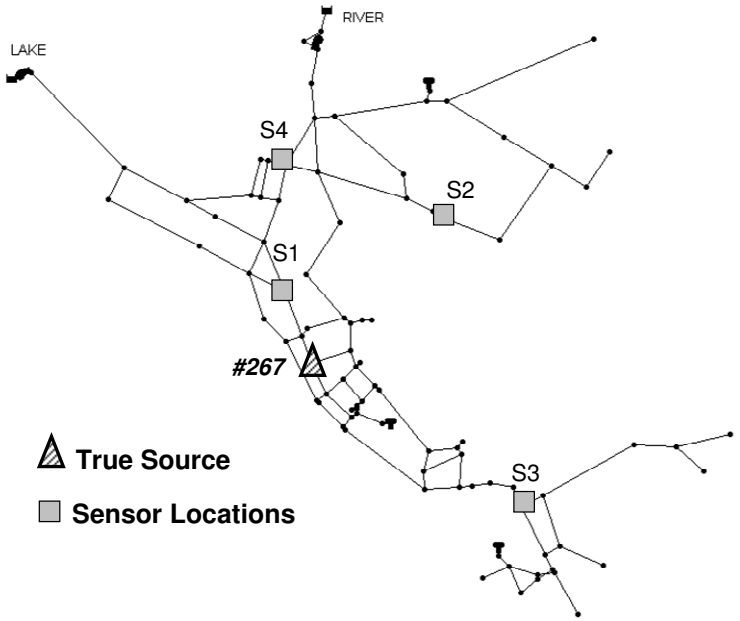
To test the improvements provided by the LRM and LS, several contamination scenarios in two WDS networks were studied. In this chapter, results for only two illustrative scenarios are presented. Scenario 1 assumes the intentional intrusion of the contaminant into a small water network. This network is one of the problem scenarios in the EPANET software (Rossman, 2000) and is comprised of 2 sources, 3 tanks, 97 nodes, and 117 pipes.

The configuration of this network is depicted in Figure 6.2 (a). Four ideal sensors are randomly placed within the network. A conservative contaminant enters the system through node #267 (denoted as a triangle in Figure 6.2 (a)) at 10:00 a.m., and the event lasts for one hour at the time-varying injection rates of 50, 60, 90, 80, 70, 40 g/min, each of which corresponds to a 10-minute interval. The first detection occurred at 12:20 p.m. by sensor S3. No non-zero concentration data were obtained at the other sensors during the observation period. This scenario is used as the baseline in the following discussions.

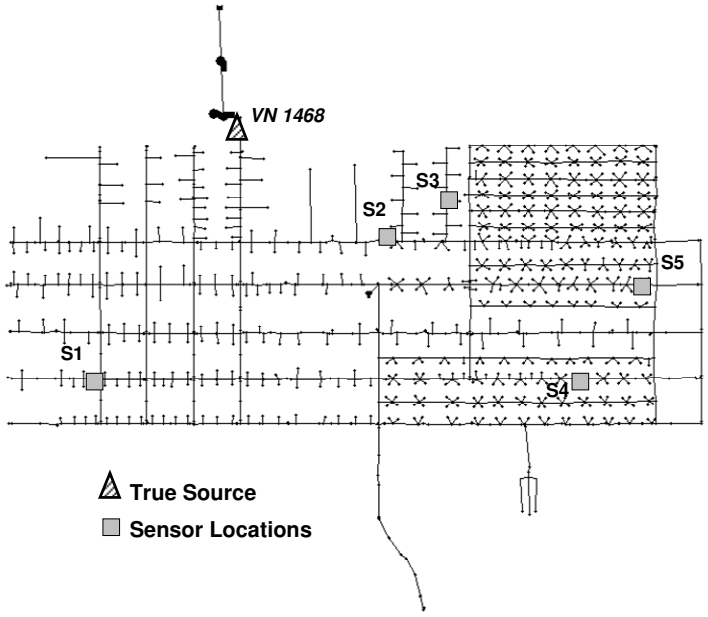
Scenario 2 represents a contamination event in a large network. This network contains 1574 junctions, 1415 pipes, 8 pumps, 2 reservoirs, and 1 tank. This network was developed for the micropolis virtual city with 5,000 residents, further details of which can be found in Brumbelow et al. (2007). The locations of five ideal sensors are chosen at random within the entire network (see Figure 6.2 (b)). The contaminant entered the network through node VN1468 at the rates of 80, 70, 60 g/min for a duration of 0.5 hour. The contamination was initially detected at sensor S2 after 6.5 hours had elapsed since the first release of contamination. In addition to sensor S2, non-zero concentration data were observed at sensors S2 and S3 during the monitoring period.

EPANET was used to run the hydraulic and water quality simulations. The contaminant transport was simulated in 10-minute intervals, and the concentration values at the sensors were observed in 10-minute increments. Given the dynamic sensor observations, the proposed approach enables an adaptive description of the contaminant, including the injection location, starting time, duration and magnitude. The allowable ranges of the decision variables are listed in Table 6.1. The parameter settings of the ES-based ADOPT are

shown in Table 6.2. To make the results statistically meaningful, each experiment was executed for 30 random trials.



(a)



(b)

Figure 6.2 Layout of the example networks. The triangle represents the injection node, and the square designates the sensor location: (a) Scenario 1; (b) Scenario 2

Table 6.1. Allowable Ranges of Source Parameters

Source Parameter	Scenario 1	Scenario 2
Location	Any node (1~97)	Any node (1~1577)
Starting Time (hr)	0~15	0~20
Duration (hr)	0~6	0~6
Mass Injection Rate (g/min)	0~100	0~100

Table 6.2 Parameter Settings of the (μ+λ) ES-based ADOPT

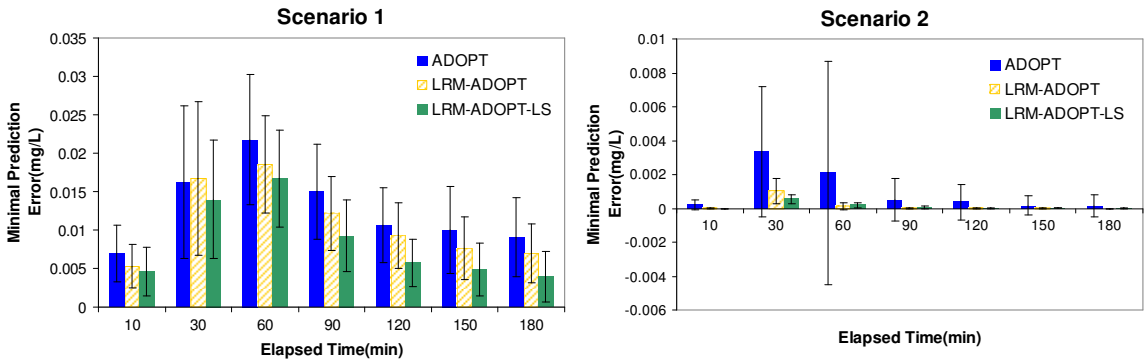
Parameter	Scenario 1	Scenario 2
Parent (μ) + mutants (λ)	100+100	100+100
Number of subpopulations	20	40
Generations	20	20

6.4.1 Performance Comparison among ADOPT, LRM-ADOPT, and LRM-ADOPT-LS

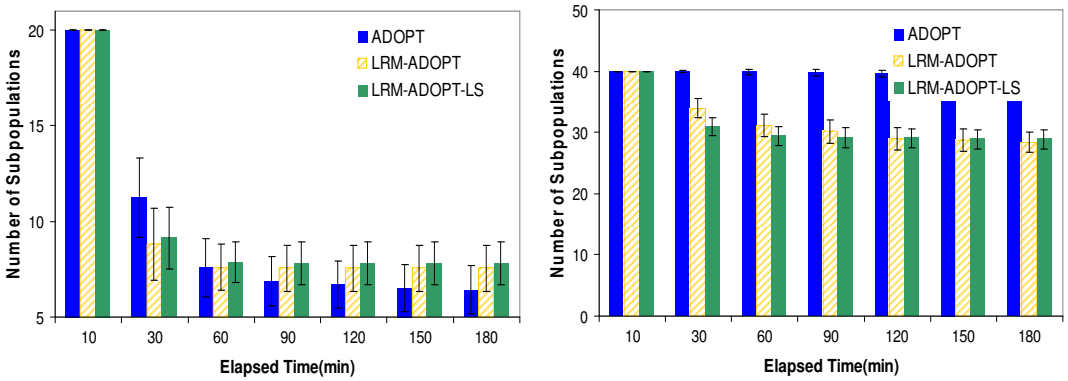
In this section, a performance analysis helps to better understand the enhancements associated with the LRMs and the LS. For each scenario, three sets of experiments were carried out with different enhancements added to ADOPT. The first set of runs focused on the pure ADOPT, which served as the baseline for the comparison. Secondly, the LR analysis was performed prior to the ADOPT runs to investigate the effects of the LRMs on the solutions. The last experiment investigated the combined improvements made by the LRMs and LS, whereby the LR analysis was carried out to reduce the search space, and the heuristic pattern move was embedded into the ADOPT operations. The pattern move was evoked when the convergence of the current population satisfied the specified criteria.

Scenario 1 includes 54 nodes that possibly contributed to the first detection and eventually decreased to 14 potential nodes. However, Scenario 2 initially contained 45 potential nodes that reduced to 24 in the end. Figure 6.3 illustrates the performance of the three strategies on two contamination scenarios by plotting the minimal prediction error,

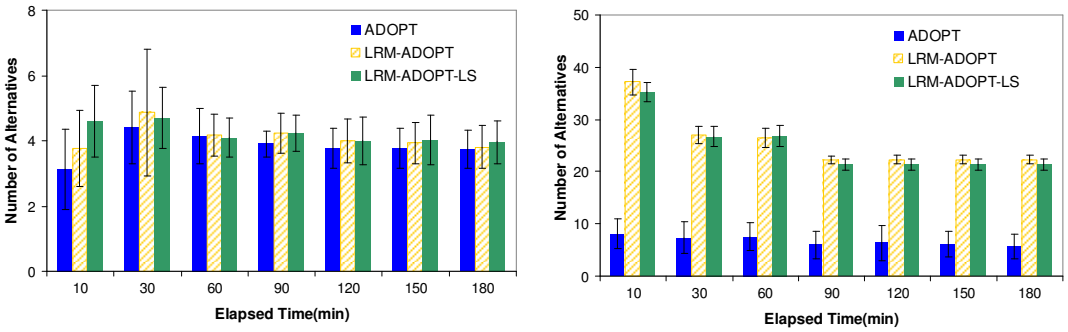
number of subpopulations, and alternatives. Overall, the solution quality improves in terms of minimal prediction errors due to the addition of enhancements, as shown in Figure 6.3 (a). To gauge computational costs, Figure 6.3 (b) presents the number of remaining subpopulations as a function of the elapsed time since the first detection. ADOPT, in combination with the LRM and LS, produces faster convergence in the first hour of detection. The outcome of Scenario 2 is the opposite of that of Scenario 1 in the later stages of characterization with respect to the discrepancy among the three approaches. As can be seen in Figure 6.3 (b), ADOPT in Scenario 1 has fewer subpopulations than the other approaches, and yet the most subpopulations are retained in Scenario 2. This observation can be explained by the fact that fewer alternatives identified by ADOPT led to a faster convergence in the later stages in Scenario 1. In contrast, the larger decision space in Scenario 2 increased the difficulty of convergence. The number of alternatives identified at each time interval is illustrated in Figure 6.3 (c). The enhancements added to ADOPT allowed it to make quick progress early on and capture more alternatives. Applying the LR analysis and LS in Scenario 1 helped to achieve the best performance with roughly equivalent computational costs. Although the LRM-ADOPT-LS in Scenario 2 performed slightly worse than the LRM-ADOPT in terms of the number of generated alternatives, ADOPT coupled with LRM and LS achieved a greater reduction in computational costs due to its faster convergence in the early stages.



(a)



(b)



(c)

Figure 6.3 Comparisons of the results from different approaches: (a) minimal prediction error; (b) number of subpopulations; (c) number of alternatives

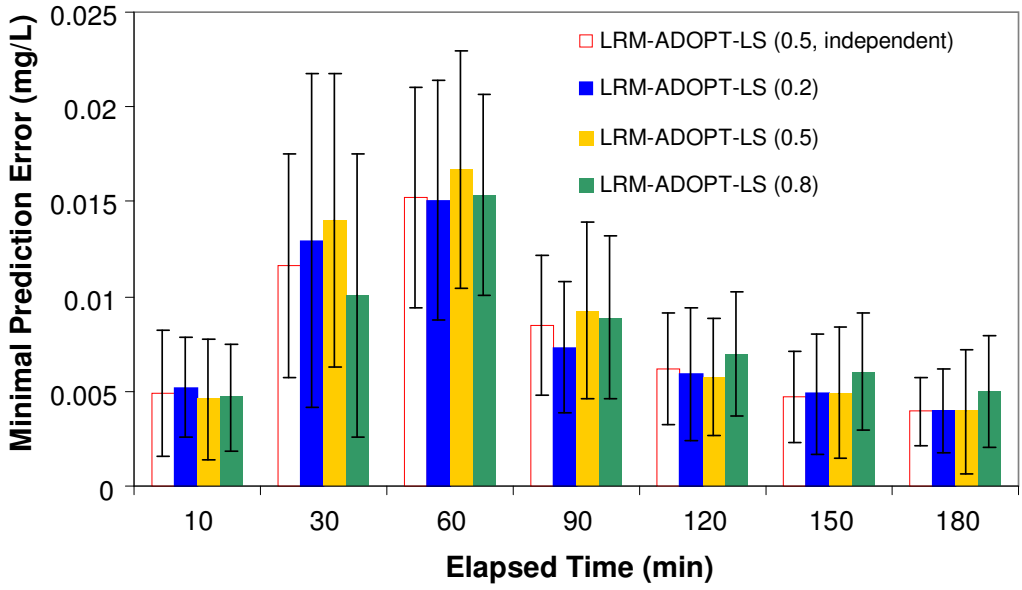
6.4.2 Local Search Selection

This section focuses on understanding the effects of LS selection on the performance of ADOPT in combination with the LRM and LS. First, an investigation was undertaken to

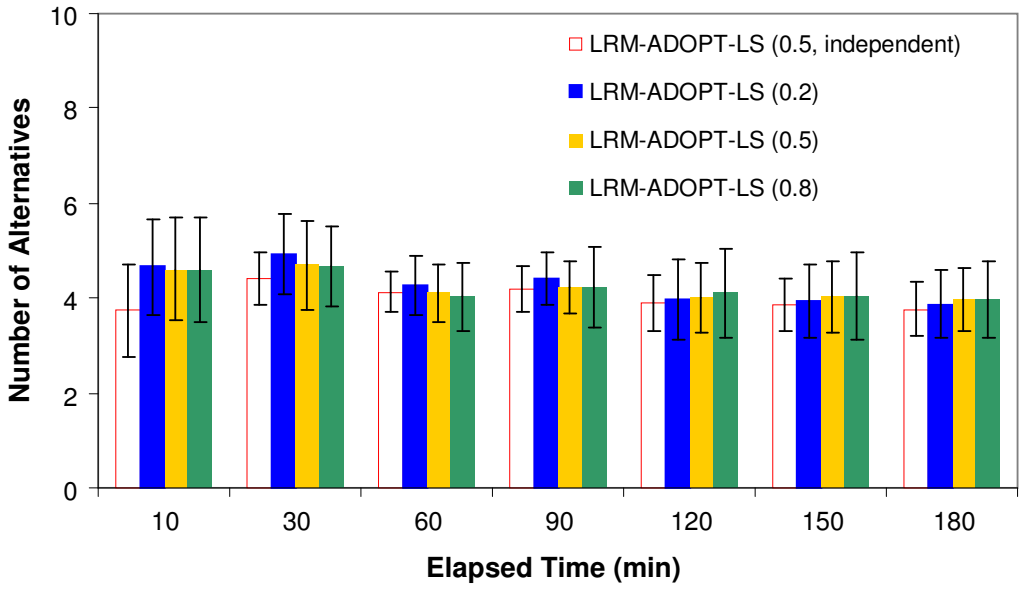
discover if it is necessary to consider other subpopulations when evaluating the need for a LS for a subpopulation. Secondly, an investigation was carried out to determine the amount of convergence that is needed to perform the LS for a subpopulation. In the first experiment, the necessity of the LS was evaluated for a given subpopulation, irrespective of the other subpopulations. The pattern move was conducted when more than 50% of the individuals converged to the same node. In the last three experiments, the LS operation required a subpopulation with the optimum located at a different node from the others or a lower prediction error if multiple subpopulations identified the optima at the same location. These three runs differed in the setting of parameter α, which, as described earlier, is the proportion of the individuals that locate at the same node in a subpopulation. Thus, this value can be used as an indicator of the level of convergence. A large value represents the infrequency of the LS. In the last three experiments, α was set as 0.2, 0.5, and 0.8, respectively.

Figure 6.4 (a) compares the minimal prediction errors over time among the four experiments. Clearly, with the same α setting, the independent LS selection yielded a lower prediction error than the dependent selection runs in the early stages. However, this significance diminished as time went by. One explanation is that excessive LS operations yield a premature convergence. When comparing the results among the last three experiments, an increase in α causes a better performance in the early stages, whereas a smaller α yields the best performance in the late stages. Again, this observation reflects that frequent LS operations may have a negative impact in the early stages owing to premature convergence.

Figure 6.4 (b) presents the number of alternatives as a function of the elapsed time since the first detection. Evidently, the smallest number of alternatives is generated by the independent LS selection due to the early convergence. Another observation from Figure 6.4 (b) is that, in accordance with Figure 6.4 (a), a higher value of α results in more alternatives in the early stages, indicating an improvement in the solution quality as a consequence of infrequent LS operations. Taking other subpopulations into consideration when assessing the need for a LS for a subpopulation is beneficial for maintaining diversity during the search. One final note on these experiments in this study is that the dependent LS selection strategy with $\alpha = 0.5$ seems to achieve the best performance in terms of both the solution quality and reduction of computational costs.



(a)



(b)

Figure 6.4 Comparison of the results between LS selection strategies: (a) minimal prediction error; (b) number of alternatives

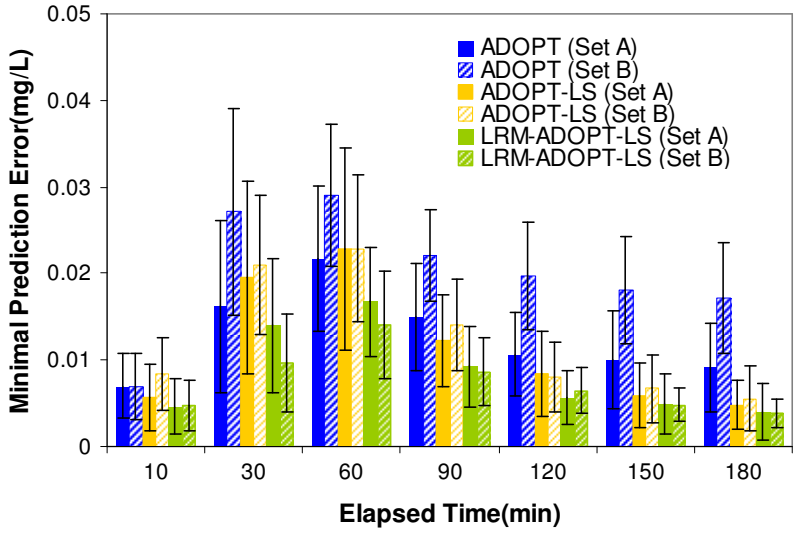
6.4.3 Effect of Mutation Sizes

Small mutation sizes may be valuable for a global optimizer because they cause a reduction in randomness and potentially allow faster convergence. However, this observation does not hold true when the search scope reduces to a small area or in the context of the EA + LS hybrid. Increasing the mutation size is expected to introduce more diversity in a small region. Moreover, in the context of an EA in combination with the LS, the LS aims to fine-tune local regions, whereas the ES works to explore the promising areas. Accordingly, too small mutation sizes would unavoidably yield premature convergence and loss of diversity in the decision space due to too much refinement. Large mutation sizes are expected when the EAs are combined with a LS, which allows the EA to focus mainly on exploration rather than local refinement.

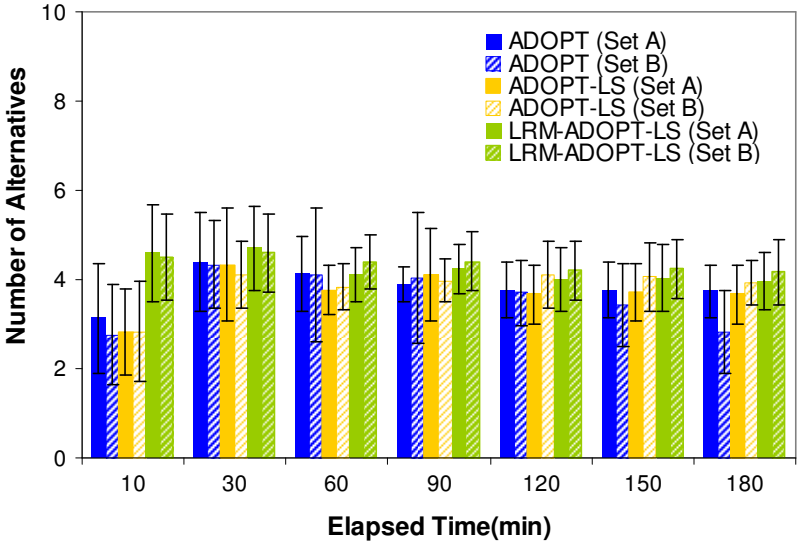
To illustrate the effects of mutation size with the addition of the LRM and LS, two sets of allowable ranges for mutation size were employed, as listed in Table 6.3. The pure ADOPT, ADOPT-LS, and LRM-ADOPT-LS methods were conducted on Sets A and B. Figure 6.5 (a) presents the minimal prediction error versus the elapsed time. It is interesting to note that increasing mutation sizes have a negative effect on the pure ADOPT, and yet they improved the performance of the hybrid approach. Further evidence of this observation is shown in Figure 6.5 (b), which presents the number of alternatives generated for the three approaches. Increasing mutation sizes aid in the discovery of more non-unique solutions for the hybrid method, whereas they cause a decrease in the alternatives generated by ADOPT.

Table 6.3 Allowable Ranges of Mutation Sizes

Mutation Size	Set A	Set B
Location	[0, 10]	[0, 20]
Starting Time	[0, 16]	[0, 32]
Duration	[0, 6]	[0, 12]
Mass Injection Rate	[0, 10]	[0, 20]



(a)



(b)

Figure 6.5 Effects of mutation sizes: (a) minimal prediction error; (b) number of alternatives

6.5. Summary

This chapter provides two enhancements to the ES-based ADOPT for characterizing contaminant sources in a WDS. Prior to employing ADOPT, an LR-based prescreening step is performed to reduce the search space. Then, a pattern move-based LS operator, which serves as the refinement operation, is embedded into ADOPT in order to accelerate the convergence in local regions. The pre-established LRMs allow fast estimation of potential source locations, given currently available sensor observations. Moreover, the pattern move as an operator subsequent to the selection operation concentrates on moving towards local optima without additional costs for objective evaluations.

The results of this research provide several interesting observations. First, the use of the LRM and LS successfully promotes the algorithmic performance of ADOPT. However, the degree of improvements made by the LRM and LS depends on the problem. Overall, both enhancements facilitate ADOPT to operate more efficiently by eliminating unnecessary nodes and fine-tuning the process. A high solution quality does not always correspond to a low computational cost, however. According to the analysis, the hybrid approach typically has a faster convergence in the early stages than ADOPT alone. However, the hybrid approach appears to contain more subpopulations in the late stages due to more identified alternatives. The reported results provide evidence that the hybrid approach can improve the solution quality as well as the computational efficiency, especially in the context of a large network.

Interestingly, large mutation sizes corresponded to the worst performance for the pure ADOPT, while they assisted in identifying more alternatives in the case of the framework

that is coupled with the LRM and LS. The decision whether or not to incorporate the LS into ES-based ADOPT establishes the role of the ES during the search process. In the context of the hybrid approach, the ES is required to concentrate on exploring new promising areas because the LS typically performs efficiently as a hill-climber. Looking at the analysis in terms of the effect of mutation sizes, it is clear that an increase in the mutation size has a positive effect on the solutions when ADOPT is coupled with the LRM and LS.

The strategy of LS selection plays an important factor in the algorithm's performance. The more frequently the LS is conducted in the early stages, the worse the performance. Assessing the need of the LS, based on the extent of individual convergence, can help avoid inappropriate or unnecessary LS operations and enable the function of the LS in an adaptive manner. Avoiding repeated LS operations at the same node increases the diversity of the solutions. Alternatively stated, excessive LR operations on the same node may yield fast convergence at the expense of a quick loss of diversity. However, a better understanding of ways to apply the LS will require studying the performance under different problem scenarios. Thus, further detailed investigations are planned to explore the effects of the LS methods, LS selection, and various degrees of complexity of the problem scenarios.

CHAPTER 7. Summary and Final Remarks

This research develops and presents several quantitative approaches to characterize water distribution system (WDS) contaminant sources in real time. Accidental and/or intentional contamination of water distribution networks is an increasingly critical issue. Therefore, the development of effective source characterization algorithms that can be used to facilitate appropriate decisions in response to a threat is necessary. Given available sensor observations, contaminant source characterization can be formulated as an optimization problem to minimize the discrepancy between the simulated concentration values and the observations. This research concentrates on simulation-optimization approaches to offer an accurate, quick, and robust characterization.

Several issues pose challenges to real-time solutions for the contaminant characterization. The sparseness of available observations and the complexities inherent to a WDS typically lead to non-uniqueness in the solutions. In addition to complicating the characterization, the dynamic system and observed data also require real-time solutions in an adaptive manner. An evolution strategy (ES)-based Adaptive Dynamic OPtimization Technique (ADOPT) has been designed to address the non-uniqueness factor and adaptively search for optimal solutions in a dynamic environment. A powerful aspect of ADOPT is that alternatives can be generated that yield similar explanations to the observations. Using current sets of solutions to seed subsequent searches enhances ADOPT's efficiency during the search process. Consequently, the proposed method is capable of adaptively describing the contaminant source by identifying one solution or a set of optimal solutions to match the

observations that are available at any given time. The application of ADOPT to two illustrative case studies with different degrees of complexity illustrates its identification accuracy and its potential to address non-uniqueness. Although a mutative self-adaptation scheme is beneficial to an ES, re-initialization of mutation sizes when an environmental change occurs is shown to be increasingly advantageous to the search as time progresses.

Furthermore, the proposed methodology is expected to be applicable to a generic dynamic optimization problem with shifting optima over time. In such a case, ADOPT is expected not only to capture the current optima, but also simultaneously preserve proper diversity. A set of potential solutions identified at the current time by ADOPT assists subsequent searches. The adaptive capability of ADOPT is revealed through its application to a moving peaks function problem and a groundwater contaminant source identification problem in dynamic environments.

Due to high levels of uncertainty associated with potential contaminant characteristics, sensor measurements, and water consumption, a probabilistic description of contaminant sources is desirable. A logistic regression (LR)-based estimation method is employed to describe the relationship between the contaminant source characteristics and their resulting sensor observations. The use of LR analysis allows direct and fast estimations of the likelihood that any given node is a source, given the sensor observations. Illustrative applications via two example water networks help to evaluate the algorithm's performance. From the analysis conducted on a large number of hypothetical contamination events, the generated results indicate that the logistic regression modes (LRMs) are able to identify the true injection node as a potential solution and effectively eliminate improbable nodes.

Additionally, more measurements across space or time are shown to improve the identification uncertainty with respect to the rank of the true source as well as the number of possible solutions. The reuse of models from a previous time is shown to produce comparable results to the independent generation of models, with a significant reduction in computational costs for model building.

In addition to yielding location-specific probability estimations, the LR analysis serves as a prescreening step for the simulation-optimization methods. The estimated probabilities are used to reduce the decision space, thus alleviating the computational burden related to the search procedure. Whereas the LRMs facilitate the removal from the search of unlikely nodes as sources, the subsequent local search (LS) procedure enables the efficient characterization of the contaminant in the localized region. The application of the LRM-LS method to two water networks includes not only various contamination scenarios but also variations imposed on the quality of the monitoring sensors.

The generated results suggest that the LRM-LS procedure can adaptively discover a set of potential locations along with corresponding release histories. However, the identification accuracy is affected by the location of the contaminant, the monitoring data, and the network size. Although identification becomes more difficult according to the uncertainties related to the monitoring data, water consumption, and problem scenarios, the possible identified solutions typically concentrate in the vicinity of the true source.

Although ADOPT is shown to be a promising application, this simulation-optimization approach is computationally expensive, which poses challenges in the context of real-time solutions. An extension of this algorithm is presented whereby it is integrated

with LR analysis and a local improvement method to expedite the convergence and potentially solve the problem faster. Prior to employing ADOPT, a LR analysis step is performed as a prescreening technique to reduce the search space by eliminating unnecessary source nodes for consideration as potential source locations. Then, a LS approach is embedded into some algorithmic steps in ADOPT to serve as a postscreening step that potentially speeds up the convergence in localized regions in the solution space. Numerical experiments for the proposed hybrid approach were performed on two illustrative example water networks. The new method is shown to reduce prediction error and augment alternatives considerably without additional, unnecessary computational costs, as compared to the standard implementation of ADOPT. Moreover, the effects of mutation size become insignificant in the context of ADOPT coupled with the LRM and LS.

In this research, the solutions generated by ADOPT and the LS assume that no uncertainty is associated with measurements and forward model simulations. In reality, however, a certain level of uncertainty is always involved in these factors. For example, binary decision sensors help only in knowing the status of the contamination. Moreover, a chance that the sensors will detect false positives or false negatives is also possible. These uncertainties may significantly affect the solutions or take the search in a wrong direction. Although the use of LRMs presents the likelihood that any given node is the contaminant source, LRMs are unable to provide the probabilistic characterization of the release history at a potential location. To incorporate these various uncertainties into contaminant characterization and add a probabilistic measure to the obtained results, future work could consider the Bayes theorem that offers the capacity to combine prior probability and current

observations. The estimated probabilities from the LRMs could serve as the prior information to infer the distribution of the solutions.

Another unrealistic assumption that is made in this research is that of deterministic demands. In a real water network, water consumption at a given node is often unknown except for limited hydraulic observed data (e.g., pressure in a junction, flow in a pipe). Consequently, solutions should be considered and adapted in a noisy environment where multiple demand realizations can be applied. To improve the algorithm's efficiency, the representative demand realizations can be generated from the hydraulic observations, and the realization that best resembled the monitored data can then be used for each generation of the solution evolution.

Although ADOPT is demonstrated herein for a test problem and two particular real-world optimization problems, the proposed methodology is general enough to be applicable to other time-dependent real-world optimization problems where a slight change is made between consecutive time steps. By coupling the simulation models, investigations into a range of realistic situations with changing optimal solutions may be carried out in the future to gain more insight into the merits and drawbacks of this algorithm.

REFERENCES

Belegundu, A. D. and T. R. Chandrupatla. (1999). Optimization Concepts and Applications in Engineering. Upper Saddle River, NJ: Prentice Hall.

Branke, J. (1999). "Memory Enhanced Evolutionary Algorithms for Changing Optimization Problems." In *Proceedings of the IEEE Congress on Evolutionary Computation*: 1875-1882.

Brank, J. (2002). Evolutionary Optimization in Dynamic Environments. Boston: Kluwer Academic Publishers

Brumbelow, K., J. Torres, S. Guikema, E. Bristow, and L. Kanta. (2007). "Virtual Cities for Water Distribution and Infrastructure System Research." In *Proceedings of World Environmental and Water Resources Congress*, Tampa, Florida, May 15-19.

Bui, L. T., J. Branke, H. A. Abbass. (2005). "Diversity as a Selection Pressure in Dynamic Environments." In *Proceedings of the Genetic and Evolutionary Computation Conference*: 1557-1558.

Bui, L. T., J. Branke, H. A. Abbass. (2005). "Multiobjective Optimization for Dynamic Environments." In *Proceedings of the Congress on Evolutionary Computation*: 2349-2356, Edinburgh, UK: IEEE Press.

Clark, R. M. and R. A. Deininger. (2000). "Protecting the Nation's Critical Infrastructure: The Vulnerability of U.S. Water Supply Systems." *Journal of Contingencies and Crisis Management*, 8 (2).

Dandy, G. C., A. R. Simpson, and L. J. Murphy. (1996). "An Improved Genetic Algorithm for Pipe Network Optimization." *Water Resources Research*, 32 (2): 449-458.

De Sanctis, A. E., F. Shang, J. G. Uber. (2006). "Determining Possible Contaminant Sources through Flow Path Analysis." In *Proceedings of Water Distribution Systems Analysis Symposium*, Cincinnatti, OH.

Espinoza, F. P., B. S. Minsker, D. E. Goldberg. (2005). "Adaptive Hybrid Genetic Algorithm for Groundwater Remediation Design." *Journal of Water Resources Planning and Management*, 131 (1): 14-24.

Ghosh, A., S. Tstutsui, and H. Tanaka. (1998). "Function Optimization in Nonstationary Environment using Steady State Genetic Algorithms with Aging of Individuals." In *Proceedings of the IEEE International Conference on Evolutionary Computation*: 667-671.

Goldberg, D. E. and J. Richardson. (1987). "Genetic Algorithms with Sharing for Multimodal Function Optimization." In *Proceedings of the 2nd International Conference on Genetic Algorithms*: 41-49.

Goldberg, D. E. and R. E. Smith. (1987). "Nonstationary Function Optimization using Genetic Algorithms with Dominance and Diploidy." In *Proceedings of the 2nd International Conference on Genetic Algorithms*: 59-68.

Grefenstette, J. J. (1992). "Genetic Algorithms for Changing Environments." In R. Maenner and B. Manderick, editors *Parallel Problem Solving from Nature*: 137-144.

Guan, J., M. M. Aral, M. L. Maslia, and W. M. Grayman. (2006). "Identification of Contaminant Sources in Water Distribution Systems Using Simulation–Optimization Method: Case Study." *Journal of Water Resources Planning and Management*, 132 (4): 252-262.

Hart, W. E. (1994). "Adaptive Global Optimization with Local Search." Ph.D. Thesis, University of California, San Diego.

Hill, J., B. G. Van Bloemen Waanders, and C. D. Laird. (2006). "Source Inversion with Uncertain Sensor Measurements." *Water Distribution Systems Analysis Symposium*, Cincinnatti, OH.

Hoffmeister, F. and T. Back. (1992). "Genetic Self-Learning." In *Proceedings of the 1st European Conference on Artificial Life*: 227-235.

Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press.

Hosmer, D. S. and S. Lemeshow. (1989). Applied Logistic Regression. Hoboken, NJ: John Wiley.

Laird, C. L., L. T. Biegler, B. G. Van Bloemen Waanders, and R. A. Bartlett. (2005). "Contamination Source Determination for Water Networks." *Journal of Water Resources Planning and Management*, 131 (2): 125-134.

Land, M. W. S. (1998). "Evolutionary Algorithms with Local Search for Combinatorial Optimization." Ph.D. Thesis, University of California, San Diego.

Lingireddy, S. and L. Ormsbee. (2002). "Hydraulic Network Calibration Using Genetic Optimization." *Civil Engineering and Environmental Systems*, 19 (1).

Liu, L., E. M. Zechman, E. D. Brill, G. Mahinthakumar, S. Ranjithan, and J. Uber. (2006). "Adaptive Contamination Source Identification in Water Distribution Systems Using an Evolutionary Algorithm-based Dynamic Optimization Procedure." In *Proceedings of Water Distribution Systems Analysis Symposium*, Cincinnatti, OH.

Lu, X., J. T. Wilson, and D. H. Kampbell. (2006). "Relationship Between Geochemical Parameters and the Occurrence of Dehalococcoides DNA in Contaminated Aquifers." *Water Resources Research*, 42 (16): 3131–40.

Mahar, P. S. and B. Datta. (1997). "Optimal Monitoring Network and Ground-water-pollution Source Identification." *Journal of Water Resources Planning and Management*, 123 (4): 199-207.

Mahinthakumar, G. K. and M. Sayeed. (2005). "Hybrid Genetic Algorithm – Local Search Methods for Solving Groundwater Source Identification Inverse Problems." *Journal of Water Resources Planning and Management*, 131 (1): 45-57.

Nelder, J. A. and R. Mead. (1965). A Simplex Method for Function Minimization. *Computer Journal*, 7: 308-313.

Oppacher, F. and Wineberg, M.(1999). The Shifting Balance Genetic Algorithm: Improving the GA in a Dynamic Environment. In *Proceedings of the Genetic and Evolutionary Computation Conference*, Vol. 1: 504–510.

Partheepan, R. (2003). "Hybrid Genetic Algorithms." M. Sc. Thesis, North Carolina State University, Raleigh, NC.

Regonda, S. K., B. Rajagopalan, and M. Clark. (2006). "A New Method to Produce Categorical Streamflow Forecasts." *Water Resources Research*, 42 (9): 9501-06.

Rossman, L. A. (2000) EPANET User's Manual. Risk Reduction Engineering Laboratory, U.S. Environmental Protection Agency, Cincinnati, OH.

Savic, D. A. and G. A. Walters. (1997). "Genetic Algorithms for Least-Cost Design of Water Distribution Networks." *Journal of Water Resources Planning and Management*, 123 (2): 67-77.

Schwefel, H. P. (1995). Evolution and Optimum Seeking. New York: John Wiley and Sons.

Ursem, R. K. (2000). Multinational gas: Multimodal optimization techniques in dynamic evironments. In P*roceedings of the Second Genetic and Evolutionary Computation Conference*.

Van Bloemen Waanders, B. G., R. A. Bartlett, L. T. Bigler, and C. D. Laird. (2003). "Nonlinear Programming Strategies for Source Detection of Municipal Water Networks." In *Proceedings of the ASCE World Water and Environmental Congress*, Philadelphia, PA, June 23-26.

Vitkosvsky, J. P., A. R. Simpson, and M. F. Lambert. (2000). "Leak Detection and Calibration using Transients and Genetic Algorithms." *Journal of Water Resources Planning and Management*, 126 (4): 262-265.

Xu, Y. G., Li, G. R., Wu, Z. P. (2001). "A Novel Hybrid Genetic Algorithm using Local Optimizer based on Heuristic Pattern Move." *Applied Artificial Intelligence*, 15: 601-631.

Yang, S. (2005). "Memory-based Immigrants for Genetic Algorithms in Dynamic Environments." In *Proceedings of the Genetic and Evolutionary Computation Conference*: 1115-1122.

Yang, S. (2005). "Population-based Incremental Learning with Memory Scheme for Changing Environments." In *Proceedings of the Genetic and Evolutionary Computation Conference*: 711-718.

Yang, S., Y. Ong, and Y. Jin, Eds. (2007). Evolutionary Computation in Dynamic and Uncertain Environments. Berlin, London: Springer.

Yeh, H. D., T. H. Chang, Y. C. Lin. (2007). "Groundwater Contaminant Source Identification by a Hybrid Heuristic Approach." *Water Resources Research*, 43, W09420.

Zechman, E. M., and S. Ranjithan. (2004). "An Evolutionary Algorithm to Generate Alternatives (EAGA) for Engineering Optimization Problems." *Engineering Optimization*, 36 (5): 539-553.

Zechman, E. M. and S. Ranjithan, (2007). "Evolutionary Computation-based Approach for Model Error Correction and Calibration," Advances in Water Resources, 30(5): 1360-1370.