# Optimal Identification of Unknown Groundwater Contaminant Sources in Conjunction with Designed Monitoring Networks

Thesis submitted by

**Shahrbanoo Hazrati Yadkoori**

**June 2018**

**For the degree of Doctor of Philosophy**

**College of Science and Engineering**

**James Cook University**

## Acknowledgments

First, I would like to sincerely honour my supervisor, Dr Bithin Datta. It has been a great privilege to be a member of his team and benefit from his knowledge and experience. Undoubtedly, without his continuous support and guidance submission of this thesis would not have been possible. I also wish to thank Associate Professor N. Sivakugan for his support as my co-supervisor.

Secondly, I am particularly grateful to the CRC-CARE, the Graduate Research School (GRS) and the College of Science and Engineering, James Cook University, Australia, for their financial support of my research and living expenses.

I would like to express my great appreciation to Professor Helene Marsh, the former Dean of Graduate Research Studies for academic and financial support. I also would like to offer my special thanks to Dr Elizabeth Tynan, senior lecturer, and coordinator at the GRS Professional Development Program for her technical writing and presentation advice throughout the course of this study.

My sincere thanks goes also to my dear husband, Mr Hooshang Shamshiri, and my beloved daughters, Eliana and Nirvana, whose love and support helped me to complete this research. I am also thankful to all my family and fellow friends, especially other PhD candidates in Engineering at James Cook University, for all their support throughout this study.

## Statement of Contributions of Others

I officially declare that all the procedures, concepts, data analysis and results presented in this thesis have been developed and written by Shahrbanoo Hazrati Yadkoori under the supervision of my primary supervisor Dr Bithin Datta. The calibrated flow and transport simulation models utilized in Chapter 6 were developed by Prakash (2014).

Financial support of this research was provided by CRC-CARE, University of Newcastle, Callaghan, Australia, and James Cook University, Australia.

Professional editing and proofreading of this thesis were carried out with the assistance of Dr John Cokley, *EduPreneur Services International.*

## Abstract

Human activities and improper management practices have resulted in widespread deterioration of groundwater quality worldwide. Groundwater contamination has seriously threatened its beneficial use in recent decades. Remediation processes are necessary for groundwater management. In the remediation of contaminated aquifer sites, identification of unknown groundwater contaminant sources has a crucial role. In other words, an effective groundwater remediation process needs an accurate identification of contaminant sources in terms of contaminant source locations, magnitudes and time-release. On the other hand, the efficiency and reliability of contaminant source identification depend on the availability, adequacy, and accuracy of hydrogeologic information and contaminant concentration measurements data. Whereas, generally when groundwater contaminations are detected, only limited and sparse measured contaminant concentration values are available. Usually, groundwater contaminations are detected after a long time, years or even decades after the starting of contaminant source activities or even after their extinction. Therefore, usually, there is not enough information regarding the number of contaminant sources, the duration of sources' activities and the contaminant magnitudes, as well as the hydrogeologic parameters of the contaminated aquifers. Simulations of groundwater flow and solute transport involve intrinsic uncertainties due to this sparse information or lack of enough hydrogeologic information of the porous medium. Therefore, for groundwater management, developing and applying an efficient procedure for identification of unknown contaminant sources is essential.

Moreover, available observed contaminant concentration values are usually erroneous and this erroneous data could cause instability in the solution results. Various combinations of source characteristics can result in similar effects at observation locations and cause non-uniqueness in the solution. Due to these instabilities and non-uniqueness in solution (Datta, 2002), the

source identification problem is known as an "ill-posed problem" (Yeh, 1986). The non-uniqueness and uncertainties involved in this ill-posed problem make this problem a difficult and complex task. Suggested methodologies to tackle this task are not completely efficient. For instance, the crux of previous approaches is highly vulnerable to the accuracy and adequacy of contaminant concentration measurements and hydrogeologic data. As a result, many of the previously suggested approaches are not applicable to real-world cases and application of relevant approaches to real-world contaminant aquifer sites is usually tedious and time-consuming. The suggested methodologies involve enormous computational time and cost due to repeated runs of the numerical simulation models within the optimisation algorithms.

Therefore, to identify the unknown characteristics of contaminant sources, different surrogate models were developed. Three different algorithms were utilized for developing the surrogate models: Self-Organising Maps (SOM), Gaussian Process Regression (GPR), and Multivariate Adaptive Regression Splines (MARS). Performance of the developed procedures was assessed for potential applicability in two hypothetical, an experimental, and a real-world contaminated aquifer sites. In the used contaminated aquifer sites, only limited contaminant concentrations data were assumed to be available. In three cases, it was also assumed that the contaminant concentrations data were collected a long time after the start of the first potential contaminant source activities.

The performance evaluations of the developed surrogate models show that these models could accurately mimic the behaviour of simulation models of groundwater flow and solute transport. These surrogate models solutions showed acceptable errors in comparison to the more robust numerical model solutions. These surrogate models were also used for identification of unknown groundwater contaminant sources when utilized to solve the inverse problem. The SOM algorithm was chosen as the surrogate model type in this study for directly addressing

the source identification problem as well. The SOM algorithm was chosen for its classification capabilities. In source identification problems, the number of actual contaminant sources is uncertain and usually, a set of a larger number of potential contaminant sources are assumed. Therefore, screening the active sources by SOM-based Surrogate Models (SOM-based SMs) may simplify the source identification problems. The performance of the developed SOM-based SMs was assessed for different scenarios. Results indicate that the developed models could also accurately screen the active sources among all potential contaminant sources with sparse contaminant concentrations data and uncertain hydrogeologic information.

For comparison purposes, MARS and GPR algorithms that are precise prediction tools were also utilized for developing MARS and GPR-based Surrogate Models (MARS and GPR-based SM) for source identification. Performance of the developed surrogate models for source identification was evaluated in terms of Normalized Absolute Error of Estimation (NAEE). For example, the performance of the developed SOM, MARS and GPR-based SMs was assessed in an illustrative hypothetical contaminated aquifer site. The results for testing data in terms of NAEE were equal to 16.3, 4.9 and 6.6%, respectively. Performance of the developed SOM, MARS and GPR-based SMs was also evaluated in an experimental contaminated aquifer site. The results for testing data in terms of NAEE were equal to 15.8, 14.1 and 16.2%. These performance evaluation results of the developed surrogate models indicate that the MARS-based SMs can be more accurate models than the SOM and GPR-based SMs in source identification problems. The most important advantage of the developed methodologies is their direct application for source identification in an inverse mode without linking to an optimisation model.

Surrogate Model-Based Optimisation (SMO) was also developed and utilized for source identification. In this developed SMO, MARS and Genetic Algorithm (GA) were utilized as the surrogate model and the optimisation model types, respectively. MARS-based SMOs performance was assessed in an illustrative hypothetical contaminated aquifer site and in a real-world contaminated aquifer site. The result of the developed MARS-based SMO for testing data in the illustrative hypothetical contaminated aquifer site in terms of Root Mean Square Error (RMSE) was equal to 0.92. Obtained solution results of the developed MARS-based SM in the real contaminated study area for testing data in terms of RMSE was equal to 42.5. The performance evaluation results of the developed methodologies in different hypothetical and real contaminated study areas demonstrate the capabilities of the constructed SOM, GPR, and MARS-based SMs and MARS-based SMO for source identification. Also, in order to increase the accuracy of source identification results, and based on the preliminary solution results of the developed SOM-based SMs, a sequential sampling method can be applied adaptively for updating the developed surrogate models. Information from a hypothetical contaminated aquifer site was used to assess the performance of this procedure. Performance evaluation results of adaptively developed MARS and GPR-based SMs in terms of NAEE were equal to 1.9 and 2.1%, respectively. The results show 3 and 4.5% improvements for source identification results by applying adaptively developed MARS and GPR-based SMs, respectively.

Another difficulty with source identification problems has been the limitation and sparsity of observed contaminant concentrations data. Previously suggested methodologies usually need long-term observation data at numerous locations which can involve large costs. Therefore, developing an effective monitoring network design procedure was one of the main goals of this study. In designing the monitoring networks, two main objectives were considered: 1.

Maximizing the accuracy of source identification results, and 2. Limiting the number of monitoring locations. It was supposed that by implementing obtained results from the designed monitoring networks for developing surrogate models, the source identification results would significantly improve. In this study, different algorithms were utilized to identify potentially important and effective monitoring locations which probably could improve source identification results. These algorithms are Random Forests (RF), Tree Net (TN) and CART. The performance of these algorithms was evaluated in different scenarios. Results indicate the potential applicability of these algorithms in recognising the most important components of prediction models. As a result, these algorithms could apply for designing monitoring networks for improving the source identification efficiency and accuracy. Concentration measurement information from a designed monitoring network and from a set of arbitrary monitoring sites was utilized to develop MARS-based surrogate models for source identification. The solution results for these two scenarios of designed monitoring and arbitrary measurements were compared for a hypothetical study area for evaluation purpose. Performance evaluation results of the developed surrogate model using information from the designed monitoring network showed improvement in source identification error in terms of RMSE for testing data by 0.7. The obtained information from the designed monitoring network was used to develop MARS-based SM for source identification of testing data in a real contaminated aquifer site. Source identification results of the developed MARS-based SM with testing data for the real contaminated aquifer site showed improvement by 35.3 in terms of RMSE compared to the solution results of MARS-based SM, which was developed by using obtained information from arbitrary monitoring locations. Performance evaluation results for the developed monitoring network procedure demonstrate the potential applicability of this procedure for source identification.

## Table of Contents:

Abstract……..iii

Table of Contents……..viii

List of Figures……..……xi

List of Tables……..……xiii

**1. Introduction.....1**

**1.1. Overview.....1**

**1.2. Objectives.....6**

**1.3. Organisation of the Thesis.....10**

**2. Literature Review.....13**

**2.1. Introduction.....13**

**2.2. Developed Methodologies for Source Identification.....14**

2.2.1. Statistical Approaches.....15

2.2.2. Approaches based on Optimisation Algorithms.....18

2.2.2.1. Response Matrix.....18

2.2.2.2. Embedded optimisation techniques.....19

2.2.2.3. Linked simulation-optimisation approaches.....20

2.2.2.4. Surrogate Models.....25

**2.3. Monitoring Network Design Procedures.....28**

**2.4. Source Identification Procedures in Conjunction with Monitoring Network Design Procedures.....31**

**2.5. Self-Organising Maps.....34**

**2.6. Motivation for this Study.....36**

**3. Contaminant Source Identification by Utilizing Adaptive Surrogate Models.....40**

**3.1. Introduction.....40**

**3.2. Developed Procedures for Source Identification.....41**

3.2.1. Surrogate Models for Contaminant Source Identification.....43

3.2.2. Numerical Simulation Models.....45

3.2.3. Self-Organising Map.....47

3.2.4. Multivariate Adaptive Regression Splines (MARS).....49

3.2.5. Gaussian Process Regression (GPR).....51

3.2.6. Assessment of the Performance of the Developed Models.....52

**3.3. Application of the Developed Surrogate Models for Source Identification.....52**
3.3.1. Study Area .....52
3.3.2. Results.....55
**3.4. Discussion.....68**
**3.5. Conclusion .....70**
**4. Application of Surrogate Model based Optimization in Conjunction with Monitoring Network Design Procedure for Source Identification.....72**
**4.1. Introduction.....72**
**4.2. An Overview of the Source Identification Problem and Previously Applied Methodologies .....73**
**4.3. Methodology .....75**
4.3.1. Surrogate Models .....76
4.3.2. Simulation Models .....79
4.3.3. Designing a Monitoring Network .....79
4.3.3.1. Random Forests (RF) .....80
4.3.3.2. Classification and Regression Trees (CART) .....81
4.3.3.3. TreeNet (TN) .....81
4.3.3.4. Designing Monitoring Network Procedure .....81
4.3.4. Optimisation Model .....87
4.3.5. Performance Evaluations of the Developed Procedures.....88
**4.4. Application of the Developed Procedure for Source Identification .....90**
4.4.1. Study Area .....90
4.4.2. Performance Evaluation Results.....93
**4.5. CONCLUSION .....99**
**5. Verification of the Developed Procedures for Source Identification by using data from an Experimental Aquifer Site.....102**
**5.1. Introduction.....102**
**5.2. Methodology .....103**
5.2.1. Surrogate Models .....103
5.2.2. Simulation Models.....105
**5.3. Application of the Developed Procedures for Source Identification.....106**
5.3.1. Study Area .....106
5.3.2. Site Description, Eastlake Experimental Site .....108

5.3.3. Tracer Test and Movement of a Conservative Element....110
5.3.4. Simulation Models....111
5.3.5. Performance Evaluation Results....115
**5.4. Conclusion....124**
**6. Source Identification by Using Surrogate Models based Optimisation in Conjunction with Monitoring Network Design Using Field Data....127**
**6.1. Introduction....127**
**6.2. Methodology....127**
6.2.1. Surrogate Models....127
**6.3. Integration of the Developed Source Identification Methodologies with the Designed Monitoring Network Approach....129**
**6.4. Contaminated Aquifer Site....130**
**6.5. Study Area....131**
**6.6. Simulation Models....133**
**6.7. Performance Evaluations of the Developed Methodologies....135**
6.7.1. Application of the Developed Monitoring Network Design Procedure....135
6.7.2. Developing Surrogate Models and SMO for Source Identification....140
**6.8. Conclusion....146**
**7. Summary and Conclusions....149**
**7.1. Introduction....149**
**7.2. Summary....149**
**7.3. Conclusions....151**
**7.4. Recommendations for Future Research....154**

## List of Figures

Figure 1.1. Annual groundwater consumption changes in Australia at three different times (Harrington & Cook, 2014) ..... 1

Figure 1.2. Groundwater consumption in different states of Australia (Harrington & Cook, 2014) ..... 2

Figure 3.1 Key elements of the ASM methodology for source identification as an inverse problem ..... 43

Figure 3.2.a).The SOM algorithm's process in classification and visualisation, b) The SOM algorithm's process in the prediction of missing values of system's new input vectors. ..... 49

Figure 3.4 Breakthrough cureves at the observation wells used for source identification ..... 55

Figure 3.5 A typical concentration plume 732 days after start of first source activity ..... 58

Figure 3.6 The result obtained from the selected SOM-based SM for estimating the contaminant concentration values for a set of test data at six observation wells (NAEE is equal to 16.4%) ..... 61

Figure.3.7 Required times for constructing various SOM-based SMs for different scenarios 62

Figure 3.8 Source identification results of the developed surrogate models ..... 64

Figure 3.10 Obtained results by using the GPR-based ASM for source identification and its 95% source estimation intervals for observed contaminant concentration values ..... 66

Figure 4.1 Schematic chart of the developed SMO in conjunction with monitoring network design approach for source identification ..... 77

Figure 4.2 Schematic diagram of the applied monitoring network design procedure using RF, TN and CART algorithms ..... 83

Figure 4.3 Illustrative study area representing typical concentration plumes 4234 days after start of first source activity (concentration values g/l) ..... 91

Figure 4.4 Breakthrough curves at initial arbitrary monitoring wells used for source identification ..... 92

Figure 4.5 Breakthrough curves at selected monitoring wells used for source identification . 96

Figure 4.6 Comparison of the obtained results for source identification by utilizing MARS-based SMO and using the information from initial arbitrary and selected monitoring wells with actual data. ..... 98

Figure 4.7 Generated hydraulic conductivity for layer 1 ..... 99

Figure 5.1. Flow chart of the main steps of developing surrogate models for source identification ..... 104

Figure 5.2. The East Lake Experimental Site location (ELE site) at the Botany Sands aquifer (Beck, 2000) ..... 107

Figure 5.3. The ELE site adjacent to the Lachlan Ponds (Beck, 2000) ..... 109

Figure 5.4. Layout of ELE site showing injection well locations, multilevel piezometers and water level piezometers (Jankowski & Beck, 2010) ..... 109

Figure 5.5 Geological cross-section of the ELE site along line D (Beck, 2000) ..... 110

Figure 5.6 Generated hydraulic conductivity for layer three, iteration two; applying the IDW interpolation algorithm (m/day) ..... 115

Figure 5.7 The performance evaluation results of the developed SOM-based SMs for various scenarios representing different numbers of SOM map units in terms of NAEE and QE values 120

Figure 5.8 The required time for constructing different SOM-based SMs representing different numbers of SOM map units ..... 120

Figure 5.9 The obtained results of the constructed surrogate model for source identification using testing data in terms of NAEE ..... 122

Figure 5.10 comparison of actual data with obtained results of selected SOM, MARS and GPR based SMs for source characterisation in terms of NAEE ..... 123

Figure 5.11 Comparison of actual data with obtained results of of the developed surrogate models for source identification in terms of NAEE ..... 124

Figure 6.1 Plan view of the study area and the contaminated area (Prakash & Datta, 2015) 132

Figure 6.2 Schematic chart of the developed monitoring network design methodology using RF, TN and CART tools ..... 136

Figure 6.3 Breakthrough curves of specific monitoring wells at specific times utilized for source identification in the contaminated aquifer ..... 144

Figure 6.4 The preliminary obtained results of the developed SOM-based SMs representing different number of monitoring locations for source identification by using the observed contaminant concentration data ..... 145

## List of Tables

Table 3.1 Aquifer characteristics and dimensions of the study area.....54
Table 3.2 Characteristics of the contaminant sources.....54
Table 3.3 Characteristics of extraction wells.....54
Table 3.4 Typical sample sets for training a surrogate model.....57
Table 3.5 Typical sample sets with missing data for testing a surrogate model.....57
Table 3.6. Normalized Absolute Error of Estimation for different developed SOM-based SMs 62
Table 4.1 Typical input vectors using in the RF, TN and CART prediction models.....84
Table 4.2 Typical results of ranking monitoring wells by using the RF, TN and CART algorithms according to their importance in improving the source identification results.....86
Table 4.3 Hydrogeologic parameter values and the dimensions (in metre (m)) of the study area 91
Table 4.4 Locations and flux magnitudes of actual contaminant sources.....92
Table 4.5 Locations of monitoring wells.....92
Table 4.6 Performance evaluation results obtained for testing data by using MARS-based surrogate model in terms of RMSE.....94
Table 4.7 Ranked potential monitoring wells according to their expected influence on source identification by using the RF, CART and TN algorithms.....95
Table 4.8 Performance evaluation results of the developed surrogate models for testing data in terms of RMSE.....97
Table 5.1 Hydrogeological information of the experimental study area.....112
Table 5.2 The monitoring locations and observed concentration values.....113
Table 5.3 A typical input for training a surrogate model.....116
Table 5.4 A typical input vector with missing data for testing the developed surrogate models 119
Table 6.1 Hydrogeologic characteristics of the contaminated study area (Prakash, 2014).....133
Table 6.2. The grid locations of potential contaminant sources.....134
Table 6.3. Typical input vectors using in the RF, TN and CART prediction models.....139
Table 6.4. Ranked potential monitoring locations based on their contributions to source identification by using RF, CART and TN.....140
Table 6.5. Typical input data for training a surrogate model.....142
Table 6.6. The performance evaluation results of the developed surrogate models of testing data in terms of RMS and NAEE.....143

Table 6.7. The obtained source identification solutions of the developed MARS-based SMOs 146

# 1. Introduction

## 1.1. Overview

Groundwater is the main source of fresh water for all types of human needs. For example, 96% of the Earth's unfrozen freshwater is groundwater, about 70% of abstracted groundwater is consumed in the agricultural sector and almost half of the world's drinking water source is groundwater (NGWA, 2016). In Australia, the consumption of groundwater has increased over the past few decades. Figures 1.1 represents changes in annual groundwater consumption in Australia at three different times (Harrington & Cook, 2014).

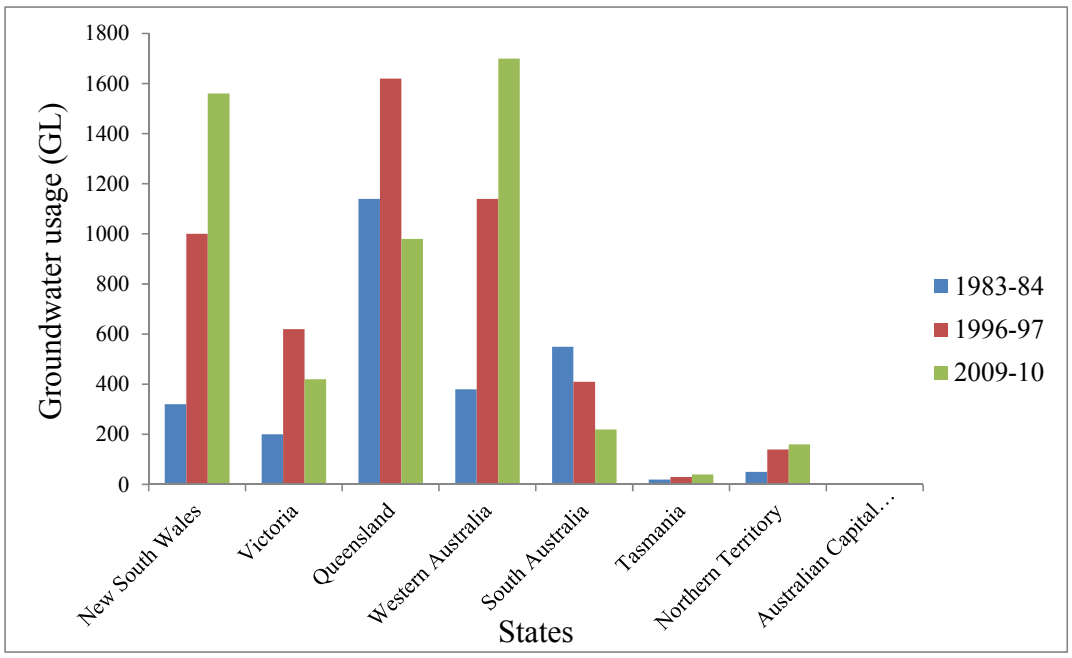



Figure 1.1. Annual groundwater consumption changes in Australia at three different times (Harrington & Cook, 2014)

As a result of possible climate change, the consumption of groundwater is expected to rise in Australia. However, the usage pattern of groundwater is different throughout the

country. For example, in some regions, groundwater is the only available source of fresh water. Figure 1.2 shows the pattern usage of groundwater in Australia.

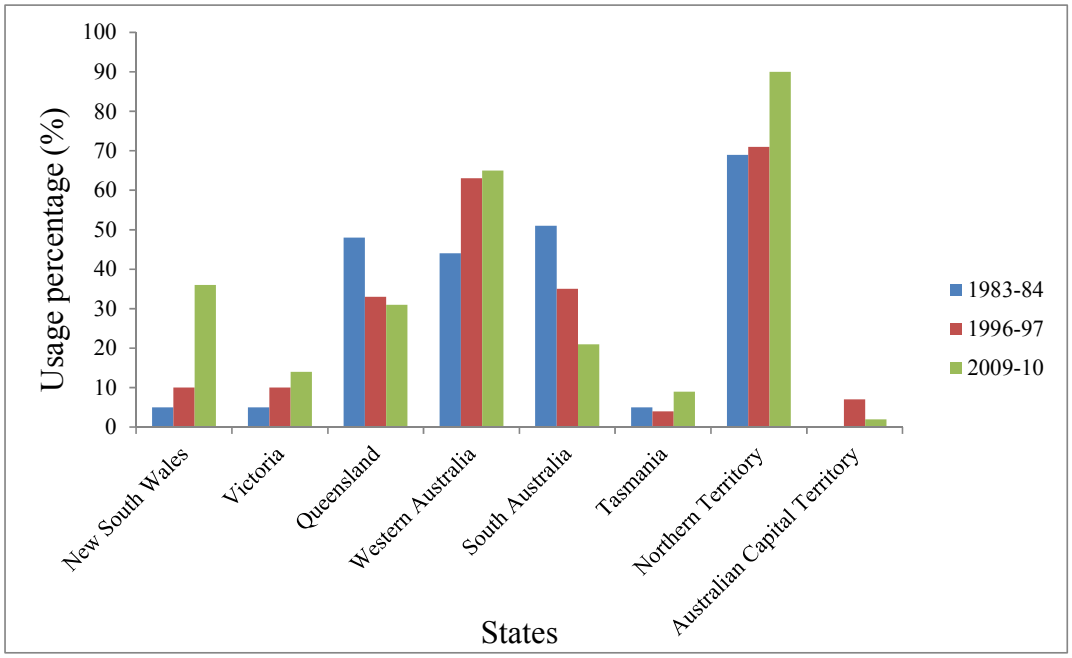



Figure 1.2. Groundwater consumption in different states of Australia (Harrington & Cook, 2014)

Widespread human activities and improper management practices have caused widespread deterioration of groundwater quality worldwide, and have seriously threatened its beneficial use in recent decades. On the other hand, groundwater pollution usually remains undetected for a long time. Therefore, enough data and information regarding the characteristics of groundwater contamination sources as well as the hydrogeologic parameters of the system are not available. As a result, identifying of unknown groundwater contaminant sources and remediation of contaminated aquifers are essential. Usually, identification of unknown groundwater contamination sources needs to be addressed in three main terms: contaminant source location(s); contaminant source fluxes; and contaminant release history. However, source identification usually is a complex problem that can be tedious and incorrect due to the uncertainties in the available

information and limitation of measurement data. The source identification solution results may also be non-unique due to the high sensitivity to the observation data and model parameters. So, developing an efficient methodology for source identification is a necessity in remediation processes and groundwater management.

In this study, three different algorithms were utilized to develop different surrogate models for source identification. Self-Organising Maps (SOM), Multivariate Adaptive Regression Splines (MARS), and Gaussian Process Regression (GPR) algorithms were utilized to develop surrogate models. Genetic Algorithm (GA) was applied as the optimisation algorithm to develop MARS-based surrogate model linked optimisation model. Each of the utilized algorithms for developing surrogate models in this study has special capabilities. For example, SOM is a powerful tool in classifying non-linear multidimensional data. On the other hand, in source identification problems, finding contaminant source location(s) is one of the main issues. Also because the source locations are very uncertain, generally several plausible locations are considered as potential contaminant source locations. The SOM-based Surrogate Model (SOM-based SM) can screen active sources among all potential contaminant sources. Therefore, in this research, the developed SOM-based SMs were applied to screen the actual source locations from an initially specified set of potential source locations. Especially with sparse measurement data and various uncertainties involved, accurately identifying actual source locations is an important and crucial step. The classification capabilities of SOM were utilized for source identification problem by classifying active and inactive source locations. This is essentially the screening of active source locations among all potential contaminant sources. As a result, the problem becomes simpler and easier to solve accurately. The preliminary source identification solutions of the SOM-based SMs not only could screen which of the potential contaminant sources are active but, also could

approximately estimate the contaminant source fluxes and the release times. These preliminary classification results can be utilized to update surrogate models by using sequential sampling methods.

In this research, MARS and GPR algorithms were also utilized for developing MARS and GPR-based Surrogate Models (MARS and GPR-based SM) for source identification, because of their capabilities in interpolating and exploring of unknown functions of multidimensional data. Comparison of the obtained source identification results of MARS and GPR-based SMs with the obtained results of the developed SOM-based SMs in terms of specified error criteria showed better accuracy for the developed MARS and GPR-based SMs' source identification results.

Application of some of the existing approaches to real-world contaminated aquifer sites generally requires enormous computational time. Typically, the computational time involved in solving the source identification problem may range between days or weeks of CPU time to obtain an optimal solution. Therefore, Surrogate Modelling-Based Optimisation (SMO) approach has been developed to increase computational efficiency. Surrogate models based on Artificial Neural Networks (ANN), Genetic Algorithm (GA), Kriging, and regression techniques have been developed as approximate simulators of the physical processes ((Bhattacharjya & Datta, 2005; Sreekanth & Datta, 2010), and (Razavi, Tolson, & Burn, 2012)). Surrogate models are trained by using numerical simulation models. Once trained, and tested the developed surrogate models can approximate the physical process simulation utilizing more rigorous numerical models. Therefore, the most commonly utilized approach for source identification, linked simulation-optimisation-models, which link with computationally intensive numerical simulation models, can be replaced by surrogate models linked to optimisation models (Singh & Datta, 2006). Replacing the numerical simulation models by surrogate models

can substantially result in computational efficiency and feasibility, as the linked simulation-optimisation models require a repeated solution of the numerical simulation models (Datta & Kourakos, 2015).

However, as mentioned earlier in this chapter, the powerful data-mining tools SOM, MARS, and GPR algorithms were utilized to develop surrogate models for source identification. The preliminary performance evaluation results of the developed SOM-based SMs showed that it is possible to apply SOM-based SMs to screen actual contaminant source locations among the specified potential contaminant source locations. The preliminary screening of the source locations can be utilized for updating the developed surrogate models. New surrogate models can be developed and updated by using the more powerful algorithm(s) in data mining i.e., using different surrogate model types (GPR, MARS), or by applying sequential sampling method. The combination of capabilities of these three algorithms (SOM, MARS and GPR) in developing surrogate models could overcome the main drawback of existing source identification problem solution, i.e., computationally intensive. Moreover, the developed surrogate models could directly apply for source identification without the necessity of using a computationally intensive linked simulation-optimisation model.

As mentioned earlier in this chapter, the accuracy and efficiency of source identification methodologies mostly are related to the quality and quantity of available data. Optimal designing of locations for monitoring contaminant concentrations could enhance the efficiency and accuracy of the source identification methodologies. Therefore, this study utilized different robust tools in data mining with capabilities in recognizing the most important and influential variables of the physical response prediction models. These tools in this study were utilized to identify the most important or relevant monitoring locations which could improve source identification results. The tools utilized to develop

a monitoring network design procedure were Random Forests (RF), Classification and Regression Trees (CART), and Tree Net (TN). In this approach, the monitoring locations which have the most influence on source identification were selected.

The performances of the developed methodologies were assessed by utilizing the developed methodologies to different contaminated study areas such as two hypothetical contaminated aquifer sites, an experimental contaminated aquifer site, and a contaminated aquifer site in Australia. In this study, in different cases, only limited contaminant concentration measurement values were assumed to be available. Also, in most of the cases, it was assumed that the contaminant concentration data were collected a long time after the start of the first potential contaminant source activities. The performance evaluation results of the developed surrogate models for source characterizing in different cases with limited concentration measurements data, parameter values, and under hydraulic conductivity uncertainties, were shown to be satisfactory in terms of source identification accuracy.

### 1.2. Objectives

The methodologies proposed by earlier researchers for source identification have various limitations. These limitations can be listed as:

1. Most of them are computationally extensive (Borah & Bhattacharjya, 2014);
2. Usually, the developed methodologies for source identification need to solve a linked optimisation model;
3. Usually, a difficult process is needed to find contaminant source location(s) (Prakash & Datta, 2015); and
4. Only a few of the proposed methodologies were evaluated under uncertain hydrogeological parameter conditions (Amirabdollahian & Datta, 2014).

Therefore, the SOM, MARS and GPR algorithms were utilized to develop surrogate models for source identification. The developed surrogate models were utilized for source identification with very limited information regarding contaminant source location(s), contaminant source magnitudes, and contaminant source activity times. It was supposed that as a result of replacing simulation models of groundwater flow and solute transport with the developed surrogate models approximating the physical processes, computational cost-effectively could be substantially reduced. The developed surrogate models could also be utilized for source identification directly in an inverse mode, without linking to an optimization model.

The developed surrogate models were also utilized for source identification with sparse and limited measurements data. The GA algorithm was also utilized to define an optimisation model in the developed MARS-based SMO for source identification. The capabilities of these different algorithms in constructing surrogate models make the complicated source identification problem easier to solve. Also, comparison of the implementation process of the developed surrogate models for source identification to the existing methodologies showed its ease of implementation. For example, SOM capabilities in classifying were utilized to screen non-active or dummy source locations among the potential contaminant source locations. Capabilities of the MARS and GPR algorithms in approximating the behaviour of non-linear multidimensional data were also utilized for source identification.

Properly designed locations for monitoring contaminant concentrations dedicated to increasing the accuracy and efficiency of the source identification process is very important, especially when the available measurement data are sparse and or, erroneous. A monitoring network design procedure was also developed in this study. The information from the designed monitoring network was utilized to develop new surrogate

models. The solution results of the developed surrogate models obtained by utilizing information from the designed monitoring network indicate the significant improvements in source identification results. Specific main objectives and the related steps of this study can be listed as follows:

1. Develop efficient methodologies for identification of unknown groundwater contaminant sources especially where data are sparse. For achieving this objective, SOM, MARS, and GPR tools were utilized to develop surrogate models for source identification. MARS-based SMOs were also developed for source identification for comparison purpose. The performances of the developed surrogate models were tested and compared in different study areas with limited contaminant concentration values. Information from two hypothetical study areas, an experimental site, and a real-world contaminated aquifer were utilized for evaluation of the performance of the developed methodologies.
2. Explore the possibility of independently using the SOM algorithm to identify unknown groundwater contaminant sources. To achieve this objective, the developed SOM-based SMs were utilized independently for source identification without linking to an optimisation model.
3. Simplify the solution process of source identification problems. This objective was achieved by using the SOM-based SMs for source identification. The solution results of the developed SOM-based SMs that can be considered as preliminary solutions to precisely screen the active contaminant sources among all potential contaminant sources. As a result, by screening the contaminant source locations, the number of unknown variables related to source characteristics that need to be addressed are decreased.

4. Evaluate the extension of the developed surrogate models to incorporate contaminant concentration measurement errors. To achieve this objective, the application of the developed surrogate models were assessed for source identification under different scenarios by using erroneous concentration measurements data.
5. Evaluate the performance of the developed methodologies by utilizing synthetic data (simulated data), and hydrogeologic data from different contaminant aquifers. As mentioned earlier, the developed methodologies were utilized for source identification in different study areas including two illustrative hypothetical study areas, as well as a contaminated experimental aquifer site, and a real contaminated aquifer site in Australia.
6. Develop an efficient and easily implementable procedure for designing a monitoring network to improve the performance of source identification process. For this purpose, the RF, CART, and TN data mining tools were utilized to design the monitoring network. Two objectives were considered in designing the monitoring network: 1. maximise the accuracy of source identification results, and 2. limit the numbers of monitoring locations.
7. Evaluate the performance of the developed monitoring network design procedure in conjunction with source identification methodologies by using information from a contaminated illustrative hypothetical study area.
8. The performance evaluation of a developed monitoring network design procedure in conjunction with the developed source identification methodologies by using the hydrogeologic and contaminant concentration data from a real contaminated aquifer site in Australia.

### 1.3. Organisation of the Thesis

This thesis consists of seven chapters including this introduction chapter. The other chapters are briefly discussed in the following paragraphs.

Chapter 2 presents a review of earlier developed methodologies (literature) for source identification. In this chapter, some of the advantages and disadvantages of the existing methodologies are explained. A review of literature is presented related to the monitoring network design procedures, surrogate models approaches and the SOM algorithm.

Chapter 3 presents the developed surrogate models for source identification by using the SOM, MARS, and GPR algorithms. Then, application of the developed methodologies to a hypothetical study area for source identification is discussed. Preliminary source identification results of the developed SOM-based SM were utilized to apply sequential sampling method to update the developed surrogate models or to develop adaptive surrogate models. Source identification was addressed in terms of contaminant source location(s), magnitudes, and release history. In this study area, contaminant concentration values were assumed to be missing for a period after the start of first contaminant source activities. The contaminant concentrations were also assumed to be available at a few observation locations. The performance of the developed methodologies was evaluated using error-free, as well as erroneous concentration measurement data.

Chapter 4 presents the application of the developed MARS-based SMO to an illustrative study area for contaminant source identification. The developed monitoring network design procedure is also explained. The performance of the developed MARS-based SMO was evaluated in a heterogeneous, multi-layered contaminated aquifer. In this study area, it was assumed that only limited concentration measurement values were available. Also, it was assumed that the contaminant concentration data were collected a long time after the start of first potential contaminant source(s) activities. The performance of the

developed MARS-based SMO was evaluated by using deterministic hydraulic conductivity values, and uncertain hydraulic conductivity values. For evaluating the applicability of the developed monitoring network procedure, obtained information from the designed monitoring network was also used for source identification. Then, these obtained solution results were compared with the preliminary solution results obtained by utilizing data from the arbitrary observation locations. The performance evaluation results indicated that by using data from the designed monitoring network, the accuracy of source identification results showed significant improvement in source identification results.

Chapter 5 briefly presents the developed surrogate models for source identification. Then, the developed methodologies were applied to an experimental contaminated aquifer site within a heterogeneous sand aquifer in Australia. The performance evaluation results of the different applied methodologies for source identification were compared. The measured contaminant concentration values and hydraulic conductivity values were not error free in this study area. For example, the distributions of hydraulic conductivity showed considerable variations in short distances. Inverse Distance Weighting (IDW) methodology was utilized to generate hydraulic conductivity values at locations where these values were unknown.

Chapter 6 presents the application of SOM-based SMs and MARS-based SMO to a contaminated aquifer site in New South Wales, Australia. The developed procedure for monitoring network design was also utilized to identify the monitoring locations that could make the most contributions to source identification. The performance evaluation results of the developed methodologies in conjunction with the designed monitoring network indicated the potential applicability of the developed methodologies in real-world cases, where sparse and limited contaminant concentrations data are available.

In Chapter 7, the developed methodologies and their performance evaluation results for source identification in different scenarios are briefly summarised. Also, the main conclusions of this study are presented in this Chapter, and some of the limitations of the developed procedures are discussed.

The next Chapter briefly reviews some of the important literature relevant to this study.

# 2. Literature Review

## 2.1. Introduction

This chapter provides a review of literature related to unknown groundwater contaminant source identification, monitoring network design, surrogate models, and Self-Organising Maps. The remediation of contaminated aquifers is one of the main challenges encountered in groundwater management. This challenge arises due to insufficient information regarding the contaminated aquifers. The availability of limited information and the uncertainties intrinsic in the numerical simulation of groundwater flow and solute transport make contaminant source identification a problem with uncertainties. Consequently, effective remediation process remains a difficult task. Therefore, developing an efficient methodology for source identification has a crucial role in the remediation process. The methodologies proposed earlier are generally highly sensitive to concentration measurement errors and need a very large amount of data and computation time. For example, the linked simulation-optimisation methodology which is the most frequently utilized methodology to tackle this problem is computationally intensive. This methodology requires an enormous amount of computational time when numerical simulation models are utilized in conjunction with optimisation models. As a result, Surrogate Models linked to Optimization (SMO) have been suggested to solve these computational problems using much smaller computation times. In this proposed method, a simpler and faster model approximates groundwater numerical flow and contaminant transport models. Moreover, developing a monitoring network design and using information obtained from the implemented monitoring network could improve source identification results and subsequently efficiency of the remediation processes significantly.

In this chapter, a briefly reviews earlier studies conducted on methodologies for source identification, monitoring network design, as well as the development of surrogate models relevant to source identification, and Self-Organising Maps.

## 2.2. Developed Methodologies for Source Identification

The source identification problem can be classified as a nonlinear problem (Mahar & Datta, 2000). This problem also can be classified as an ill-posed non-unique and inverse problem (Datta, 2002), as a result of erroneous measurements data and the necessity of using optimisation models (Amirabdollahian & Datta, 2013). Inverse problems are classified as well-posed if they have specific characteristics such as a unique solution and stability in the solution. In source identification problems, these solutions may be non-unique and unstable due to high sensitivity to the accuracy of recorded data and hydrogeologic parameters which required in the simulation models. The source identification results can also be inaccurate in terms of contaminant source location(s), magnitudes, and timing due to limited measured data and various uncertainties in the hydrogeologic parameters' values. The procedures have been proposed previously for source identification can be classified into two major groups:

1. Statistical or numerical estimation approaches: In this approach, related equations were solved backwards in time as an inverse problem. In these procedures, techniques that can overcome the non-uniqueness and instability of source identification solution results were applied (Pinder, Ross, & Dokou, 2009); and
2. Approaches based on optimisation algorithms in conjunction with simulation models of groundwater flow and solute transport.

An extended review of literature related to source identification procedures can be found in Prakash and Datta (2015), Amirabdollahian and Datta (2013), Jha and Datta (2013) , Chadalavada, Datta, and Naidu (2011a), J. Atmadja and A. Bagtzoglou (2001); Sun,

Painter, and Wittmeyer (2006a, 2006b), Bagtzoglou and Atmadja (2005), J. Atmadja and A. Bagtzoglou (2001) and Mahar and Datta (1997). In the next two sections (2.2.1 and 2.2.2), a review of some of the important procedures that have been applied for source identification is explained.

### 2.2.1. Statistical Approaches

The random walk particle method was applied by Bagtzoglou, Dougherty, and Tompson (1992); Bagtzoglou, Tompson, and Dougherty (1991) to solve unknown contamination source characteristics in backward time without using the optimisation concept. Their procedure was based on stochastic methods which increased the probability of identifying unknown pollution sources properties by applying measurement data from the designed monitoring locations. Although the proposed approach could handle heterogeneity of the aquifer, it involved extensive computation. An alternative methodology, an inverse analytical method, was proposed by Ala and Domenico (1992) to solve various equations simultaneously. This method could calculate the unique magnitudes of different unknown parameters of contaminant sources which affect the characteristics of pollution plumes. Also, an inverse model which combined simulation models of groundwater flow and solute transport with a non-linear maximum likelihood optimisation model was utilized by Wagner (1992). This method was utilized to calculate unknown simulation model parameters and the characteristics of pollution sources.

Tikhonov Regularization (TR) was utilized by Skaggas and Kabala (1994) to recover the release history of a contaminant plume. They assumed the system as a one-dimensional homogeneous system with single contaminant source. For achieving a unique solution, they utilized TR to transform the ill-posed inverse problem to a minimisation problem. The results indicated that TR was sensitive to measurement errors. For example, even small errors in measurement data or input data can produce large errors in the solution

results. The results also demonstrated that this approach might be an effective approach to recover release and evaluation history of pollution plumes when sufficient data are available.

The Minimum Relative Entropy (MRE) method was applied by Woodbury et al. (1996). They utilized prior information for recovering release histories of pollution plumes in a one-dimensional steady groundwater system as a linear inverse problem. In this approach, the pollution source is characterized as a Probability Density Function (PDF). This function enabled them to forecast the future behaviour of the plume. In addition, a geostatistical method in a Bayesian framework was applied to estimate the release history of a conservative groundwater contamination by Snodgrassa and Kitanidis (1997). They also utilized the TR procedure to reconstruct the plume characteristics in a one-dimensional, homogeneous system by transforming the ill-posed mass flux problem to a well-posed problem. Their results illustrated that the accuracy of their solution significantly related to plume measurement errors. They concluded that this approach could be an effective method in the presence of enough accurate measurement data. However, the developed approach was not examined for more complicated systems such as heterogeneous multidimensional systems.

The Monte Carlo approach was utilized by Skaggas and Kabala (1998) in a procedure to reconstruct plume history characteristics. In other research, an inverse technique based on correlation coefficient optimisation was developed by Sidauruk, Cheng, and Ouazar (1998). They applied this methodology to a two-dimensional example to delineate the groundwater contamination plume and its transport parameters. In this research, the analytical solutions were based on simplifying the problem by assuming uniform flow and a homogeneous aquifer in a two-dimensional example. The MRE strategy was also developed by Woodbury, Sudicky, Ulrych, and Ludwig (1998) to calculate the release

history of a three-dimensional contaminant plume in a steady and uniform groundwater system. Their results indicated that the monitoring of an earlier time of plume histories is important and useful because this method poorly reconstructed the release history of earlier duration.

The TR method, suggested by Skaggas and Kabala (1994), was also utilized by Liu and Ball (1999). They utilized measured concentration data in a low permeability field at Dover Air Force Base. The contaminant boundary concentration was estimated by assuming a simple mass-conserving two-layer diffusion model. They also utilized a least squares method in addition to a regularisation term of the objective functions. Their results indicated that inverse problems are inherently ill-posed and converting them to well-posed problems could affect the analysis.

The relative effectiveness of TR and MRE procedures again were assessed by Neupauer, Borchers, and Wilson (2000). They applied this procedure to reproduce the release history of a conservative pollution in a one-dimensional field. Their results indicated that with error-free data both methods were useful in constructing source release history functions while MRE results were better. However, TR achieved a more appropriate result with data that contain measurement errors. Moreover, MRE could identify the region of pollution with limited data regarding release history while the TR technique was not able to achieve this. The Backward Beam Equation (BBE) technique was applied by J. Atmadja and A. C. Bagtzoglou (2001) to identify the contaminant source location and recover release history of the pollution in a heterogeneous system. They also developed a hybrid method called the Marching-Jury Backward Beam Equation (MJBBE) which effectively increased the accuracy of previous solutions and enabled them to solve the actual problem.

Multiple mathematical tools were utilized by Dokou and Pinder (2009) to identify dense non-aqueous phase liquids (DNAPLs) source location(s) of a synthetic example. These tools were a Monte Carlo stochastic model of groundwater flow and solute transport, Kalman filter, Choquet integral and Latin Hyper Cube Sampling (LHS). They also assumed that there was enough hydrogeologic information to model the system. While the hydraulic conductivity values were under uncertainty, their results indicated that the developed algorithm was able to find the DNAPL source(s) by applying the existing water quality information. Also, their developed algorithm was able to distinguish the best-water quality sampling points among possible locations.

The developed methodologies in this group were applied mostly to characterize one or two-dimensional homogeneous contaminated study areas. In most of them, contaminant sources were considered as a single contaminant source. The validation results of the applied procedures demonstrated the potential applicability of them in the presence of sufficient and accurate measurements data.

### 2.2.2. Approaches based on Optimisation Algorithms

In the approaches based on optimisation models, consisting of the embedding technique, response matrix, linked simulation-optimisation approaches, and surrogate models linked to optimisation models were utilized to incorporate simulation models with optimisation models for source identification ((Mahar & Datta, 2000), (Amirabdollahian & Datta, 2013) , (Borah & Bhattacharjya, 2014), and (Prakash & Datta, 2015)).

#### *2.2.2.1. Response Matrix*

One of the techniques for identifying unknown groundwater contaminant sources was suggested by Gorelick, Evans, and Remson (1983). They utilized a response matrix approach for source identification. They utilized least square regression, a linear programming technique and stepwise multiple regressions techniques. Each of these

techniques was combined with a groundwater solute transport numerical simulation model to identify contaminant sources. The developed approach was applied to two hypothetical groundwater systems. One of these systems was steady-state and the other was a transient system. They assumed that all the contaminated aquifers' parameters are known without any uncertainty.

A combination of statistical pattern recognition and optimisation tools was utilized to identify pollution sources by Datta, Beegle, Kavvas, and Orlob (1989). An expert-system approach based on a statistical pattern recognition algorithm for source identification was developed by them. The expert system utilized the solution results obtained by using the statistical pattern recognition to choose a set for contaminant source location(s) and magnitude(s). They utilized a response matrix technique to simulate groundwater flow and solute transport processes (Mahar & Datta, 2001). The main disadvantages of the response matrix approaches can be:

1. These approaches need relatively high information about the aquifer system. For example, for developing the response matrix, aquifers' parameters should be known (Mahar & Datta, 2001);
2. In the developed methodologies based on a response matrix, groundwater systems were assumed to be linear (Singh, Datta, & Jain, 2004), and
3. The approach is also highly sensitive to measurement errors (Amirabdollahian & Datta, 2013).

#### *2.2.2.2. Embedded optimisation techniques*

This approach is a type of optimisation-based methodology which was utilized for source identification. The governing equations of groundwater flow and solute transport were embedded as constraints to an optimisation model (Mahar & Datta, 2001). Mahar and Datta (1997, 2000, 2001) had presented embedded optimisation techniques. They utilized

a nonlinear optimisation model incorporating finite difference discretised governing equations of groundwater flow and solute transport to identify unknown groundwater contaminant sources. The Performance evaluations of their procedure by applying information from an illustrative study area indicated the potential applicability of this approach when the aquifer's parameters were known, and measurements data were error-free. The embedded techniques have some limitations. For instance, for obtaining the optimal solutions, repeated solutions of a set of discretised groundwater and transport governing equations are required. As a result, these procedures are computationally intensive and especially for large-scale areas, although this approach may deliver most accurate solutions. The linked simulation optimization approach and then the replacement of numerical models with trained surrogate models were developed to overcome some of these issues with computational feasibility for large-scale real-world study areas ((Singh & Datta, 2006, 2007; Singh et al., 2004), (Jha & Datta, 2013), and (Prakash & Datta, 2015)).

#### *2.2.2.3. Linked simulation-optimisation approaches*

Among the proposed methodologies noted earlier, the linked simulation-optimisation approach is widely utilized for source identification due to its efficiency in source identification problems. The linked simulation-optimisation methodology is externally linked to the simulation models of groundwater flow and solute transport with an optimisation model. Some of the prominent algorithms which were utilized in this procedure are discussed in the following paragraphs.

A Progressive Genetic Algorithm (PGA) was applied by Aral, Guan, and Maslia (2001) to solve the nonlinear optimisation model in a source identification problem. The pollution source location and its activity time were assumed to be unknown variables. Compared with the embedded techniques, using PGA decreased the required numbers for

the solution of the simulation models of groundwater flow and solute transport. A hybrid approach based on Genetic Algorithm (GA) as a global search method and a local searching method was applied by Mahinthakumar and Sayeed (2005). They developed this approach by considering that the GA algorithm may not be very efficient in finding solutions near the global solutions. They developed their hybrid optimisation method for source identification as an inverse problem. The developed approach was utilized in two- and three-dimensional heterogeneous contaminated aquifer systems with a single pollution source. Later, Mahinthakumar and Sayeed (2006) developed their previous hybrid optimisation method to reconstruct groundwater unknown plume release histories in a three-dimensional heterogeneous aquifer system with single and multiple pollution sources. Their performance evaluation results demonstrated that this strategy can be an effective technique for a source identification problem.

A linked simulation-optimisation method based on the GA algorithm was developed by Singh and Datta (2006) to characterize unknown groundwater pollution sources. They applied the developed methodology to a complex study area with several contamination sources. Different scenarios representing different data availability conditions and concentration measurement errors were also considered. Their results demonstrated the importance of the numbers and locations of observation bores in source identification problems. The main advantage of the developed procedure was its potential applicability to complex contaminant aquifer systems with multiple contaminant sources. In another research project, a hybrid approach based on Simulated Annealing (SA), TABU Search (TS) and a three-dimensional solute transport model, was applied by Yeh, Chang, and Lin (2007). They utilized this method to solve and reconstruct the unknown pollution sources' locations and release history. First, they applied TS to select potential source locations in the suspected area. Then, they utilized SA to generate a release history and

characteristics. The developed methodology was utilized in different homogeneous and heterogeneous study areas. In only one scenario tested, the flow was considered to be transient.

A simulation-optimisation approach was utilized by He, Huang, and Lu (2009) to plan the remediation of a petroleum-polluted site under uncertainties in soil porosity. Their results indicated that this approach decreased the optimisation process cost, and possessed some characteristics such as 1. It addressed the stochastics' parameters of the numerical simulation models of flow and solute transport; 2. It connected a direct and rapid relationship between remediation procedures (pumping rate) and remediation efficiency; and 3. It had a confidence level for various optimal identification solutions.

A linked simulation-optimisation approach was developed by Ayvaz (2010) for source identification in terms of contaminant source locations and release history. In this approach, simulation models of groundwater flow and solute transport, MODFLOW and MT3DMS respectively, were integrated with a hybrid optimisation model. The optimisation model consisted of a binary genetic algorithm and the generalised reduced gradient optimisation technique. Information from a hypothetical study area with different contaminant source distributions was used to assess the performance of the developed approach.

The classical nonlinear optimisation algorithm externally linked with the simulation models of flow and solute transport by Datta, Chakrabarty, and Dhar (2011). They applied this methodology for source identification in a homogenous, isotropic illustrative study area. Multiple unknown contamination sources in this one-layer confined aquifer were considered. They concluded that this approach was applicable to large study areas with multiple unknown contaminant sources. The developed approach computationally was more efficient compared with the embedded techniques in source identification

problems. The linked simulation-optimisation approach, based on SA and Adaptive Simulated Annealing (ASA) algorithms, was applied by Jha and Datta (2012, 2013). They combined these algorithms as optimisation models with the simulation models of groundwater flow and solute transport for source identification in illustrated study areas. The results demonstrated that the SA and ASA-based linked simulation-optimisation approaches can be computationally more efficient and more accurate compared with the GA-based linked simulation-optimisation procedures.

A linked simulation-optimisation procedure by using SA was developed by Prakash and Datta (2014). They applied the developed model to characterize unknown contaminant sources when the starting activity times were not known. They applied the developed methodology to a hypothetical site with multiple contaminant sources. Their results indicated the potential applicability of the developed procedure for identifying contaminant release history and initiation times for an illustrative study area. Later, they utilized the developed methodology for source identification in a contaminated real aquifer site (Prakash & Datta, 2015). To enhance the accuracy of source identification results, they integrated the source identification procedure with a sequential monitoring network design procedure.

A direct search algorithm was utilized to solve an optimisation model linked to simulation models of groundwater flow and transport for source identification by Borah and Bhattacharjya (2014). The developed methodology was utilized in an illustrative study area and a contaminated homogeneous and isotropic confined aquifer. The performance assessment results demonstrated that the developed procedure was not computationally efficient as it took several days to solve a relatively easy real-world case. An alternative approach for source identification was developed and applied in an illustrative study area by Amirabdollahian and Datta (2014). They applied the ASA in conjunction with fuzzy

logic in the linked simulation-optimisation approach for source identification under parameter uncertainty. Their results indicated that this procedure was effective in estimating unknown pollution sources' characteristics with uncertainties in hydrologic parameters. Later, they evaluated the developed methodology by using it in different study areas such as an experimental site and a contaminated real-world aquifer site (Amir Abdollahian, 2016). The performance assessment results demonstrated the potential applicability of the developed approach for source identification.

The main advantages of the linked simulation-optimisation approach compared with the other ones include:

1. In this approach, some complex simulation models of groundwater flow and transport such as MODFLOW and MT3DMS can be utilized. This issue is important as the efficiency and accuracy of the source identification results are highly dependent on the performance of simulation models of groundwater flow and solute transport; and
2. The number of decision variables of the optimisation model can be decreased in this approach by eliminating the embedded equations as binding constraints (Datta, 2002), so the solutions can be easier and less intensive in terms of feasibility.

However, the main disadvantage of the developed linked simulation-optimisation approaches is their computational times which are very high (Borah & Bhattacharjya, 2014). For example, for solving a real-world case, they may need several days of the iterative solution. To overcome this drawback, recently computational simulation models of groundwater flow and transport have been replaced by surrogate models. In the next section, surrogate models and their application in engineering fields included in source identification problems are briefly discussed.

#### *2.2.2.4. Surrogate Models*

In most engineering fields, computer simulations are commonly utilized to simulate complex processes by using mathematical formulations. Generally, implementation of simulation models for real-world cases is complex and extensively time-consuming. To reach an optimal solution using a linked simulation-optimisation model, typically simulation models need to run thousands of times. Therefore, the solutions of these problems involve significant time and cost. To reduce these computing costs, these computationally intensive simulation models have been replaced by surrogate models or by response surface methodologies (Koziel, Ciaurri, & Leifsson, 2011; Razavi et al., 2012). The main reason for using and applying surrogate models is to make more efficient use of the available limited computational cost in the desired fields (Razavi et al., 2012). Some of the most important research that has been done in surrogate model fields are briefly explained in the following paragraphs.

A classification of surrogate models using global optimisation methods was investigated by Jones (2001). To better illustrate different surrogate model techniques, they utilized seven different techniques for different numerical examples. The utilized techniques were 1. Minimising a quadratic surface, 2. Minimising an interpolating surface, 3. Minimising a statistical lower bound, 4. Maximising the probability of improvement, 5. Maximising expected improvement, 6. One-stage approach for goal seeking, and 7. One-stage approach for optimisation. By comparing the obtained results, they found that the first two techniques were the simplest approaches. These two techniques easily could miss the global minimum. They also found that methods three to five were highly dependent on the quality of initial sampling. However, the results indicated that method four was the most practical and reliable among these seven. The last two techniques, methods six and seven, were classified as computationally extensive approaches if Kriging was utilized as

the surrogate model type. They also concluded that defining constraints for improving results in models six and seven can be useful in non-Kriging surfaces. Defining constraints in Kriging cases can increase their complexity and decrease their efficiency. Moreover, basic issues in constructing surrogate-based optimisation models such as the design of experimental data, surrogate model selection and analysis, optimisation methods evaluations, and surrogate model validation were explored by Queipo et al. (2005). Basic principles of constructing surrogate model-based optimisation were also discussed in Koziel et al. (2011).

However, different types of surrogate models such as surrogate model-based optimisation and Adaptive Surrogate Models (ASM) have been suggested to increase surrogate models' efficiency (Wang et al., 2014). For example, different methods for developing surrogate models such as Quadratic polynomial regression, Tree, Multivariate Adaptive Regression Spline (MARS), Gaussian Process Regression (GPR), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) were utilized by Wang et al. (2014). They utilized these techniques to develop an adaptive surrogate model-based optimisation to solve two benchmark problems: 1. The Hartman function, and 2. Calibration of the SAC-SMA hydrologic model. Their results demonstrated that the GPR algorithm was the best surrogate model type. The results also indicated that the minimum interpolation surface technique was the best adaptive sampling method. The Low Discrepancy Quasi-Monte Carlo technique was also mentioned as the most suitable experimental data design method in this study. They noted that the best sample size might be 15-20 times the dimension of the problem. An adaptive surrogate model-based sampling strategy was also developed by Gong and Duan (2017) for parameter optimisation and distribution.

An extensive literature review (48 studies) related to surrogate models that had been applied to water resources is contained in Razavi et al. (2012). The 48 series of research projects were analysed and categorised by them. They noted some important properties of this modelling such as:

1. Introducing any special type for surrogate models as the best one may not be easy, and it depends on the problem and software availability;
2. Surrogate models decrease time-consuming computational simulation models; and
3. By increasing the numbers of variables, the efficiency of the surrogate models could decrease.

Recently, different surrogate models were developed for source identification. Different algorithms such as ANN and Genetic Programming (GP) were utilized to develop surrogate-based optimisation models for source identification. For example, ANN was utilized as the surrogate model type for source identification by Singh et al. (2004). This procedure was evaluated by using various data availability and measurement errors in various locations and time steps. They applied this methodology to identify unknown groundwater contaminant source(s) problems in a simple case with a single source of contaminant and with multiple contaminant sources. Later, they applied the developed methodology for source identification (Singh & Datta, 2007) when concentration data were missing over a period of contaminant source(s) activity times. The available measurements data were also considered to be erroneous. They suggested this method was an acceptable practical method in source identification problems. The obtained results of Sreekanth and Datta (2010) demonstrated that a GP-based surrogate model showed better performance compared to a Modular Neural Network (MNN) based surrogate model in the management of saltwater intrusion problems. Recently, a GP-

based surrogate model linked to an optimisation model was utilized by K. Esfahani and Datta (2016) for reactive contaminant source identifications. They applied the developed methodology to a contaminated mining site. Their results demonstrated the potential applicability of GP in approximating groundwater flow and chemically reactive multiple species transport process in a contaminated aquifer site.

In some of the suggested approaches for source identification problems, an optimal monitoring network design approach was integrated with source identification methodologies to enhance the accuracy of source identification results. In the next section (2.3), a review of monitoring network design procedures is briefly presented. Some of the developed groundwater monitoring network design methodologies integrated with source identification procedures are discussed in section 2.4.

### 2.3. Monitoring Network Design Procedures

In source identification problems, the quality and quantity of contaminant concentrations data play an essential role in the accuracy of solution results. The complexity of source identification problems also arises due to insufficient, sparse and uncertain hydrogeological data. On the other hand, sparse and limited data are usually available because of the huge cost of long-term monitoring worldwide. Monitoring and collecting data in a contaminated porous medium is an expensive and time-consuming procedure. So, designing a monitoring network might be one of the essential steps of source identification problems and subsequently remediation processes. As a result, designing a monitoring network could improve groundwater management worldwide. In the following paragraphs, examples of related research are explained.

A Dynamic Programming (DP) algorithm was applied by Cleveland and Yeh (1991) for designing a monitoring network and schedule for estimating aquifer characteristics. They utilized this procedure to estimate data of an aquifer's model parameters. In this work,

their criterion was to obtain the maximum information from the study area within a specified budget. Their results demonstrated that sampling before contaminant concentrations reached a sufficient level may not be cost-effective. They applied the developed methodology in a small-scale case. However, they found that the developed methodology can be applied to larger and more real areas by applying modifications. In other research, a review of the most outstanding methodologies which had been applied to design the groundwater quality monitoring networks was published by Loaiciga et al. (1992). They found that the dynamic nature and institutional programs were the most ideal for optimal monitoring network design in most research. Also, in most works, multi-objective functions that included cost and health criteria were utilized. A mathematical model for groundwater quality monitoring network was developed by Datta and Dhiman (1996). They utilized mixed integer programming procedure to design an optimal monitoring network by minimizing undetected contaminant concentrations. Eight nonlinear multi-objective optimisation methods were applied by Lee and Ellis (1996) to design monitoring networks. They suggested that SA and TABU search methods were superior to other methods in designing a monitoring network.

Also, a multi-objective optimal monitoring network design was developed (Reed & Minsker, 2004) for long-term groundwater monitoring. In this developed methodology, quantile Kriging and Non-Dominated Sorted Genetic Algorithm-II (NSGA-II) was utilized to solve this multi-objective problem. The aims of this study were: 1. Minimise sampling cost, 2. Maximise the accuracy and quality of interpolated plume maps, 3. Maximise the accuracy of estimated contaminant concentrations, and 4. Minimise estimation uncertainty. Seven important steps in Long-Term Monitoring Optimisation (LTMO) was suggested by EPA. (2005). These steps included properly defining and documenting the existing monitoring procedure, analysing available data, examining the

potential use of the site for LTMO, selecting the LTMOA technique, applying the selected LTMO technique and performance evaluation of the implemented methodology. In other research, a cost-effective long-term monitoring network designed by preserving accuracy for a sampling of contaminant aquifer was implemented by Wu, Zheng, and Chien (2005). They utilized Ordinary Kriging (OK) and Inverse Distance Weighting (IDW) for interpolating plume. According to their results, they recommended the OK procedure for future works even though this method was more time-consuming. GA and SA were utilized by Mugunthan and Shoemaker (2010) to design a cost-effective sampling monitoring network. This methodology was utilized for long-term monitoring over multiple monitoring periods under uncertain flow conditions. They utilized and compared two different methodologies to solve optimisation models: 1. myopic heuristic algorithm with an error-reducing search neighbourhood and 2. SA as the error-reducing neighbourhood and GA. The first approach performs considerably better than the second one. This strategy could save 25% in project costs by using all possible locations and samples.

A multi-objective optimal long-term groundwater monitoring network design under hydraulic conductivity uncertainty was developed by Luo, Wu, Yang, Qian, and Wu (2016). The main aims of this study were to minimise total sampling costs for monitoring contaminant plume, mass estimation error, the first-moment estimation error, and the second-moment estimation error of the contaminant plume. A Probabilistic Pareto Genetic Algorithm (PPGA) combined with simulation models of groundwater flow and transport were utilized. The developed procedure was evaluated by using Monte Carlo analysis. The developed procedure performance was applied to a two-dimensional hypothetical study area and a three-dimensional real case. The performance assessment

results indicated the potential usage of the suggested approach for optimal designing for a long-term monitoring network.

## 2.4. Source Identification Procedures in Conjunction with Monitoring Network Design Procedures

Optimal monitoring network design procedures in conjunction with source identification methodologies could improve the accuracy of source identification results significantly ((Amirabdollahian & Datta, 2013) and (Prakash & Datta, 2015)). In this chapter, some of the earlier works in this field are briefly discussed in the following paragraphs.

A three-step approach combining optimal source identification procedure with a designed optimal monitoring network was introduced by Mahar and Datta (1997). In the first step, an embedded nonlinear optimisation model by using existing information of contaminant concentrations was used for preliminary source identification. In the next step, these preliminary results were used to design an optimal monitoring network. In the final step, recorded contaminant concentration data from the designed monitoring network were also used for source identification. Comparison of the results showed a significant improvement in the accuracy of source identification results when information from designed monitoring network was utilized. However, parameter uncertainty was not sufficiently considered in this study.

A dynamic monitoring network procedure that conforms with the transient nature of solute transport was designed by Dhar and Datta (2007). They considered two main objectives: 1. the cost of installing monitoring wells and monitoring contaminant concentrations at these locations, and 2. minimising estimated variances of contaminant concentrations at unmonitored locations. The obtained results indicated that their method was applicable for designing an economically efficient groundwater monitoring network. An optimal groundwater monitoring network design procedure was also developed by

Chadalavada and Datta (2008) in a transient flow to detect the pollution transport process in a hypothetical groundwater system. They applied GA as the optimisation algorithm to solve the optimisation model with these two objectives: 1. minimise the sum of unmonitored pollution values at various potential monitoring places, and 2. minimise estimated variances of pollution concentration values at locations without monitoring wells.

Also, an unknown source identification strategy was combined with an optimal monitoring network procedure by Datta, Chakrabarty, and Dhar (2009). In this method, the limited pollution concentration values were utilized to estimate pollution source magnitudes. In the next step, the results of a previous stage were applied to design a monitoring network. Then, collected data from the designed monitoring wells were utilized for source identification. The results demonstrated that the identifying unknown contaminant sources process was improved. This procedure was continued until it reached the desired accuracy. They evaluated their methodology in areas with known contaminant sources and the results were satisfactory.

Moreover, an optimal search strategy that identified non-aqueous phase liquids was developed by Dokou and Pinder (2009). The developed search study consisted of a Monte Carlo stochastic groundwater flow and transport model, an existing set of potential contaminant source locations and a Kalman filter. The Kalman filter was utilized to update simulated contaminant concentrations by using contaminant concentration data. In this methodology, they combined simulation models with expert knowledge to develop an integrated optimal procedure for identifying a DNAPL source location. The developed methodology utilized synthetic data in different scenarios which could represent real field conditions.

A multi-objective monitoring network design procedure combined with source identification procedure was developed by Bashi-Azghadi, Kerachian, Bazargan-Lari, and Solouki (2010). NSGA-II linked to the simulation models of groundwater flow and transport were utilized. The objective functions of the optimisation model were considered to minimise the total number of monitoring locations and maximise the accuracy and reliability of unknown groundwater contaminant source identification result. Another strategy was developed to design an optimal monitoring network by Dhar and Datta (2010). They considered redundancy reduction that results in economic inefficiency in the network. They utilized the branch-and-bound algorithm to solve the linear optimisation model. This methodology was also tested in a real case study area and the evaluation results were satisfactory. An optimal monitoring network based on uncertainty in estimating concentration values was applied by Chadalavada, Datta, and Naidu (2011b). They considered two criteria in designing the monitoring networks: 1. Minimise the spatial concentration estimation values variances at monitoring location; and 2. The number of monitoring wells. They evaluated this procedure in terms of concentration estimation errors. Later, they utilized a feedback base methodology for source identification (Chadalavada, Datta, & Naidu, 2012). The developed methodology was a sequential optimal monitoring network design and source identification procedure that applied in a hypothetical contaminated area and in a real contaminated aquifer site. Their results demonstrated that a feedback-based strategy can be useful in source identification.

An optimal monitoring network design methodology by using GP was developed by Prakash and Datta (2013). In this methodology, the GP algorithm and the linked simulation-optimisation procedure were utilized to reconstruct the plume history of unknown contaminant source(s) by using limited contaminant concentrations data. They

considered the maximum number of monitoring locations to minimise the probability of missing contamination sources. They evaluated this methodology by applying it to an illustrative study area. The solution results demonstrated that this procedure improved the efficiency of identifying unknown pollution sources by using the designed monitoring network data. The integrated linked simulation-optimisation approach with optimal monitoring design procedure was also utilized by Datta, Prakash, Campbell, and Escalada (2013) to improve source identification results. GP and SA algorithms were utilized to solve optimisation models in the optimal monitoring network design and the linked simulation-optimisation approaches. Limited performance evaluations of the developed methodology indicate its capability in the source identification.

A feedback-based methodology, integrated of a sequential monitoring network design procedure with the linked simulation-optimisation approach for improving source identification results, was developed by Prakash (2014); Prakash and Datta (2015). The SA algorithm was utilized as optimisation models in the linked simulation-optimisation procedure and optimal designed monitoring network procedure. The developed methodology was evaluated by using a real contaminated aquifer site in Australia.

A two-objective monitoring network design procedure integrated with a linked simulation-optimisation approach was developed by Amir Abdollahian (2016). The main objectives of this methodology were: 1. Reduce the uncertainty of reconstructed plume history; and 2. Reduce the redundancy of observation locations. The NSGA-II algorithm was utilized to develop an optimisation model in this methodology. The developed procedure was applied in a contaminated aquifer site in Australia.

### 2.5. Self-Organising Maps

The standard Self-Organising training method is a type of Neural Network (NN). This modified NN can be classified as an unsupervised technique because it does not need to

have a specific target output. This algorithm was introduced by T. Kohonen in 1982 (Kohonen, Oja, Simula, Visa, & Kangas, 1996). This algorithm is widely utilized to visualise and cluster multidimensional data due to its efficiency and easy implementation. The main features of this algorithm are its capability of transforming complex non-linear high-dimensional data space into simple geometric relationships. Usually, the relationships are presented in two dimensions by preserving the topological structure of the input data (Di Mauro, Maggioni, Grasso, & Colosimo, 2016; Kohonen et al., 1996). In other words, the most important characteristics of SOM are visualisation, classification, and abstraction of raw data for high-dimensional systems (Kohonen et al., 1996). The SOM algorithm can be utilized for different purposes such as decreasing total numbers of training data, accelerating the learning process, nonlinear interpolation, generalisation, and reliable abstraction of information for transmission (Kohonen et al., 1996).

In the past two decades, SOM has been applied in different fields to classify and visualise multidimensional data. A comprehensive review of SOM and its potential applicability was reported by Thi et al. (2014). This research illustrated that because of the inherent characteristics of the SOM algorithm, dimensionality reduction, and data compression, SOM was broadly utilized in data mining and machine learning. In the following paragraphs, some of the existing studies are explained briefly.

SOM was applied to predict a non-linear time sequence data of kinetic trajectories in a set of potentials by Walter, Ritter, and Schulten (1990). Also, the ability of SOM in the modelling of complex systems and the application of this potential in predicting future states were shown by Simula, Vesanto, Alhoniemi, and Hollmen (1999). Their results indicated that by applying SOM to solve a complex system, there was no need to define the problem with an analytical function. In another research, SOM was utilized in

ecological modelling to create linear regression by Whigham (2005). In addition, a hybrid strategy that combined SOM with NSGA II was applied to solve a multi-objective optimisation water distribution problem by Norouzi and Rakhshandehroo (2011). Their results indicated that application of SOM in this study increased the efficiency of standard NSGA-II in the same case. Thi et al. (2014) also utilized SOM as an optimisation algorithm. SOM capability in clustering was utilized by Dragomir, Dragomir, and Radulescu (2014) to classify consumers' daily load profiles. Vatanen et al. (2015) utilized SOM and Generative Topographic Mapping in the presence of missing data to analyse their learning results for high-dimensional data. SOM algorithm was also utilized by Barbariol et al. (2016) to characterize the extremes of a sea wave.

### 2.6. Motivation for this Study

All the developed source identification procedures had addressed at least one of the three main questions related to contaminant source characteristics. These three questions address the contaminant source location(s), contaminant source fluxes, and contaminant release histories. Among the existing methodologies for unknown contaminant source identification, the linked simulation-optimisation method in conjunction with optimal monitoring network design is more efficient compared with other techniques. This procedure is more practical and efficient especially for large-scale and complex aquifer systems (Prakash & Datta, 2015). Many of the previous approaches are very sensitive to concentration measurement errors and are only applicable to simple sites such as Neupauer and Wilson (1999); Skaggas and Kabala (1994); Snodgrassa and Kitanidis (1997).

Moreover, significant numbers of the earlier developed methodologies for achieving a reliable source identification solution need considerable input data including contaminant concentrations data (Prakash, 2014). Also, a significant number of previously proposed

approaches considered that all the hydrogeological parameter values are known. Only a few previously developed methodologies such as Amirabdollahian and Datta (2014) and Dokou and Pinder (2009) were evaluated under uncertain hydrogeological parameter conditions. Among the earlier developed methodologies, a few such as Ala and Domenico (1992), Liu and Ball (1999), and Prakash and Datta (2015) applied their developed methodologies to real-world cases.

On the other hand, application of the linked simulation-optimisation approach for source identification to real-world contaminated aquifers can be very computationally intensive due to repeated runs of the simulation models within the optimisation algorithms. In some real-world cases, it may need several days for source identification. Therefore, three different algorithms, SOM, MARS, and GPR, with different capabilities for comparison purpose were utilized to develop surrogate models for source identification. In other words, the linked simulation-optimisation model was replaced by trained surrogate models for source identification. It was supposed that by replacing the simulation models of groundwater flow and transport by approximate surrogate models, computational time of solving source identification problems would be decreased. The constructed surrogate models approximate the simulation models of groundwater flow and solute transport accurately. These surrogate models are also able to eliminate the need for using a formal optimisation model for source identification in terms of location, magnitude, and release history. It was supposed that by decreasing the required numbers of simulation and optimisation models runs, the computational cost would be decreased significantly. Surrogate models based optimisation was also developed for comparison purpose.

Moreover, one of the challenges in source identification problems is the uncertainty related to the numbers and locations of contaminant sources (Prakash & Datta, 2015). Only in the cases that the numbers of contaminant source locations are estimated to be

known with some degrees of certainty, the source identification results in terms of contaminant source locations, contaminant source fluxes, and contaminant release histories can be meaningful. The developed SOM-based surrogate models due to the capabilities of the SOM algorithm in clustering could screen the active contaminant sources among all potential contaminant sources. Identifying the contaminant source locations may simplify the source identification problem. As a result, more algorithms and methodologies can be utilized to obtain more reliable solutions for source identification.

The performance of the developed surrogate models in this study was assessed by using erroneous contaminant concentration data, limited measured contaminant concentrations data, an experimental contaminated aquifer site data and a real-world contaminated aquifer site in Australia. The obtained results indicate the potential applicability of the developed procedures for source identification.

Another challenge of the source identification problem is the availability of contaminant concentrations data. Usually, these data are sparse in the contaminated aquifer sites. The existing methodologies usually need significant numbers of contaminant concentration data at different monitoring locations which can cause significant cost worldwide. Therefore, designing a monitoring network can significantly improve source identification process. In this research, three different algorithms were utilized to identify monitoring wells that have the most contributions in source identification. Random Forests (RF), Tree Net (TN) and CART were the three algorithms utilized for designing a monitoring network in this study. The performance of the developed monitoring network procedure was evaluated in an illustrative hypothetical and a real-world contaminated aquifer. The results indicated the potential applicability of these algorithms

in designing monitoring network. The information obtained from the designed monitoring networks could improve source identification results.

In the next chapter, first, the developed methodologies for source identification are explained in detail. Then, the performance evaluation results of the developed procedures for an illustrative study area is presented.

# 3. Contaminant Source Identification by Utilizing Adaptive Surrogate Models

## 3.1. Introduction

Some contents of this chapter have been released in the following Journal paper:

- Hazrati-Yadkoori, S., & Datta, B. (2017). Adaptive Surrogate Model Based Optimization (ASMBO) for Unknown Groundwater Contaminant Source Identification Using Self-Organizing Maps. Journal of Water Resource and Protection, 9, 23. doi:10.4236/jwarp.2017.92014

Identifying unknown groundwater contaminant sources in terms of contaminant source location(s) magnitudes and release history is a complex problem. The methodologies proposed earlier to solve this complex problem are usually computationally intensive. For example, the most effective approach to tackle source identification problems is the linked simulation-optimisation approach. However, application of this approach to real-world cases may require days or weeks of CPU time to obtain an optimal solution when simulation models are linked to the optimisation algorithm. Therefore, Surrogate Modelling Based Optimisation (SMO) as an alternative has been developed to decrease these computational costs and time associated with repeated runs of the numerical simulation models within the optimisation algorithm.

One of the main goals of this study was to introduce efficient methodologies for source identification. Therefore, Self-Organising Maps (SOM), Gaussian Process Regression (GPR) and Multivariate Adaptive Regression Splines (MARS) algorithms were utilized

to construct surrogate models and adaptive surrogate models for source identification. The developed models mimic the behaviour of simulation models of groundwater flow and solute transport.

In this chapter, the application of the developed surrogate models in a hypothetical study area is presented. Error-free and erroneous contaminant concentrations data were used to assess the performance of the constructed surrogate models. In this study area, it was assumed that the first group of contaminant concentrations were collected 1.5 years after the start of the first potential contaminant source activity. Therefore, the starting or time of initiation of the source activities is assumed known. The performance evaluation results show that the developed surrogate models could accurately mimic the behaviour of the simulation models of groundwater flow and solute transport, and even substitute the optimisation model for source identification in terms of contaminant source location(s), magnitudes and release history.

### 3.2. Developed Procedures for Source Identification

Generally, implementation of the simulation models for real-world cases is complex and extensively time-consuming. Therefore, to decrease the high computational cost of the complex simulation models, these computationally intensive simulation models have been replaced by response surface methodologies. It is supposed that by accurately constructing these models, the behavior of more sophisticated simulation models can be approximately emulated with much reduced computational time (Gorissen, Couckuyt, Demeester, Dhaene, & Crombecq, 2010). Several types of surrogate models have been constructed by using Kriging, Artificial Neural Network (ANN), MARS and Gaussian Process (GP) as approximate simulators of the physical processes (Razavi et al., 2012). Surrogate Models based Optimization (SMO) is one of the popular surrogate models which has been suggested to reduce computational burden. This approach replaces the

computationally intensive simulation models with a cheaper to run trained surrogate model. Therefore, for obtaining global optimal solution there is no need to run the computational intensive simulation models tens of thousands times (Wang et al., 2014).

The SOM, MARS, and GPR algorithms were utilized as surrogate model types in this study. Utilizing the SOM algorithm as a surrogate model type for contaminant source identification is a new use of this algorithm. The MARS and GPRS algorithms were utilized in different fields as surrogate model types and their potential applicability has been proven. For example, Wang et al. (2014) applied six different algorithms to construct different Adaptive Surrogate Model-based Optimization (ASMO). These algorithms were Quadratic, Regression Tree method, Random Forests (RF), Support Vector Machines (SVM), MARS, ANN, and GPR. The developed surrogate models were utilized to solve two benchmark problems: the Hartman function and calibration of the SAC-SMA hydrologic model (Wang et al., 2014). The study results demonstrate that the GPR-based ASMO generally produce better results than the other developed models. Performance evaluation results of the MARS-based ASMO also demonstrated acceptability. Therefore, in this study, the MARS and GPR algorithms were also selected as the surrogate model types. The results of these surrogate models were compared and utilized for further evolution of the developed surrogate models. The constructed surrogate models represented the simulation models of groundwater flow and solute transport to approximate these models' inputs-outputs values. Once the models were constructed and validated, they could be utilized to accurately predict output values for any new points within the domain of validation. One of the main advantages of these developed methodologies is that they are also capable for source identifications without linking to time-consuming optimisation algorithms.

### 3.2.1. Surrogate Models for Contaminant Source Identification

As mentioned earlier in this chapter, the SOM, GPR and MARS algorithms were utilized to develop surrogate models and Adaptive Surrogate Models (ASM). The developed surrogate models were used for source identification. Figure 3.1 presents the main steps of developing an ASM for source identification. In the next few paragraphs, these steps are explained.

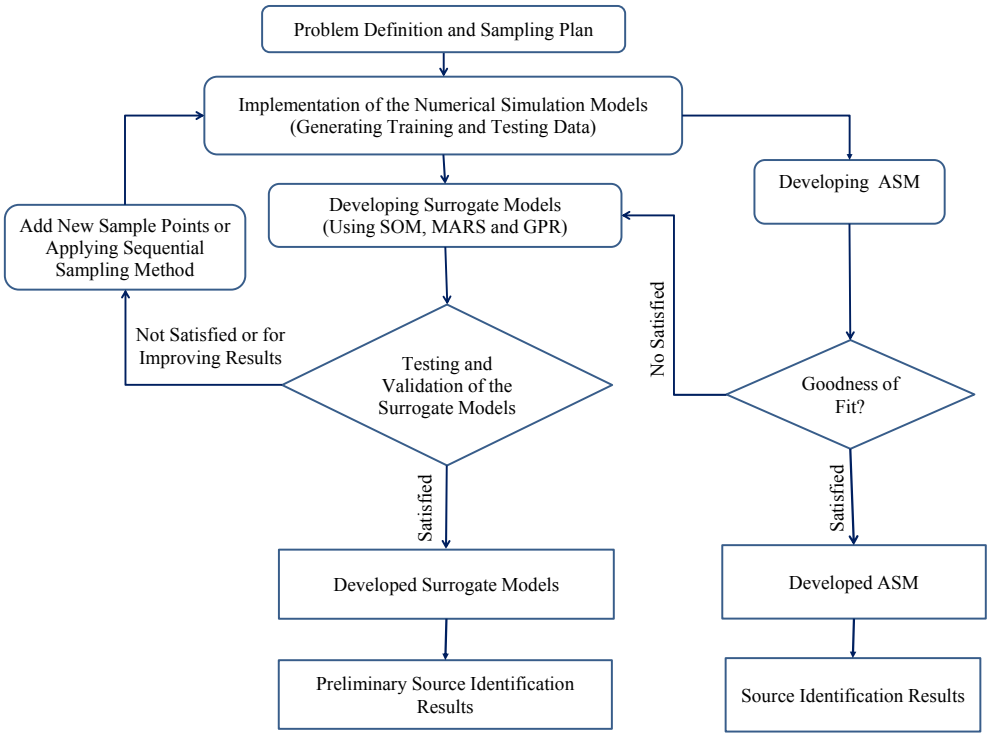



Figure 3.1 Key elements of the ASM methodology for source identification as an inverse problem

1. Problem Definition and sampling plan: The most important variables of the system which are highly dependent on the complexity of origin system are defined (Forrester & Keane, 2009). In source identification problems, contaminant source locations, contaminant source fluxes, and contaminant source release history are the main characteristics of unknown groundwater contaminant sources which are required to be addressed. These characteristics need to be defined as parts of surrogate models' important variables. Moreover, in source identification

problems, measured concentrations data at specific observation locations at specific times are used for source identification. Therefore, the information of concentrations data at specified observation locations at specified times also needs to be defined as do the balance of important variables of the system. Then, for generating qualified sampling points for training and evaluation of surrogate models a suitable random generating methodology need to be selected and utilized. Latin Hypercube Sampling (LHS) was suggested as an appropriate and suitable methodology for this step (Queipo et al., 2005). In source identification problems to generate training data, the LHS could be utilized to randomly generate source fluxes at potential contaminant source locations at possible activity times. For training an accurate surrogate model, an adequate number of sample sets which cover all possible ranges of potential contaminant sources need to be generated.

2. Implementing numerical simulation models: The flow and groundwater simulation models for the contaminated aquifer site are solved at this step. These models are solved to randomly generated contaminant source fluxes at the previous step. As a result, the contaminant concentration values are obtained as the solution of the simulation models of the groundwater flow and solute transport.
3. Construction of surrogate models: The type(s) of the surrogate model(s) need to be addressed in this step. The surrogate model types make surrogate models to represent the simulation model input-output values. The other important question in this step is how surrogate models could be designed to accurately approximate the simulation models of groundwater flow and solute transport with limited numbers of inputs.

4. Model evaluation: This step tests and validates the developed surrogate models for potential applicability by using new sample data sets which are independent of the training data. The model evaluation results can be used for changing the surrogate model type(s) or design(s).
5. Sequential sampling: For improving the results of the developed surrogate models, sequential sampling strategy are applicable. There are various sequential sampling methods such as Maximising Expected Improvement (MEI), Maximising the Probability of Improvement (MPI) and Minimising a Statistical Lower Bound (MSL). Each of the above mentioned three methodologies can lead the surrogate models to go back and find the samples points related to the preliminary results.
6. Developing ASMs: The ASMs are developed by adding the new generated training data (in the previous stage) to the initial training data to effectively improve the accuracy of source identification results.
7. Stop/step 3: If the source identification results for testing data are satisfied, the developed ASMs are ready to identify unknown groundwater contaminant sources as an inverse problem. Otherwise, go to step 3 and modify the architecture or types of the constructed surrogate models.

### 3.2.2. Numerical Simulation Models

The numerical simulation model MODFLOW (Harbaugh, 2005) was used for numerical flow simulation. The three-dimensional equation of groundwater flow through porous media is utilized by MODFLOW which is a partial differential equation that represents the groundwater flow in non-equilibrium, anisotropic and heterogeneous conditions (Harbaugh, 2005). The general governing equation of the groundwater flow through porous media is described by equation (3-1).

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_{zz}\frac{\partial h}{\partial z}\right)\pm W=S_s\frac{\partial h}{\partial t} \quad (3\text{-}1)$$

Where:

$K_{xx}$, $K_{yy}$ and Kzz are the hydraulic conductivity values along the x, y, and z coordinate axes (L/T);

h is the potentiometric head (L);

$S_S$ is the specific storage of the porous media ($L^{-1}$);

t is time (T); and

W is a volumetric flux per unit volume from aquifer as sources (sinks); the negative value represents withdrawal of the groundwater system and vice versa ($T^{-1}$).

MT3DMS (Zheng & Wang, 1999) is the numerical simulation model of mass transport used in this study. Equation (3-2) represents the governing equation of MT3DMS. The MT3DMS uses a partial differential equation. This model has the capability of simulating the advection, dispersion, and chemical reaction processes of the groundwater contaminants transport (Zheng & Wang, 1999).

$$\frac{\partial(\theta C^k)}{\partial t}=\frac{\partial}{\partial x_j}\left(\theta D_{ij}\frac{\partial C^k}{\partial x_j}\right)-\frac{\partial}{\partial x_i}\left(\theta v_i C^k\right)+q_s C_s^k+\sum R_n \quad (3\text{-}2)$$

Where:

$x_i$ *and* $x_j$ represent the distances along the Cartesian coordinate axes (L);

$\theta$ is the subsurface porous media porosity (dimensionless);

$C^k$ is the dissolved concentration of species k ($ML^{-3}$);

t is time (T);

$D_{ij}$ is the hydrodynamic dispersion coefficient tensor ($L^2\ T^{-1}$);

$v_i$ represents the seepage velocity ($LT^{-1}$); it is related to the Darcy flux through the relationship; $v_i = \frac{q_i}{\theta}$;

$q_s$ is volumetric flow rate per unit volume of the groundwater system which represents fluid source (positive) and sinks (negative) ($T^{-1}$);

$C_s^k$ is the concentration of the source or sink flux for species k ($ML^{-3}$); and

$\sum R_n$ is the chemical reaction term ($ML^{-3}T^{-1}$).

### 3.2.3. Self-Organising Map

The Self-Organizing Map (SOM) is an unsupervised learning method that was introduced by T. Kohonen in 1982 to visualise multidimensional data (Kohonen et al., 1996). The main features of this algorithm are its capability in visualising complex non-linear multidimensional input data into a simple geometric relationship (Kohonen & Oja, 2001; Kohonen et al., 1996). This algorithm is widely used to visualize and cluster multidimensional data due to its easy implementation (Kohonen & Oja, 2001; Le Thi & Nguyen, 2014). Usually, SOM results are represented in two dimensions by preserving the topological structure of the input data (Simula et al., 1999).

The main processes of Kohonen's SOM algorithm can be summarised as initialisation, competition, cooperation and adaptation (Bullinaria, 2004a, 2004b, 2014; Dragomir et al., 2014), which are described as follows:

1. Initialisation: A group of high-dimensional input data is quantized by a few weight vectors to a discrete space usually two-dimensional grid (Chalasani & Principe, 2015) and (Amauri, Júnior, Barreto, & Corona, 2015). If X is an m-dimensional continuous input data pattern $\{X = x_1, x_2, \dots, x_m\}$, these data are mapped to output neurons which usually is a two-dimensional discrete space by

the weight matrix $\{W = w_{j1}, w_{j2}, \dots, w_{jm}\}$, where m is the size of the input data and j= 1…, n, where n defines the number of the output space neurons.

2. Competition: For each random sample of input space, the output neurons compete to declare the winner neuron. The winning neuron which has the most similarity to the input data is called the Best Matching Unit (BMU). The distance between the random sample of input space and all weight vectors are calculated by using equation (3-3) which is a squared Euclidean measure.

$$d_j(x) = \min\left(\sum_{i=1}^{m}\left(x_i - w_{ji}\right)^2\right), \qquad \forall\, i = 1, \dots, m \tag{3-3}$$

BMU command in SOM algorithm by searching to find the most similar output neuron to the input vector can be used for finding missing values of an input vector (Figure 3.2). This command in this study was applied for source identification (Hazrati-Yadkoori & Datta, 2017).

3. Cooperation: once the winner neuron is obtained, the weight vector of the winning neuron and all other neurons are updated according to equation (3-4) and are moved to reduce their distance with the input units (Chalasani & Principe, 2015).

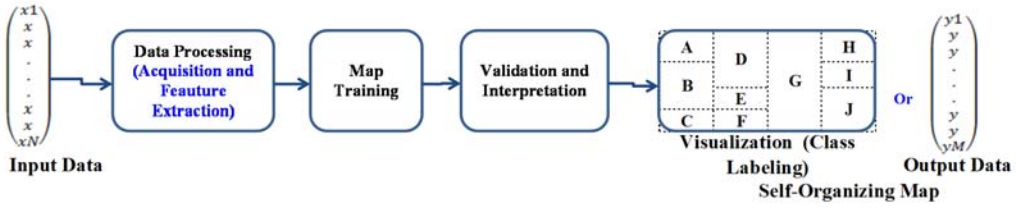
$$W_{ji} = w_{ji}(t) + \eta(t)\, K(j,t)\left[X_i - W_{j,i}(t)\right] \tag{3-4}$$

Where $\eta(t)$: is the learning rate at iteration t; and $K(j,t)$ is a suitable neighbourhood function. This neighbourhood function has the responsibility of preserving the topology of input data (Chalasani & Principe, 2015).

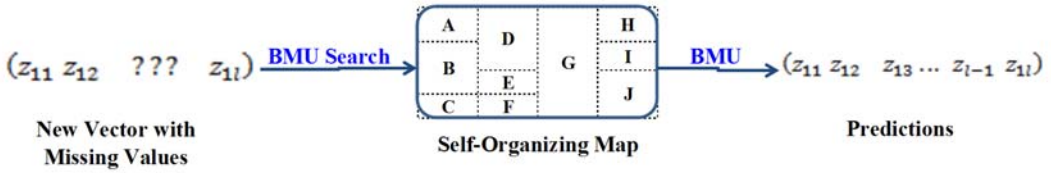
4. Adaptation: The weight adjusting is repeated until a stable map is obtained or the map is converged (Amauri et al., 2015).

The SOM algorithm can also apply for generalisation. This algorithm is capable to interpolate between the initial data and estimate missing values of the system's vectors (Simula et al., 1999). The SOM algorithm's process in clustering is presented in Figure

3.2 (a). Figure 3.2 (b) illustrates how this algorithm is used for predicting the missing values of a new vector (Z) of the system. The software "SOM Toolbox for Matlab 5" (Vesanto, Himberg, Alhoniemi, & Parhankangas, 2000) was used for constructing the SOM-based Surrogate Models (SOM-based SMs).




a)




b)

Figure 3.2.a).The SOM algorithm's process in classification and visualisation, b) The SOM algorithm's process in the prediction of missing values of system's new input vectors.

### 3.2.4. Multivariate Adaptive Regression Splines (MARS)

MARS first was introduced by Friedman (1991). This procedure is for fitting relationships between a response dependent variable which calls as a target variable and a set of predictors (Wang et al., 2014; Zhang & T.C. Goh, 2016). MARS is a nonparametric statistical algorithm which in the training input data is divided into separate piecewise linear segments (splines) with various gradients (slope) (Zhang & T.C. Goh, 2016). In this procedure, usually, the splines are smoothly connected through piecewise curves together. These curves are also known as Basis Functions (BFs) (Zhang

& T.C. Goh, 2016). The MARS model can be expressed as equation (3-5) (Wang et al., 2014):

$$\hat{y} = \beta_0 + \sum_{j=1}^{M} \beta_j B_j(\vec{x}) \quad (3\text{-}5)$$

Where:

$\vec{x} = (x_1, x_2, ..., x_p)$: The predictor vector.

$B_j$ : The jth basis function which can be a spline function or interactions of two or more BFs.

$\beta_0 \; and \; \beta_j$ : Constant coefficients which are calculated by minimising the sum of the squared residuals.

In the MARS algorithm, a forward-backwards approach is performed to develop the final model. The MARS evaluates the performance of constructed models by using Generalised Cross Validation (GCV). GCV is the mean squared residual error divided by a penalty depending on the model complexity. Finally, the best model is selected as one that has the least GCV (Wang et al., 2014).

The main advantages of the MARS algorithm are 1. its accuracy in approximating the behaviour of non-linear multidimensional data; 2. reducing the scale of large-scale problems by selecting the effective variables; and 3. self-testing capability at a high speed ("SPM User Guide, Introduction to MARS," 2013).

The Salford Predictive Modeler 8.0 software was utilized to use the MARS algorithm to develop MARS-based Surrogate Models (MARS-based SM) and adaptive surrogate model ("Salford Predictive Modeller 8 ", 2017).

### 3.2.5. Gaussian Process Regression (GPR)

GPR is a supervised learning regression Model. GPR models are flexible nonlinear interpolating techniques which are based on the training data (Belyaev et al., 2016). This technique can explore unknown functions of multidimensional data which map input data to output data (explore their interactions) (Schulz, Speekenbrink, & Krause, 2016). This technique can approximate any multidimensional data (Retherford & McDonald, 2010). These capabilities make GPR a popular and widely utilized surrogate models' technique. The GPR models are defined by two functions: mean function $m(\vec{X})$ and covariance function $k\left(\vec{X}, \overrightarrow{X'}\right)$. These functions can be described by equations (3-6) and (3-7), respectively (Wang et al., 2014):

$$m(\vec{X}) = \mathrm{E}[f(\vec{X})] \tag{3-6}$$

The mean function represents the expected function value for input X (Schulz et al., 2016).

$$k = \left(\vec{X}, \overrightarrow{X'}\right) = \mathrm{E}\left[\left(f(\vec{X}) - m(\vec{X})\right)\left(f(\overrightarrow{X'}) - m\left(\vec{X'}\right)\right)\right] \tag{3-7}$$

The covariance function models the interactions between the function values at different input points X and $X'$ (Schulz et al., 2016).

A GP model can be written as equation (3-8) (Wang et al., 2014):

$$f(\vec{X}) \sim GP\left(m(\vec{X}),\ k = \left(\vec{X}, \overrightarrow{X'}\right)\right) \tag{3-8}$$

### 3.2.6. Assessment of the Performance of the Developed Models

The performance of the developed surrogate models was assessed by considering two assumptions regarding errors in concentration measurements:

1. All the model parameters and measured contaminant concentrations are precisely known; and
2. Simulated contaminant concentrations were used as measured concentration values after perturbing the simulated concentrations with different amounts of random errors, i.e., 5, 10, 15, 20, 25 and 30%.

Normalised Absolute Error of Estimation (NAEE) (Equation (3-9)) was used as a measure to calculate a normalised error of estimation (Jha & Datta, 2013):

$$\mathrm{NAEE}(\%) = \frac{\sum_{i=1}^{S}\sum_{j=1}^{N}\left|\left(q_i^j\right)_{est} - \left(q_i^j\right)_{act}\right|}{\sum_{i=1}^{S}\sum_{j=1}^{N}\left(q_i^j\right)_{act}} \times 100 \qquad (3\text{-}9)$$

Where S and N are the numbers of contaminant source(s) and transport stress periods, respectively. $\left(q_i^j\right)_{act}$ and $\left(q_i^j\right)_{est}$ are actual and estimated source flux at source number i in stress period j, respectively.

## 3.3. Application of the Developed Surrogate Models for Source Identification

### 3.3.1. Study Area

Information from a homogeneous confined aquifer (Figure 3.3) was applied to assess the performance of the developed surrogate models. In this study area, the east and west boundaries were assumed to be variable head boundaries. The north and south boundaries were assumed to be specified head boundaries. The specified heads for north and south boundaries were considered to be 35 metres and 25 metres, respectively. The locations and fluxes of the potential contaminant sources (PCS1, PCS2, and PCS3) are shown in

Table 3.2. Table 3.1 shows the Hydrogeologic parameter values of this illustrative study area (Hazrati-Yadkoori & Datta, 2017).

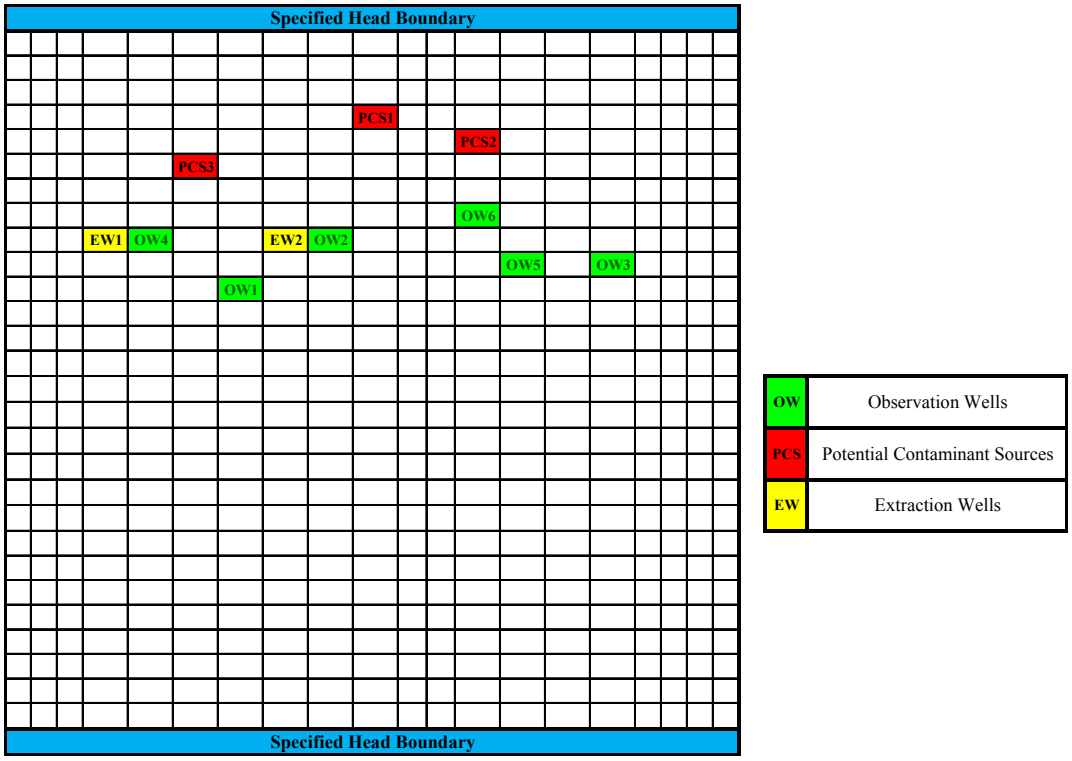



Figure 3.3 Important features of the used illustrative study area

Six observation wells (OW1 to OW6) and two extraction wells (EW1 and EW2) were considered to be in this study area (Figure 3.3). The total time of simulation duration was divided into five separate stress periods (ST1 to ST5). The duration of each of the ST1 to ST4 was 183 days. The ST5 was of 2200 days duration. PCS1 to PCS3 were assumed to be active only in the stress periods ST1 to ST4. Table 3.3 present the extraction rates for each stress period at the extraction wells. It was assumed that the contamination was detected at the end of ST3, or just 1.5 years after the start of first source activity. The breakthrough curves at the selected observation wells utilized for source identification are shown in Figure 3.4.

Table 3.1 Aquifer characteristics and dimensions of the study area

| Parameter | Unit | Value |
|---|---|---|
| Maximum length | metre | 1000 |
| Maximum width | metre | 1500 |
| Saturated thickness, b | metre | 7.6 |
| Grid spacing in x-direction | metre | 50 |
| Grid spacing in y-direction | metre | 50 |
| Horizontal Hydraulic Conductivity | metre /day | 18 |
| Porosity | - | 0.25 |
| Longitudinal Dispersivity | metre | 35 |
| Ratio: H/L Dispersivity | - | 0.2 |
| Specific Yield | - | 0.2 |
| Confined Storage Coefficient | - | 0.2 |
| Initial Contaminant Flux | Kg/day | 0-100 |

Table 3.2 Characteristics of the contaminant sources

| Potential contaminant source location (row, column) | Contamination source flux (Kg/day) | | | | |
|---|---|---|---|---|---|
| | ST1 | ST2 | ST3 | ST4 | ST5 |
| PCS1 (5,10) | 0 | 0 | 0 | 0 | 0 |
| PCS2 (6,13) | 60 | 20 | 45 | 50 | 0 |
| PCS3 (7,6) | 80 | 58 | 22 | 30 | 0 |

Table 3.3 Characteristics of extraction wells

| ID | ROW | Column | Extraction rate ($m^3$/day) | | | | |
|---|---|---|---|---|---|---|---|
| | | | ST1 | ST2 | ST3 | ST4 | ST5 |
| EW1 | 10 | 4 | -100.25 | -100.25 | -68 | -16 | -49 |
| EW2 | 10 | 8 | -100.25 | -80.2 | -96 | -100.25 | -88 |

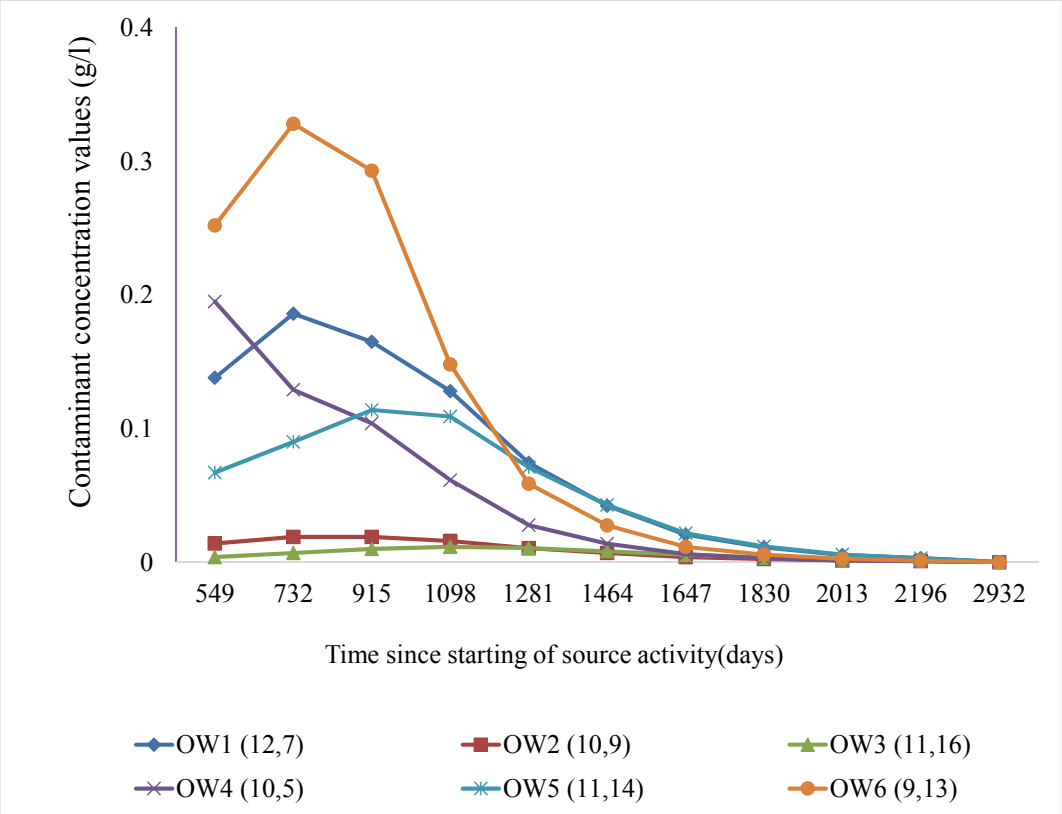



Figure 3.4 Breakthrough cureves at the observation wells used for source identification

### 3.3.2. Results

In this evaluation process, three different data mining tools were utilized to develop different surrogate models. The SOM, GPR and MARS tools were utilized to construct the SOM, GPR and MARS-based Surrogate Models (SOM, GPR and MARS-based SMs) and Adaptive Surrogate Models (SOM, GPR and MARS-based ASMs) for source identification. The following steps were followed to develop surrogate models and apply them to the illustrative study area. Then, the sequential sampling method was utilized and adaptive surrogate models were developed for source identification.

1. Problem definition and sampling plan: Contaminant source fluxes at PCS1 to PCS3 at ST1 to ST4 and their corresponding contaminant concentration

magnitudes at six observation wells and specific times (Figure 3.4) were considered as the important variables of the defined study area. The LHS was used to generate random initial sample sets (one group of 1000 initial sample sets). These sample sets were generated by assuming that PCS1 to PCS3 were active through the ST1 to ST4. Also, three groups of 100 sample sets were generated by assuming that in each group at least one of the potential contaminant sources was inactive. It was assumed that source fluxes varied in the range of 0-100kg/day for PCS1 to PCS3.

2. Generating training data: The numerical simulation models of flow (MODFLOW) and solute transport (MT3DMS) (within GMS 7) were implemented to obtain adequate sample sets for training the surrogate models. The training sample sets consist of randomly generated contaminant source fluxes at the previous step and their corresponding contaminant concentration values at observation wells at specified times (Figure 3.4). Figure 3.5 presents a typical contaminant plume 732 days after the start of the first source activity. Table 3.4 shows a typical input for training surrogate models in this study. This input consists of five sample sets. Each sample set consists of randomly generated contaminant source fluxes at PCS1 to PCS3 at ST1 to ST4. Also, each set consists of corresponding contaminant concentration magnitudes at six observation wells (OW1 to OW6) at three stress periods (ST3 to ST5).

Table 3.4 Typical sample sets for training a surrogate model

| Contamination source flux (Kg/day) | | | | | | | | | | | | Contaminant concentration (g/l) | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PCS1-ST | | | | PCS2-ST | | | | PCS3-ST | | | | OW1 | | | OW2 | | | OW3 | | | OW4 | | | OW5 | | | OW6 | | |
| Stress Period (ST) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 3 | 4 | 5 | 3 | 4 | 5 | 3 | 4 | 5 | 3 | 4 | 5 | 3 | 4 | 5 | 3 | 4 | 5 |
| 42 | 44 | 41 | 97 | 11 | 16 | 58 | 23 | 43 | 40 | 29 | 35 | 0.1 | 0.1 | 0.0 | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 | 0.1 | 0.0 | 0.2 | 0.3 | 0.0 |
| 56 | 73 | 24 | 54 | 35 | 27 | 35 | 22 | 3 | 62 | 87 | 87 | 0.1 | 0.2 | 0.0 | 0.2 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.2 | 0.0 | 0.0 | 0.1 | 0.0 | 0.2 | 0.2 | 0.0 |
| 73 | 51 | 59 | 59 | 0 | 36 | 48 | 10 | 59 | 95 | 21 | 39 | 0.1 | 0.2 | 0.0 | 0.2 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 | 0.0 | 0.1 | 0.0 | 0.2 | 0.3 | 0.0 |
| 65 | 69 | 5 | 49 | 32 | 50 | 39 | 17 | 50 | 29 | 23 | 2 | 0.1 | 0.1 | 0.0 | 0.2 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.0 | 0.1 | 0.1 | 0.0 | 0.3 | 0.3 | 0.0 |
| 30 | 47 | 9 | 32 | 55 | 48 | 8 | 46 | 71 | 84 | 17 | 9 | 0.2 | 0.2 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.1 | 0.0 | 0.1 | 0.1 | 0.0 | 0.3 | 0.2 | 0.0 |

Table 3.5 Typical sample sets with missing data for testing a surrogate model

| Contamination source flux (Kg/day) | | | | | | | | | | | | Contaminant concentration (g/l) | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PCS1 | | | | PCS2 | | | | PCS3 | | | | OW1 | | | OW2 | | | OW3 | | | OW4 | | | OW5 | | | OW6 | | |
| Stress Period (ST) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 3 | 4 | 5 | 3 | 4 | 5 | 3 | 4 | 5 | 3 | 4 | 5 | 3 | 4 | 5 | 3 | 4 | 5 |
| | | | | | | | | | | | | 0.10 | 0.14 | 0.00 | 0.04 | 0.10 | 0.00 | 0.00 | 0.01 | 0.00 | 0.14 | 0.10 | 0.00 | 0.07 | 0.08 | 0.00 | 0.23 | 0.25 | 0.00 |
| | | | | | | | | | | | | 0.08 | 0.16 | 0.00 | 0.03 | 0.08 | 0.00 | 0.01 | 0.01 | 0.00 | 0.14 | 0.14 | 0.00 | 0.13 | 0.20 | 0.00 | 0.63 | 0.62 | 0.00 |
| | | | | | | | | | | | | 0.13 | 0.16 | 0.00 | 0.23 | 0.31 | 0.00 | 0.01 | 0.01 | 0.00 | 0.15 | 0.09 | 0.00 | 0.10 | 0.14 | 0.00 | 0.45 | 0.41 | 0.00 |
| | | | | | | | | | | | | 0.09 | 0.22 | 0.00 | 0.13 | 0.16 | 0.00 | 0.01 | 0.01 | 0.00 | 0.18 | 0.20 | 0.00 | 0.09 | 0.09 | 0.00 | 0.21 | 0.26 | 0.00 |
| | | | | | | | | | | | | 0.15 | 0.24 | 0.00 | 0.21 | 0.25 | 0.00 | 0.01 | 0.01 | 0.00 | 0.20 | 0.18 | 0.00 | 0.10 | 0.12 | 0.00 | 0.33 | 0.39 | 0.00 |

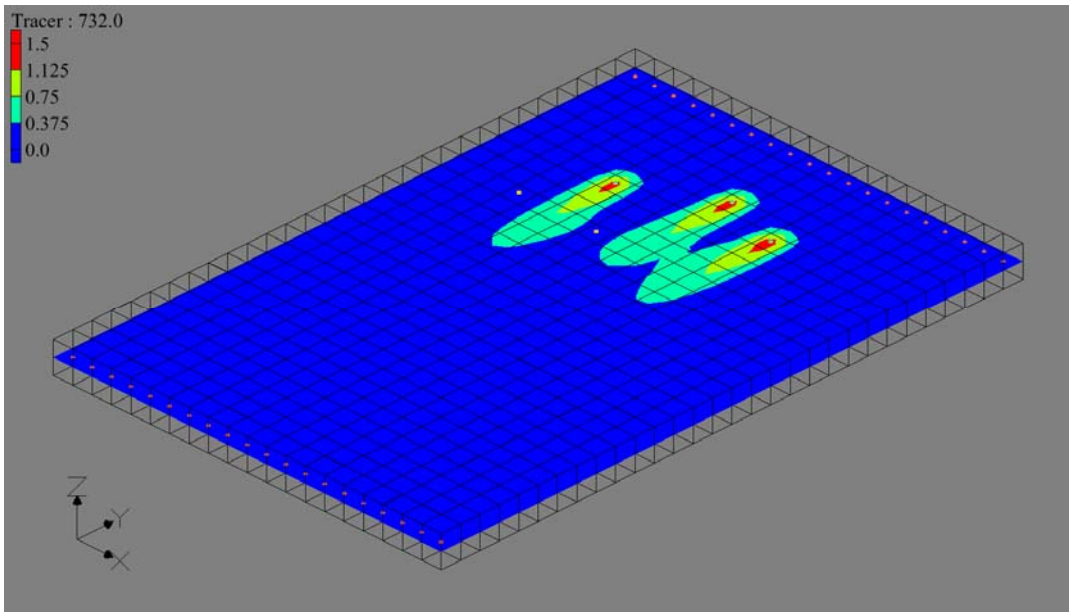



Figure 3.5 A typical concentration plume 732 days after start of first source activity

3. Developing the surrogate models: The SOM, MARS and GPR algorithms were used to develop surrogate models.

Same training data was utilized to develop different surrogate models. However, the design of variables in the SOM-based SM is different, because of different characteristics of the SOM algorithm compared with the GPR and MARS-based SMs. The SOM-based SMs were developed by using training data (Table 3.4) in a single run. Then, the developed SOM-based SMs were applied for source identification without using an optimisation model. As previously mentioned, in this chapter, the optimisation model was not solved for source identification. In the SOM-based SM case, the source identification based on concentration measurement data was accomplished by running the SOM-based SM in an inverse mode. The SOM-based SM was used to estimate the contaminant source characteristics as the output, while the concertation measurements resulting from the unknown contaminant sources were utilized as inputs. The same

constructed SOM-based SMs also applied for estimating contaminant concentration values at selected observation wells at specific times.

For developing the MARS/GPR-based SMs which could independently apply for source identification without using an optimisation model, the known variables or predictors were considered to be measured contaminant concentrations at observation wells at specific times. The unknown or target variables were also considered to be the contaminant source fluxes at potential contaminant locations at specific times. Then, the MARS/GPR algorithms were utilized to develop the MARS/GPR models for all the target variables. By using the MARS or GPR algorithm, developing a prediction model for each target variable is a necessity. Once all the MARS or GPR models related to all the target variables were developed, the MARS or GPR-based SMs could be developed by integrating all the MARS prediction models or all the GPR prediction models, respectively. Finally, by using the simulated or measured contaminant concentration values, the source identification results could be obtained.

4. Evaluation of the developed models: The performance of the developed surrogate models were assessed by using 120 new random sample sets. The contaminant source fluxes of these sample sets were generated randomly by using the LHS method in the range of 0-100kg/day. Then, the corresponding contaminant concentration values at specific times at specific observation wells were obtained by using the simulation models.

In the SOM-based SM case, since the definition of the BMU of the SOM algorithm (equation (3-3)) is similar to the objective function of the source identification problem, the BMU of the SOM algorithm was used for estimating unknown characteristics of potential contaminant sources. Consequently, using the BMU command eliminated the

necessity for using any complex and explicit optimisation model. This algorithm, by using the information of known components of the input vector, estimated the unknown components of the input vector. By searching for the BMU or the most similar vector and using information of known components of the input vector, the most similar vector was recognised and missing values of the input vector were estimated.

In this evaluation for source identification, the contaminant concentration values at specified observation wells at specific times were considered to be known variables of an input vector. The input vector needs to have the same dimension as the input vectors of the training data. Table 3.5 shows a typical input for testing data when the SOM, MARS and GPR-based SMs were used for source identification as an inverse problem. In this table, magnitudes of contaminant concentration values at six observation wells (OW1 to OW6) at the end of three stress periods (ST3 to ST5) were assumed to be known variables of the developed surrogate models. The contaminant source fluxes at PCS1 to PCS3 at four stress periods (ST1 to ST4) were considered to be unknown variables.

In the SOM algorithm, the SOM Map quality could be assessed by the Quantisation Error (QE) which is a widely utilized criterion for evaluating the SOM Maps. The QE gradually decreases by increasing the map sizes. The earlier studies indicate that suitable numbers of SOM map units have an essential role in the accuracy and performance of the SOM algorithm (Di Mauro et al., 2016). Therefore, different SOM-based SMs representing different numbers of SOM map units were constructed. In these scenarios, the number of observation wells and the number of initial sample sets were maintained constant at six and 1300, respectively. The developed SOM-based SMs were also used to estimate contaminant concentration values at specified locations at specific times when the contaminant sources and their characteristics were known (Figure 3.6).

The obtained solution results for source identification and estimating contaminant concentrations at observation wells of testing data are presented in Table 3.6. The performance evaluation results lead to select the best candidate SOM-based SM among the developed SOM-based SMs for the illustrative study area. The results indicate a consistency in the solution result, and the best results were reached by using $130 \times 130$ map units. An important constraint in these evaluations of various scenarios was the CPU capacity, which was exceeded by increasing the numbers of SOM map units beyond 120×120 (Figure 3.7). Figure 3.7 also presents the QE values for various SOM-based SMs representing different numbers of SOM map units. The SOM-based SM, which consisted of 1300 initial sample sets and 100×100 map units, was selected as SOM-based SM among the constructed SOM-based SMs. This surrogate model was selected regarding the performance evaluation results of the developed surrogate models in terms of NAEE values, QE values, and the required times for constructing the SOM-based SMs in different scenarios.

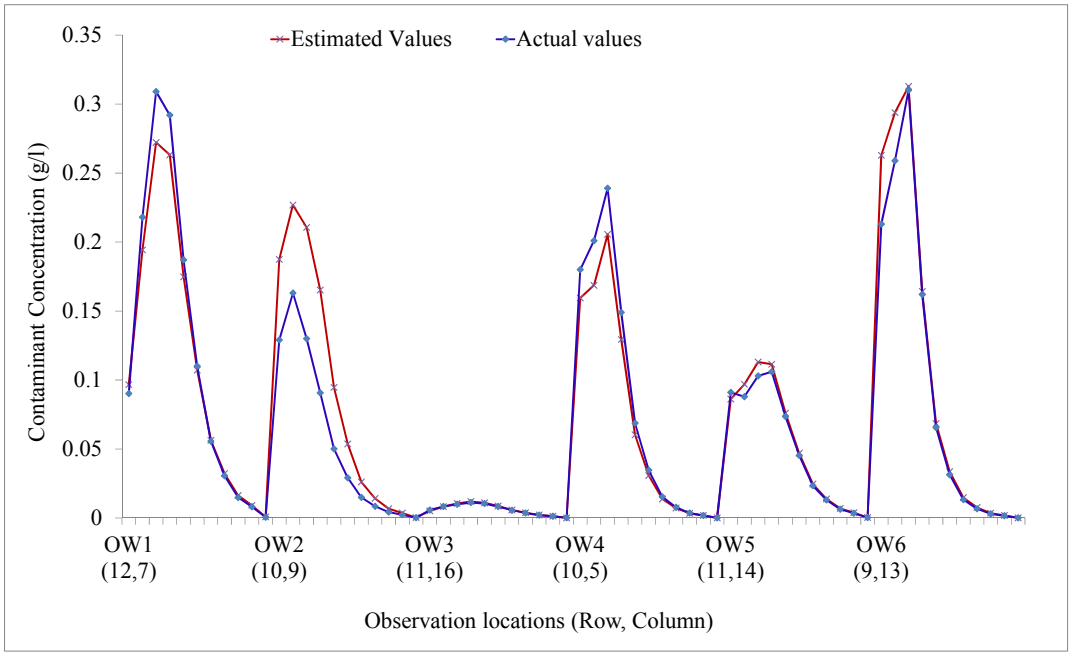



Figure 3.6 The result obtained from the selected SOM-based SM for estimating the contaminant concentration values for a set of test data at six observation wells (NAEE is equal to 16.4%)

Table 3.6. Normalized Absolute Error of Estimation for different developed SOM-based SMs

| SOM's Map characteristics | | | Source identification | Estimation of contaminant concentration |
|---|---|---|---|---|
| Map shape | Neighbourhood function | Numbers of map units | NAEE (%) | NAEE (%) |
| Rectangular | Gaussian | 50×50 | 40.9 | 18.1 |
| | | 75×75 | 40.0 | 17.7 |
| | | 100×100 | 40.0 | 16.3 |
| | | 110×110 | 40.5 | 16.5 |
| | | 120×120 | 40.2 | 16.5 |
| | | 130×130 | 39.7 | 15.8 |

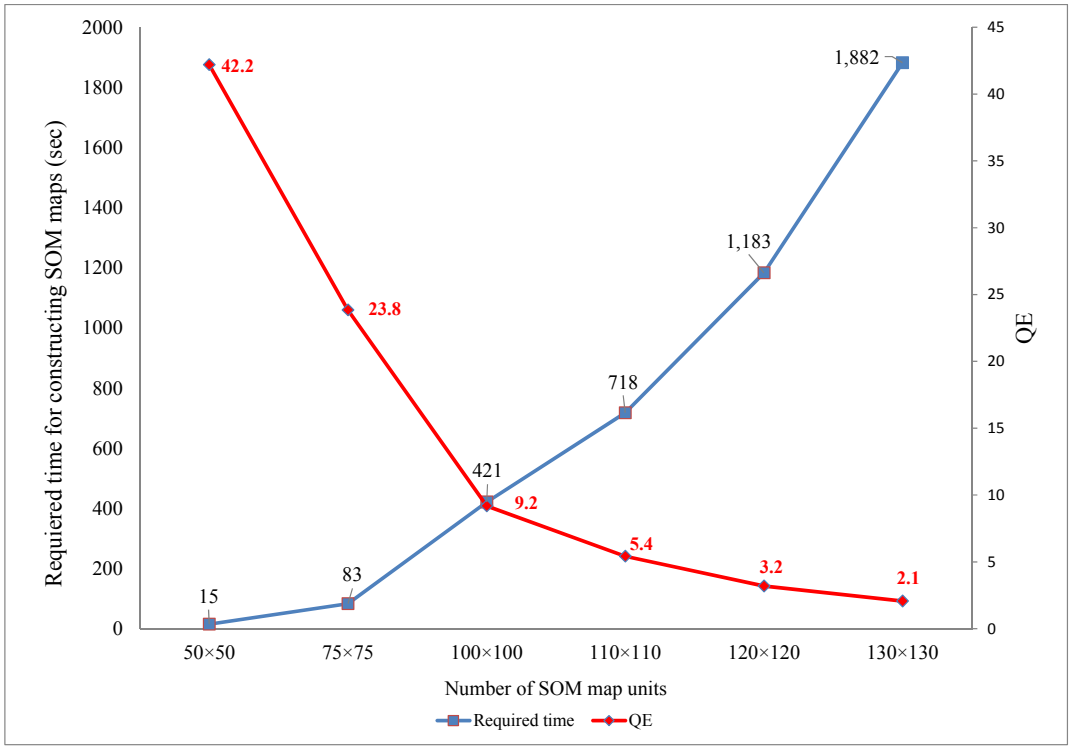



Figure.3.7 Required times for constructing various SOM-based SMs for different scenarios

As mentioned at the previous stage, the developed MARS and GPR-based SMs by using the simulated contaminant concentration values of testing data directly could be utilized for source identification. Therefore, the developed MARS and GPR-based SMs were also evaluated by using same testing data. The average NAEE obtained for the testing data

(120 sample sets) were equal to 4.9 and 6.6% by using the MARS and the GPR-based SMs, respectively. The results of the MARS and the GPR-based SMs indicate that the accuracy of these surrogate models is better than the accuracy of the SOM-based SMs (Table 3.6). However, comparing the developed SOM-based SMs to the developed MARS- and GPR-based SMs shows three advantages for the developed SOM-based SMs, such as:

a. The GPR and the MARS-based SMs are not able to screen dummy sources while the SOM-based SMs do it properly; and
b. The development process of GPR and MARS-based SMs is not as easy as the development process of the SOM-based SMs. For example, implementation of the SOM algorithm for developing a SOM-based SM for a complex system is easy and can be developed at one shot or in a single run, unlike the MARS and the GPR algorithms.
c. As previously mentioned, by using the GPR and MARS algorithms, for each target variable, a separate prediction model is required to be developed. Then, by integrating all the developed prediction models, the MARS- or GPR-based SMs can be developed.

5. Source identification: The developed and evaluated surrogate models using the observed contaminant concentration (synthetically generated for known source fluxes) values (Figure 3.4) were utilized for source identification. The obtained results of the developed surrogate models for source identification are presented in Figure 3.8. The obtained source identification results of the MARS, GPR and SOM-based SMs in terms of NAEE equal to 3.7, 4 and 28.5%, respectively.

Illustrated results in Figure 3.8 indicate that the MARS- and the GPR-based SMs show more accuracy compared with the SOM-based SMs.

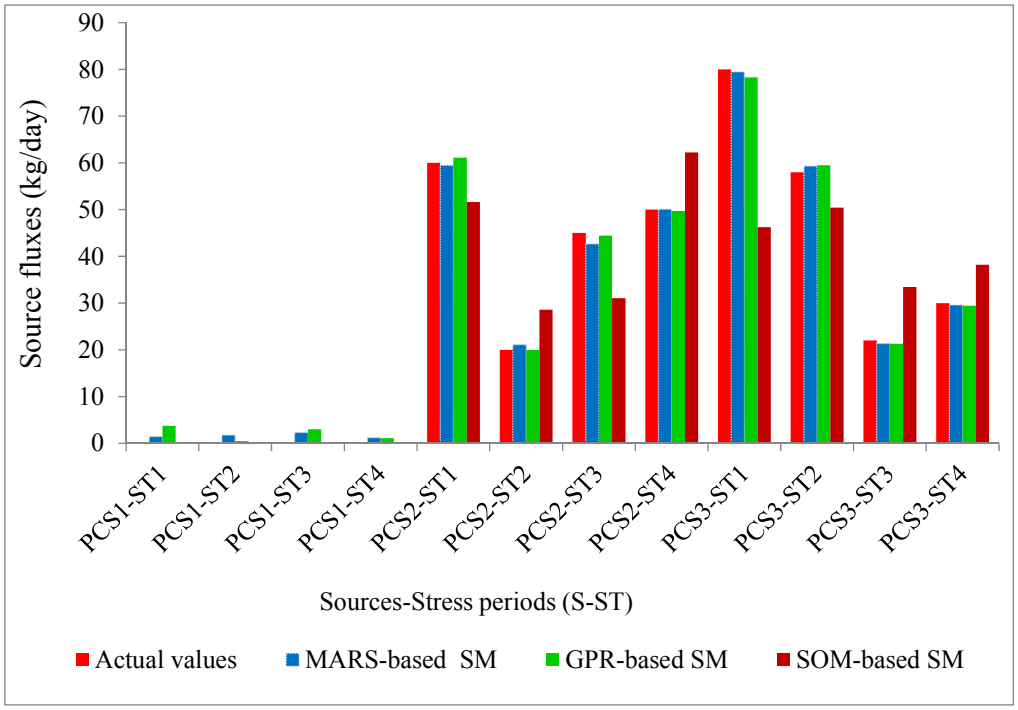



Figure 3.8 Source identification results of the developed surrogate models

However, the capability of the SOM algorithm in the classification of multidimensional input data leads the SOM-based SMs to screen the dummy source(s) i.e., not actual sources but included as potential sources precisely. For example, for testing sample sets, the SOM-based SMs accurately could screen the dummy source(s) among all the potential contaminant sources in all the cases. Therefore, for updating the surrogate models and improving source identification results, new sample sets could be generated. These sample sets were generated based on the preliminary source identification results of the selected SOM-based SM by using observed contaminant concentrations data. By considering the results of the selected SOM-based SM at this stage, which indicated that PCS1 was not an actual source, new sample sets could be generated and utilized for updating the developed surrogate models.

6. Sequential sampling method and adding new sample points: Based on the result of the previous stage, 500 new sample sets were generated by considering PCS1 as a dummy source. These 500 new sample sets were randomly generated by using the LHS. Then, these new sample sets were added to the initial sample sets to develop Adaptive Surrogate Models (ASM).
7. Developing ASM: ASMs were constructed for contaminated aquifers by using the MARS and GPR algorithms. These algorithms were selected because the developed surrogate models obtained by using the MARS and GPR algorithms showed more accurate results compared to the SOM-based SMs results for source identification. The obtained results for this stage are shown in Figure 3.9. The obtained source identification results of the MARS-based ASM and the GPR-based ASM in terms of NAEE equal 1.9 and 2.1%, respectively.

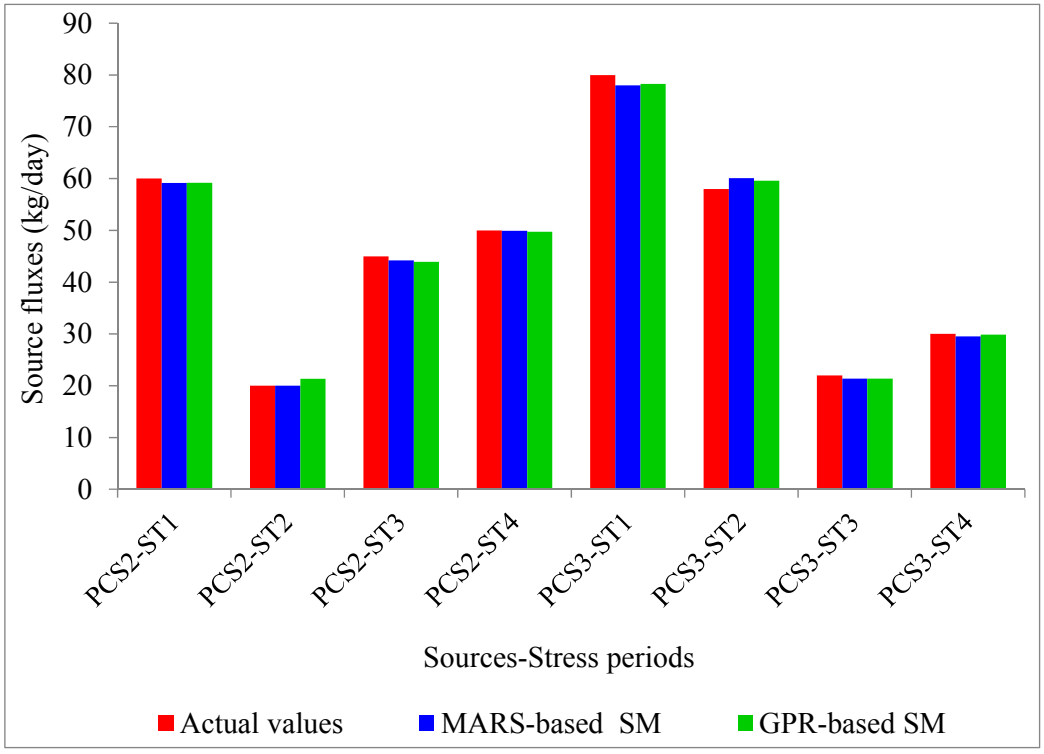


Figure 3.9 Source identification results of the developed ASM

One of the advantages of the GPR algorithm is its capability in estimating the 95% prediction intervals which can be useful for erroneous observation data. Figure 3.10 presents the 95% sources estimation intervals corresponding to observe contaminant concentrations data by using the GPR-based ASM.

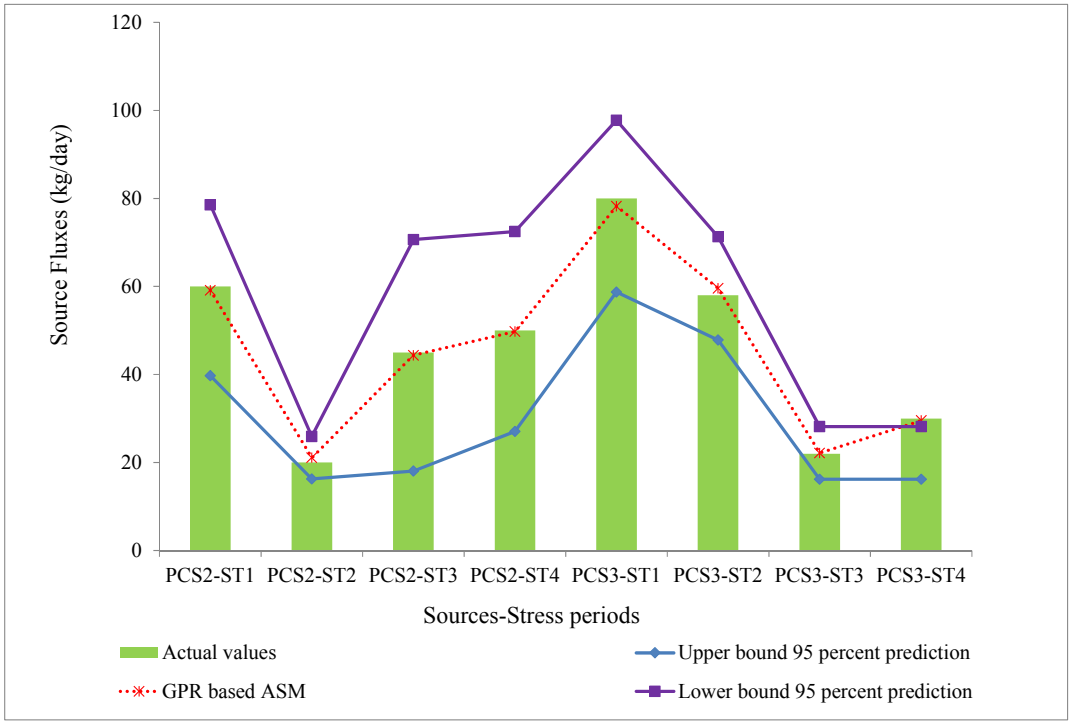



Figure 3.10 Obtained results by using the GPR-based ASM for source identification and its 95% source estimation intervals for observed contaminant concentration values

Moreover, for further evaluations of the developed models, synthetic erroneous concentration measurements data were used for performance evaluation purposes. For this purpose, simulated contaminant concentrations were perturbed with varied amounts of random errors, i.e., 5, 10, 15, 20, 25 and 30% of simulated values. The simulated contaminant concentrations measurements at observation wells were assumed to incorporate 5, 10, 15, 20, 25 and 30% random errors. The equation (3-10) was applied

for synthetically generating the perturbed concentration measurement values with random errors (Jha & Datta, 2013).

$$C_{per} = C_S + a \times b \times C_S \quad (3\text{-}10)$$

Where $C_{per}$ and $C_S$ are perturbed concentration measurement values and simulated concentration values, respectively. a and b are maximum deviation expressed as a percentage and a random fraction between +1 and -1 obtained by using the LHS.

The source identification results obtained with these erroneous concentration measurements are illustrated in Figure 3.11. These solution results (Figure 3.11) indicate that the accuracy of the MARS- and GPR-based ASMs results significantly worsened when the contaminant concentration values are erroneous, especially when the incorporated errors are 10% or larger. The deterioration of accuracy of results increased by incorporating larger errors. However, obtained results of the MARS- and the GPR-based ASMs are more accurate than the SOM-based ASM results for scenarios with error free, 5% and 10%. The source identification performances of the SOM-based ASM do not substantially change for all the scenarios with erroneous concentration measurements. It can be concluded that the source identification results are more sensitive to the measurement errors, when the utilized surrogate models are more precise.

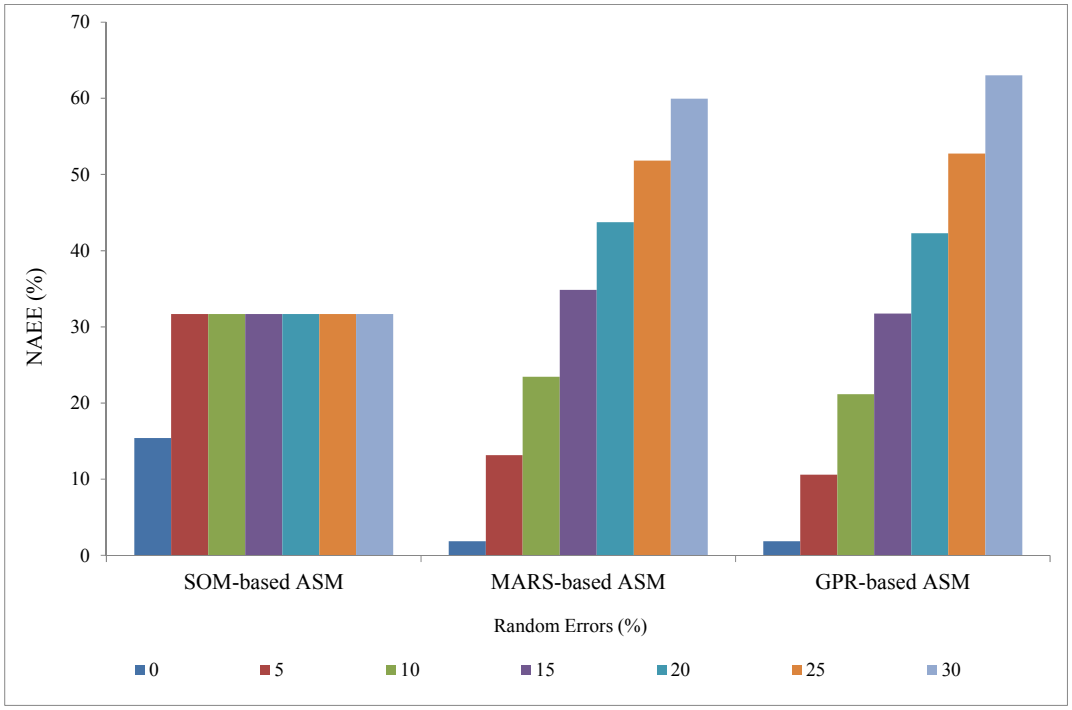



Figure 3.11 NAEE of the ASMs for source identification by using perturbed concentration measurement values

### 3.4. Discussion

The developed SOM, GPR and MARS-based SMs and ASMs were applied for source identification in an inverse mode without using an optimisation model. The developed surrogate models were used to estimate the contaminant source characteristics as the output, while the concertation measurements resulting from the unknown contaminant sources were utilized as the inputs. The performance evaluation results of the developed surrogate models indicate the potential applicability of the developed surrogate models for source identification. The results demonstrate that the SOM, MARS and GPR algorithms are powerful tools to develop surrogate models and ASMs (Figure 3.8 and 3.9). However, each of these tools has its advantages and disadvantages.

For example, the capability of the SOM algorithm in classifying leads the SOM-based SMs to be very efficient in screening dummy sources among all potential contaminant

sources. Consequently, this capability may make the SOM algorithm a potentially powerful tool in source identification problems. In source identification problems, detecting the dummy sources among all potential contaminant sources is one of the main questions of these problems that need to be addressed. The performance evaluation results of the developed SOM-based SMs indicate that the SOM-based SMs could accurately find the solution of this question. Another advantage of the developed SOM-based SMs is the consistency of the solution results for ideal (with error-free data) and real (with erroneous data) scenarios. The other advantage of the SOM-based SMs is that the developing and using process of these surrogate models for source identification is considerably easier than the MARS and GPR SMs' processes. For example, by applying the MARS and GPR algorithms, for each unknown variable, separate models need to be built. Then, these models need to be integrated to develop surrogate models for source identification. The SOM-based SMs can be constructed in a single run.

On the other hand, the MARS- and GPR-based SMs and ASMs comparatively show more precise results than the SOM-based SMs. For instance, the source identification results of the MARS- and the GPR-based SMs in terms of average NAEE for testing data were equal to 4.9 and 6.6%, respectively. The other advantage of using the GPR algorithm is its capability in estimating the 95% prediction intervals ("Gaussian Process Regression," 2017). This capability can be useful for scenarios which its observations data incorporate with erroneous data. These limited results showed that the developed surrogate models could approximate groundwater flow and transport simulation models properly. The developed surrogate models were also capable to characterize unknown contaminant sources independently without linking to an optimisation model.

### 3.5. Conclusion

This chapter presents applications of different surrogate models for source identification. The performance of the constructed surrogate models was assessed for an illustrative aquifer site with missing contaminant concentration data. The randomly generated source fluxes at potential contaminant sources at potential activity times and corresponding simulated contaminant concentrations at six observation wells at limited specific times were utilized to develop the surrogate models. Different surrogate models were developed by using the SOM, GPR and MARS algorithms. Same training data were utilized to develop the SOM, MARS and GPR-based SMs. These surrogate models were utilized in an inverse mode for source identification, for an ideal scenario of error-free concentration data, as well as scenarios with different degrees of erroneous concentration measurements data. In addition, an improved version of surrogate models based on the information obtained at the preliminary stage, i.e. Adaptive Surrogate Models (ASM) were developed for source identification. The main conclusions that can be drawn from these limited performance evaluation results are:

1. The SOM, MARS and GPR-based SMs are potentially efficient methods to approximate the simulation models of groundwater flow and transport, MODFLOW and MT3DMS, respectively.
2. The constructed methodology can be utilized as an alternative approach for source identification, which can potentially eliminate the necessity for other widely utilized procedures, i.e., the linked simulation-optimisation procedure. When additional information based on earlier (preliminary) source identification results were incorporated in the training stage, it could increase the efficiency of the developed methodology in terms of decreasing computational time and increasing accuracy of source identification results.

3. The SOM-based SMs could identify the active sources among all the potential contaminant sources more accurately. Therefore, the SOM-based SMs' capabilities in screening dummy sources and decreasing the number of surrogate models' variables, may decrease the complexity of source identification problems.
4. The developed ASMs can efficiently characterize unknown groundwater contaminant sources (Figure 3.9).
5. In developing the process of the SOM-based SMs, the size of SOM map units is important (Table 3.6 and Figure 3.7). The best size needs to be chosen due to the memory of the PC utilized, the number of variables, and initial sample sizes.
6. Assessment results of the performance of the developed surrogate models demonstrate potential applicability of the SOM, MARS and GPR-algorithms as the surrogate model types for source identification problems with error-free and erroneous data (Figure 3.11).
7. The performance evaluation results show that the accuracy of the MARS- and the GPR-based ASMs results significantly decreased when the contaminant concentration values were incorporated with larger errors (Figure 3.11).
8. The SOM-based SMs seem to perform satisfactorily when concentration measurement data were erroneous.
9. The developed surrogate models may provide a feasible procedure for source identification without the necessity for a linked simulation-optimisation model.

In the next chapter, first, developed procedures for source identification and monitoring network design are explained in detail. Then, the application of the developed methodologies to a hypothetical study area is discussed.

# 4. Application of Surrogate Model based Optimization in Conjunction with Monitoring Network Design Procedure for Source Identification

## 4.1.Introduction

Some contents of this chapter have been released to present in the following journal papers:

- Hazrati. Y, S., & Datta, B. (2017b). Self-Organizing Map based Surrogate Models for Contaminant Source Identification under Parameter Uncertainty. *International Journal of GEOMATE*, 13(36), 8. doi:http://dx.doi.org/10.21660/2017.36.2750
- Hazrati. Y, S., & Datta, B. (2017).Characterization of Groundwater Contaminant Sources by Utilizing MARS based Surrogate Model Linked to Optimization Model. Springer Book Series "Advances in Intelligent Systems and Computing". To be published.

Also, some parts of the next two sections (4.2 and 4.3) have been presented in the literature review chapter since they concern concepts relevant to the chapter. In this chapter, a sequential approach that combines developed surrogate model based optimisation model with monitoring network design methodology for source identification is presented. Performance evaluations of the developed approach for source identification in a heterogeneous, multi-layered aquifer are also discussed.

Identification of unknown groundwater contaminant sources is a complex problem. The complexities arise mainly from uncertainties related to the hydrogeologic information,

sparsity of measurement data and unavoidable concentration measurement errors. The process of contaminant source identification with sparse and limited concentration measurement data, especially when the hydrogeologic parameters are uncertain, requires an efficient procedure. Existing methodologies to tackle this problem in real-world cases usually require huge computational time and the solutions may be non-unique. Therefore, one of the main objectives of this chapter is to evaluate a developed methodology to characterize the groundwater contamination sources in a heterogeneous, multi-layered aquifer. Multivariate Adaptive Regression Splines (MARS) algorithm was utilized to design Surrogate Model-based Optimization (SMO) for source identification. In this SMO, the developed MARS-based surrogate model was also linked to a Genetic Algorithm (GA) based optimisation model.

The other specific, main goal of this study was to develop an efficient procedure to design a monitoring network. It was supposed that by using information from the designed monitoring network to develop the surrogate models, the accuracy of source identification results would improve. In designing the monitoring network, two main objectives were considered: 1. maximising the accuracy of source identification results, and 2. limiting the number of monitoring wells.

## 4.2. An Overview of the Source Identification Problem and Previously Applied Methodologies

Human activities and improper management practices have caused widespread deterioration of groundwater quality worldwide, and have seriously threatened its beneficial use in recent decades. However, when groundwater contamination is detected a long time after the contaminant source(s) became active, often there is not enough information regarding the characteristics of the contamination sources as well as the hydrogeologic parameters of the system. On the other hand, the efficiency and reliability

of contaminant source identification depend on the availability, adequacy and accuracy of hydrogeologic information and contaminant concentration measurements data. For instance, the main disadvantage of previous approaches is that they are highly vulnerable to the accuracy and adequacy of contaminant concentration measurements and hydrogeologic data. A significant number of previously proposed approaches considered that all the hydrogeological parameter values are known. These approaches include the embedded optimisation method (Mahar & Datta, 1997, 2000); and the linked simulation-optimisation method which is the most effective approach to contaminant source identification. In the linked simulation-optimisation approach, different optimisation algorithms were utilized such as the Genetic GA (Jha & Datta, 2013; Singh & Datta, 2006), Simulated Annealing (SA) (Prakash & Datta, 2015) and Adaptive Simulated Annealing (ASA) (Amirabdollahian & Datta, 2015; Jha & Datta, 2013). Only a few previously developed methodologies such as (Amirabdollahian & Datta, 2015; Jha & Datta, 2013) were evaluated under uncertain hydrogeological parameter conditions.

To characterize the unknown characteristics of contaminant sources a new approach was developed and evaluated for potential applicability in practical scenarios. In this approach, MARS algorithm was utilized to design SMO for source identification. The trained surrogate model for source identification approximates the flow and transport simulation models. The developed SMO applied to identify unknown groundwater contaminant sources in terms of contaminant source locations, magnitudes and release history.

However, in this approach and the other methods, the accurate analysis of the process of groundwater flow and transport requires accurate and adequate information on hydrogeologic parameters and contaminant concentration values. The simulation of groundwater flow and solute transport involves intrinsic uncertainties due to the sparsity

or lack of enough hydrogeologic information about the porous medium. For example, hydraulic conductivity plays the main role in the process of groundwater flow and transport and this parameter may be the most uncertain parameter in the groundwater flow and transport models. It is not possible to measure this parameter in every location or discretisation node, where the simulation models of groundwater flow and transport need hydraulic conductivity values. Generally, in real-world cases limited numbers of measured hydraulic conductivity are available. The values of this parameter for other locations are subject to uncertainty and these values need to be estimated.

Therefore, using a proper method to estimate the unknown hydrogeologic parameters based on limited available data is essential in any contaminant source identification strategy. If these estimations do not approximate the hydrogeologic parameters accurately, more errors and uncertainty in the simulation models of groundwater flow and transport will be propagated. Thus, one of the specific objectives of this chapter is to develop an efficient approach for characterising unknown groundwater contaminant sources. This developed approach was evaluated especially where contaminant concentration measurements data was missing for long intervals, and hydraulic conductivity values were only known at limited sample points. As mentioned in the introduction section, the other important goal of this study was to develop an efficient methodology for designing a monitoring network. Random Forests (RF), Classification and Regression Trees (CART), and Tree Net (TN) were the algorithms utilized to develop the monitoring network design approach. These algorithms were selected for their capabilities in recognising the most important components of prediction models.

### 4.3. Methodology

A sequential approach including the developed SMO with monitoring network design approach was utilized for source identification. Some components of the developed

approach such as MARS have been explained in detail in Chapter 3. However, in this chapter, the developed approach and some of its components are also briefly explained.

### 4.3.1. Surrogate Models

Surrogate models or Response Surface Models (RSM) are compact analytical models. These compact models are based on limited numbers of input and output sets obtained from computationally extensive simulation models. If these models are precisely constructed, surrogate models are able to approximate the behaviour of complex systems at reduced computational times (Gorissen et al., 2010). A Surrogate Model-based on Optimisation (SMO) is one of the most practical types of surrogate models that have been utilized to solve nonlinear complex problems. Figure 4.1 presents the schematic chart of the developed SMO in conjunction with a monitoring network design approach for source identification. The main steps in constructing an SMO for source identification in conjunction with the designed monitoring network approach are explained in the following paragraphs (Forrester & Keane, 2009).

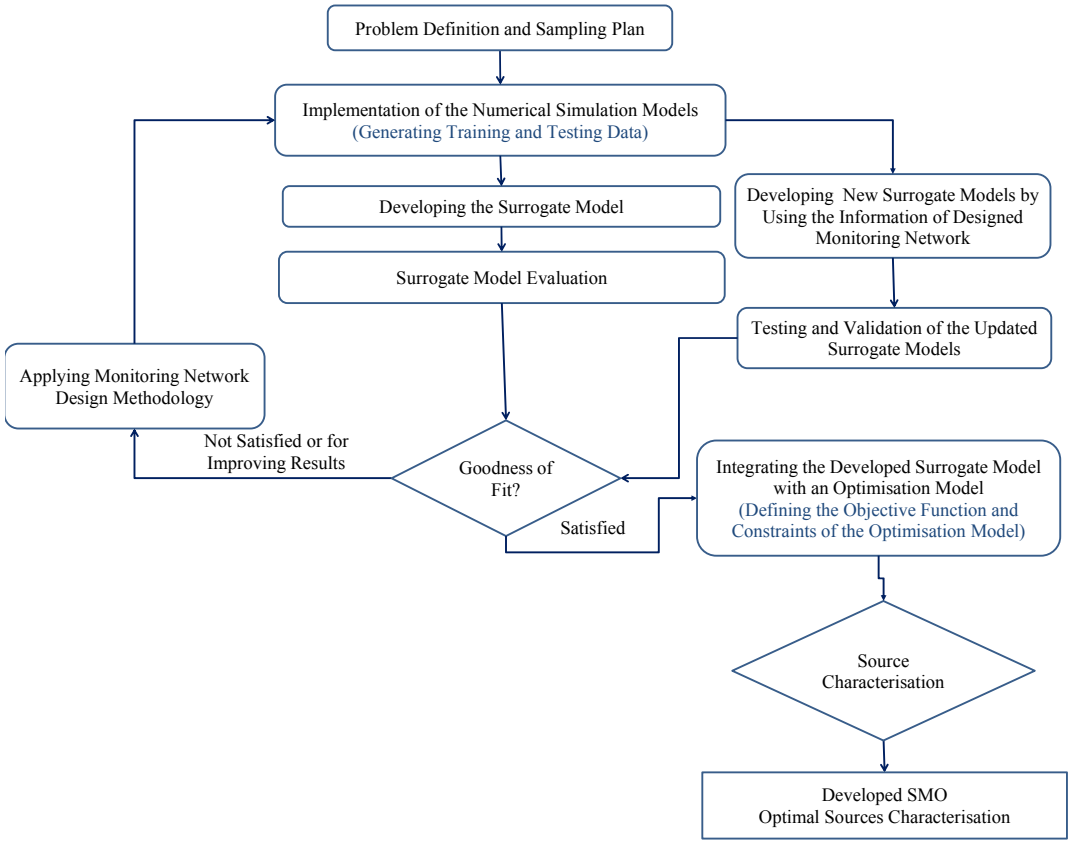



Figure 4.1 Schematic chart of the developed SMO in conjunction with monitoring network design approach for source identification

1. Problem definition and sampling plan: First, the problem and the most important variables of the system which are highly dependent on the complexity of origin system are defined. The important variables of the system must be chosen considering the limitation of surrogate models, which is the number of variables (Razavi et al., 2012). This limitation requires choosing fewer important variables to obtain more accurate results by using the same training data sets. In source identification problems, the characteristics of unknown contaminant sources are considered as parts of the important variables and their numbers are certain. Therefore, the numbers of the remaining important variables of the system must be limited. In source identification problems, these variables are mostly related to

recorded contaminant concentration values at monitoring wells. Therefore, designing a monitoring network and using the information from these locations in developing surrogate models probably could improve the accuracy of source identification results. The Latin Hypercube Sampling (LHS) technique as suggested in (Queipo et al., 2005) was utilized to generate adequate numbers of contaminant source fluxes of the potential contaminant sources.

2. Implementation of the numerical simulation models: Solution results of the simulation models of the groundwater flow and transport for randomly generated contaminant source fluxes in the previous stage were obtained.
3. Developing the surrogate model: The type of surrogate models and the architecture of them should be addressed.
4. Model evaluation: This step assesses the eligibility and predictive accuracy of the developed surrogate model. The results can be utilized in model selection and selection of the model architecture.
5. Developing the SMO: The evaluated surrogate model was integrated into an optimisation model. The objective function and constraints of the optimisation model need to be addressed at this step.
6. Stop/step 3: If the termination criteria are satisfied, stop; otherwise, if the results are not entirely satisfactory, or to improve the results, the monitoring network design methodology could be applied.
7. Updating the developed surrogate models: By using the obtained information from the designed monitoring network and going to step 3, the developed surrogate models could be updated. Then, the updated surrogate models could be utilized for further evaluations.

### 4.3.2. Simulation Models

To solve the flow equation, the numerical simulation model MODFLOW (Harbaugh, 2005) was utilized. MODFLOW uses a three-dimensional equation to represent groundwater flow through porous media (equation(3-1)) (Harbaugh, 2005).

In addition, a Modular Three-Dimensional Multi species Transport Model (MT3DMS) (Zheng & Wang, 1999) was utilized. The MT3DMS uses a partial differential equation (equation (3-2)) to simulate the advection, dispersion, and chemical reaction processes of contaminants to calculate contamination concentration values in groundwater systems (Zheng & Wang, 1999).

### 4.3.3. Designing a Monitoring Network

The quality of contaminant concentration data has a crucial role in the accuracy of source identification results. On the other hand, for solving the source identification problems, only limited and sparse information is usually available. For example, usually, only limited contaminant concentrations are available, which may be collected at different locations. Moreover, the process of collecting data is classified as difficult and expensive. The difficulty is due to the complexity of contaminant movements. Also, collecting data usually is classified as an expensive task because of the necessity for the long duration of collection and the large numbers of contaminated aquifers worldwide. Therefore, designing an effective monitoring network or identifying the monitoring wells that can improve the accuracy of source identification results and subsequently remediation process is essential. Also, the most important limitation of the surrogate models is when their dimensions are large (Razavi et al., 2012). Therefore, the limitations of the surrogate models can be overcome by selecting the most important variables of the system and using the information of these variables (monitoring wells). However, one of the objects is to identify the most important monitoring wells that make the greatest contributions to

the source identification process. Therefore, RF, CART, and TN techniques were applied to identify the most important monitoring wells that can improve source identification results. The Salford Predictive Modeller 8 software was utilized for using the RF, CART, TN and MARS techniques ("Salford Predictive Modeller 8 ", 2017).

#### *4.3.3.1. Random Forests (RF)*

RF is a robust learning machine technique among data mining tools and was introduced by Leo Breiman in 1999. Later, he further developed it with Adele Cutler ("Random Forest for Beginners," 2014). This technique can work with continuous and discrete data ("SPM User Guide, Introduction to Random Forests," 2012). One of the characteristics of RF is that it is effective with small learning data. This tool can also identify the most important or eligible predictors among thousands of potential predictors for predicting a target variable in a prediction model ("SPM User Guide, Introduction to Random Forests," 2012). Once the predictors are identified for each target variable, RF begins to randomly grow decision trees. In each node of a tree, randomly selected predictors are utilized. Then, RF assembles and combines the information of learning trees to generate accurate predictive models. The selection of the most important predictors among the potential predictors is based on the damage that each predictor could do to the prediction models if its values are inaccurate ("SPM User Guide, Introduction to Random Forests," 2012). RF also is capable of self-testing the constructed models by using “out of bag data” by repeating it 100 times ("SPM User Guide, Introduction to Random Forests," 2012). RF capability to recognise the most important variables or effective predictors was utilized to select the most important monitoring wells in source identification among the potential monitoring wells.

#### *4.3.3.2. Classification and Regression Trees (CART)*

The CART method as a data mining tool was introduced in 1984 by Breiman, Freidman, Olshen, Stone (Timofeev, 2004). The CART technique is a robust decision tree tool that is applicable for classification and prediction ("SPM User Guide, Introducing CART," 2013). This algorithm broadly is applicable in different areas such as bioinformatics and risk management ("SPM User Guide, Introducing CART," 2013). One of the capabilities of this algorithm is in identifying the most important variables (predictors) among the potential predictors. This characteristic was utilized to identify the most important monitoring wells that could maximise the accuracy of source identification results.

#### *4.3.3.3. TreeNet (TN)*

TN is one of the advanced technologies in data mining developed by Jerome Friedman ("SPM User Guide, Introducing Tree Net," 2013). TN is fast and easy to use compared to other data mining tools ("SPM User Guide, Introducing Tree Net," 2013). This algorithm also can deal with data that have missing values or even with erroneous data ("SPM User Guide, Introducing Tree Net," 2013). One of the advantages of this algorithm is its capability for ranking the important variables of a prediction model among the potential predictors. This characteristic was utilized to recognise the most important monitoring wells that had the greatest impact on the source identification results.

#### *4.3.3.4. Designing Monitoring Network Procedure*

As mentioned previously, RF, CART, and TN were utilized to design the monitoring network. These algorithms can predict specified target variables by using the information of the predictor variables. These algorithms are also capable of identifying and ranking the predictors based on their contribution to predicting the target variables. These capabilities were utilized to specify the most important monitoring wells among the potential monitoring wells. In this chapter, two main aims were considered in designing

the monitoring network: 1. Maximising the accuracy of source identification; and 2. Limiting the number of monitoring wells. Figure 4.2 presents the schematic diagram of the designing monitoring network procedure using RF, TN and CART. The process of selecting the most important monitoring wells among the potential monitoring wells can be listed as:

1. Defining the important variables of the system: Predictor variables and Target variables of the system should be addressed. In source identification problems, the unknown characteristics of the groundwater contaminant sources are considered as the target variables. Variables related to measured contaminant concentrations are considered to be the predictors.
2. Generating the target variables of the training data of the prediction models: LHS was utilized to randomly generate enough sample sets to train the prediction models. Therefore, LHS was utilized to randomly generate sufficient numbers of target variables, contaminant source fluxes at all potential contaminant sources at all potential activity times.

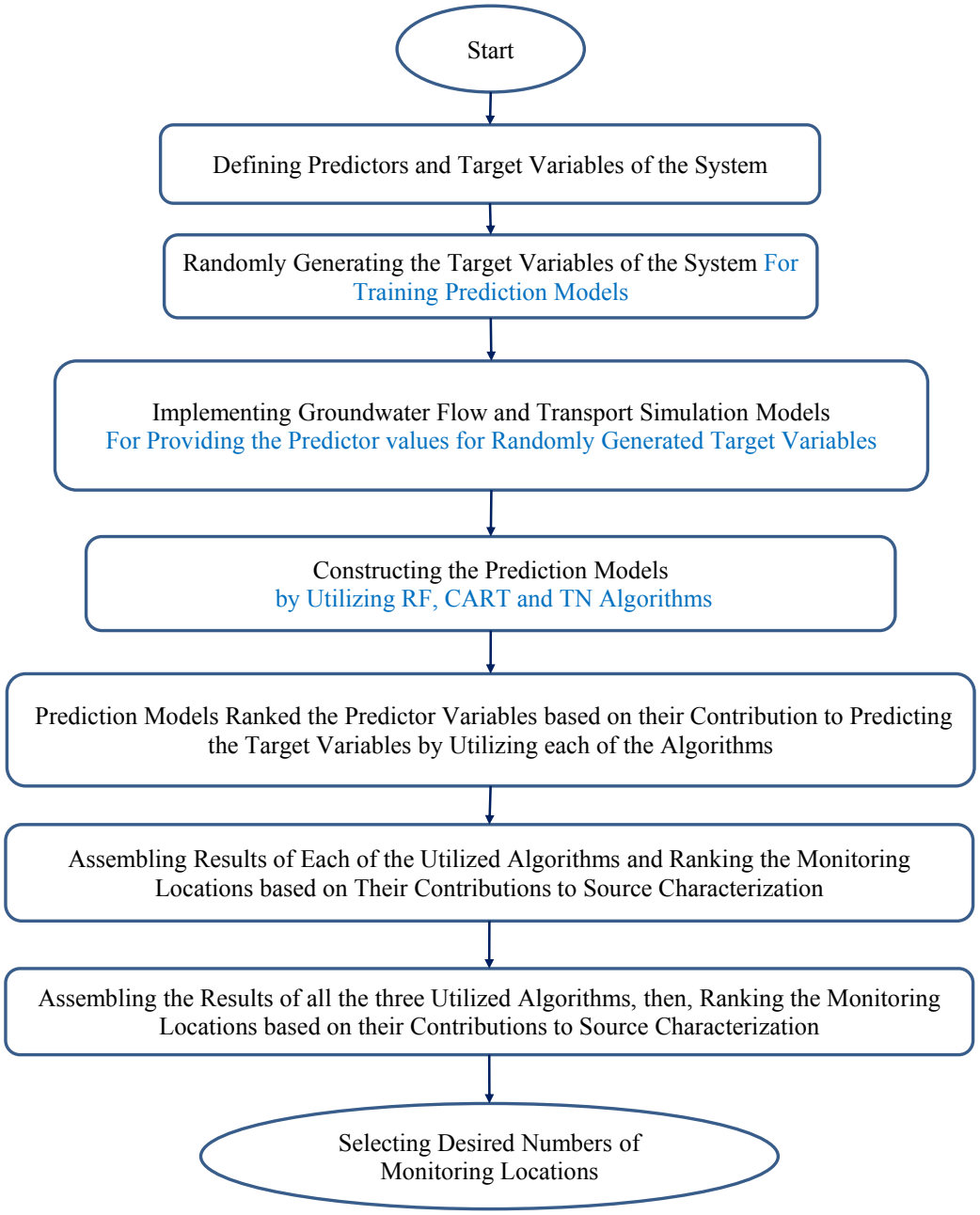



Figure 4.2 Schematic diagram of the applied monitoring network design procedure using RF, TN and CART algorithms

3. Solving the simulation models to generate the predictor variables of the training data: Simulation models of groundwater flow and transport were solved for randomly generated source fluxes. The solution provides corresponding contaminant concentration values at specified potential observation locations at

specified times. These three steps are same as the first three steps of developing surrogate models. So, in this study, for designing a monitoring network, repeating these steps was not necessary.

4. Constructing the prediction models: One separate prediction model needs to be designed and built for each target variable. Table 4.1 shows a typical input for a prediction model by using each of the RF, CART, and TN tools. In this table, the simulated contaminant concentration values at two observation locations at five different times were assumed to be the predictors of prediction models. The contaminant source fluxes at a specific location at a specific time were considered to be a target variable. In this typical input, just 10 sample sets are considered as the training data.

Table 4.1 Typical input vectors using in the RF, TN and CART prediction models

| ID | Target Variable | Predictors | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Source fluxes (g/s) | Contaminant concentration values (g/l) | | | | | | | | | |
| | Source 1 | Monitoring location 1 | | | | | Monitoring location 2 | | | | |
| | Stress period 1 | Time after the start of first source activity | | | | | | | | | |
| | | 3723 | 4015 | 4380 | 4745 | 5110 | 3723 | 4015 | 4380 | 4745 | 5110 |
| 1 | 6.06 | 1.52 | 0.76 | 0.28 | 0.10 | 0.03 | 1.05 | 1.26 | 1.40 | 1.42 | 1.34 |
| 2 | 0.27 | 1.34 | 0.79 | 0.16 | 0.07 | 0.03 | 0.39 | 0.62 | 0.55 | 0.70 | 0.80 |
| 3 | 3.89 | 1.16 | 0.70 | 0.14 | 0.06 | 0.02 | 0.68 | 0.85 | 0.61 | 0.64 | 0.66 |
| 4 | 0.46 | 2.23 | 1.32 | 0.27 | 0.05 | 0.02 | 0.19 | 0.39 | 0.46 | 0.36 | 0.51 |
| 5 | 2.10 | 2.10 | 1.06 | 0.39 | 0.13 | 0.05 | 0.52 | 0.75 | 1.04 | 1.28 | 1.40 |
| 6 | 0.05 | 0.40 | 0.20 | 0.08 | 0.02 | 0.01 | 0.64 | 0.97 | 1.27 | 0.87 | 0.81 |
| 7 | 8.13 | 1.20 | 0.62 | 0.16 | 0.03 | 0.01 | 0.76 | 0.72 | 0.55 | 0.33 | 0.33 |
| 8 | 9.68 | 0.85 | 0.27 | 0.11 | 0.04 | 0.02 | 1.07 | 0.94 | 0.89 | 0.80 | 0.70 |
| 9 | 8.78 | 2.15 | 1.10 | 0.40 | 0.14 | 0.05 | 1.39 | 1.58 | 1.60 | 1.52 | 1.43 |
| 10 | 3.44 | 0.66 | 0.48 | 0.51 | 0.25 | 0.07 | 0.47 | 0.60 | 1.02 | 1.23 | 1.18 |

5. Ranking the predictors based on their contributions to predicting the target variables: All the prediction models developed by using the RF, TN and CART algorithms ranked the predictors. The predictors were ranked by their contributions to predicting the target variables. For each predictor, due to its importance in prediction process, one weight value was assigned. In this case, the predictor variables that represent the recorded contaminant concentration values at specific observation locations at specified times were ranked based on their influence on improving the accuracy of source identification results.
6. Assembling the results of each algorithm for source identification: The results of each utilized algorithm for constructing the prediction models for all the target variables were assembled. Then, for each prediction algorithm, the monitoring wells were ranked according to their contributions in source identification. The monitoring wells that contributed more were assigned larger weights. The weights vary from one to n, in which n is the total number of potential monitoring wells. So, for each potential monitoring location by using the RF, TN and CART algorithms, three weights were allocated. Table 4.2 presents a typical result of using RF, TN and CART to rank 34 monitoring wells based on their influence in source identification.
7. Assembling the results of three utilized algorithms: The assigned weights of the monitoring wells for all three prediction algorithms were added together. As a result, at this stage, one single weight for each of the potential monitoring wells was available. The monitoring wells were then sorted based on their weights and importance in source identification.
8. Selecting the desired number of monitoring wells: Due to the imposed constraint on the maximum permissible number of monitoring wells, the most important

monitoring wells in terms of their respective ranking were selected. It was assumed that, by using the information from the selected monitoring wells, the source identification results could improve.

Table 4.2 Typical results of ranking monitoring wells by using the RF, TN and CART algorithms according to their importance in improving the source identification results

| Ranked by RF | | Ranked by TN | | Ranked by CART | |
|---|---|---|---|---|---|
| Potential monitoring wells | Weight | Potential monitoring wells | Weight | Potential monitoring wells | Weight |
| M12 (12, 40) | 34 | M17 (38, 24) | 34 | M18 (38, 28) | 34 |
| M11 (12, 38) | 33 | M10 (12, 34) | 33 | M10 (12, 34) | 33 |
| M10 (12, 34) | 32 | M12 (12, 40) | 32 | M17 (38, 24) | 32 |
| M9 (12, 30) | 31 | M5 (38, 29) | 31 | M12 (12, 40) | 31 |
| M8 (12, 28) | 30 | M34 (39, 37) | 30 | M5 (38, 29) | 30 |
| M7 (12, 26) | 29 | M33 (36, 42) | 29 | M9 (12, 30) | 29 |
| M20 (12, 43) | 28 | M8 (12, 28) | 28 | M11 (12, 38) | 28 |
| M29 (38, 38) | 27 | M20 (12, 43) | 27 | M29 (38, 38) | 27 |
| M18 (38, 28) | 26 | M11 (12, 38) | 26 | M2 (12,35) | 26 |
| M33 (36, 42) | 25 | M19 (38, 32) | 25 | M8 (12, 28) | 25 |
| M28 (38, 35) | 24 | M18 (38, 28) | 24 | M28 (38, 35) | 24 |
| M30 (38, 41) | 23 | M16 (38, 21) | 23 | M19 (38, 32) | 23 |
| M17 (38, 24) | 22 | M29 (38, 38) | 22 | M34 (39, 37) | 22 |
| M5 (38, 29) | 21 | M26 (13, 46) | 21 | M7 (12, 26) | 21 |
| M19 (38, 32) | 20 | M30 (38, 41) | 20 | M30 (38, 41) | 20 |
| M34 (39, 37) | 19 | M9 (12, 30) | 19 | M15 (38, 19) | 19 |
| M31 (38, 44) | 18 | M31 (38, 44) | 18 | M31 (38, 44) | 18 |
| M16 (38, 21) | 17 | M2 (12,35) | 17 | M16 (38, 21) | 17 |
| M21 (12, 46) | 16 | M7 (12, 26) | 16 | M20 (12, 43) | 16 |
| M26 (13, 46) | 15 | M21 (12, 46) | 15 | M33 (36, 42) | 15 |
| M2 (12,35) | 14 | M35 (39, 45) | 14 | M21 (12, 46) | 14 |
| M15 (38, 19) | 13 | M13 (38, 10) | 13 | M1 (12, 21) | 13 |
| M32 (38, 47) | 12 | M24 (10, 53) | 12 | M4 (38, 16) | 12 |
| M35 (39, 45) | 11 | M28 (38, 35) | 11 | M24 (10, 53) | 11 |
| M22 (12, 49) | 10 | M4 (38, 16) | 10 | M22 (12, 49) | 10 |
| M27 (11, 51) | 9 | M15 (38, 19) | 9 | M14 (38, 17) | 9 |
| M24 (10, 53) | 8 | M27 (11, 51) | 8 | M32 (38, 47) | 8 |
| M23 (12, 52) | 7 | M1 (12, 21) | 7 | M35 (39, 45) | 7 |
| M14 (38, 17) | 6 | M14 (38, 17) | 6 | M13 (38, 10) | 6 |
| M4 (38, 16) | 5 | M22 (12, 49) | 5 | M26 (13, 46) | 5 |
| M1 (12, 21) | 4 | M23 (12, 52) | 4 | M27 (11, 51) | 4 |
| M25 (13, 54) | 3 | M32 (38, 47) | 3 | M23 (12, 52) | 3 |
| M13 (38, 10) | 2 | M25 (13, 54) | 2 | M25 (13, 54) | 2 |
| M3 (26, 28) | 1 | M3 (26, 28) | 1 | M3 (26, 28) | 1 |

### 4.3.4. Optimisation Model

The objective function of the source identification problem can be defined by equation (4-1). This equation is defined to minimise the difference between the estimated and the observed contaminant concentration values at possible observation points at specified times (Mahar & Datta, 1997).

$$Minimize\ E = \sum_{t=1}^{T}\sum_{l=1}^{L}(Cest_l^t - Cobs_l^t)^2. \quad (4\text{-}1)$$

Where:

$Cest_l^t$ and $Cobs_l^t$ are estimated and observed contaminant concentration values at observation well l and at time t, respectively. T and L are the total numbers of concentration observations times and Observation wells, respectively. $w_l^t$ is a weight related to the possible observation point l and time t, this parameter can be defined as (Mahar & Datta, 1997):

$$\boldsymbol{w_l^t = \frac{1}{(Cobs_l^t+\eta)^2}} \quad (4\text{-}2)$$

Where, η is defined as a constant coefficient. This coefficient needs to be large enough to prevent the solution being dominated by errors corresponding to very small measured concentrations (Mahar & Datta, 1997). The developed surrogate model linked to the optimisation model is the main constraint defining the approximate description of the flow and transport processes in the optimisation model.

The main constraints of the optimisation model can be defined as (Prakash & Datta, 2014):

$$Cest_l^t = f(x, y, z, v_x, q_s, C_s, t) \quad (4\text{-}3)$$

Where, $f(x, y, z, v_x, q_s, C_s, t)$ represents the simulation model or the surrogate model linked to the optimisation identification model at time step t.

$x, y, z$ : Cartesian coordinates of the monitoring wells;

$v_x$ : Groundwater velocity along the x coordinate axis ($LT^{-1}$);

$q_s$ : Volumetric flux of water per unit volume of aquifer ($T^{-1}$);

$C_s$ : Concentration of the sources or sinks ($ML^{-3}$); and

$q_s C_s$ : Contaminant source fluxes ($ML^{-3}T^{-1}$).

### 4.3.5. Performance Evaluations of the Developed Procedures

The performance of the developed surrogate model based optimisation was assessed for an illustrative contaminated aquifer study area (Figure 4.3). The performance evaluation was carried out for two different scenarios based on two different assumptions:

1. All the hydrogeologic parameters of the model were precisely known; and
2. Uncertainties were associated with the hydraulic conductivity of the study area, and these parameter values were known only at limited sparse locations.

As for the first assumption, the study area considered was heterogeneous and the actual hydraulic conductivity values were assumed to be random variables. Therefore, to generate hydraulic conductivity throughout the entire study area, the values of hydraulic conductivity (K) were assumed to follow the Lognormal distribution (Freeze, 1975). Thus, it is possible to define a new parameter such as Y = log K, which is normally distributed. Also, the LHS method was utilized to randomly generate the hydraulic conductivity field throughout the study area following the method utilized in (Dokou & Pinder, 2009).

The second assumption implies that the hydraulic conductivity measurements were available only at limited locations, while the simulation models need this parameter

values at all their nodes. Therefore, hydraulic conductivity values should be estimated at other nodes. According to Boman, Molz, and Guven (1995), the Inverse Distance Weighting (IDW) methodology could be the most suitable method to generate hydraulic conductivity because of its simplicity and associated computational ease. This study also demonstrated that the more complicated interpolation methods such as Kriging or fractal-based methods perform little better compared to simplified methods such as the IDW. Also, these two methods are not suitable if measurement data is sparse. Therefore, IDW was utilized to generate hydraulic conductivity values at locations where these values were unknown.

Moreover, Normalised Absolute Error of Estimation (NAEE) and Root Mean Square Error (RMSE) were utilized to quantify the performance evaluation of the developed procedure. The NAEE, which calculates a normalised error of estimation, was represented by equation (3-9) (Jha & Datta, 2013). The RMSE can be defined as:

$$RMSE = \sqrt{\frac{1}{(N \times S)} \left(\sum_{i=1}^{S} \sum_{j=1}^{N} \left(\left(q_i^j\right)_{est} - \left(q_i^j\right)_{act}\right)\right)^2} \quad (4\text{-}4)$$

Where:

S and N are the number(s) of potential contaminant sources and transport stress periods, respectively.

$\left(q_i^j\right)_{act}$ and $\left(q_i^j\right)_{est}$ are actual and estimated source flux at source number i in stress period j.

## 4.4. Application of the Developed Procedure for Source Identification

### 4.4.1. Study Area

Performance of the developed procedures was assessed by using information from an illustrative heterogeneous aquifer site. This aquifer consists of three unconfined layers. The study area is presented in Figure 4.3. The north and south boundaries of this study area are considered no-flow boundaries, while the east and west boundaries are assumed to be specified head boundaries. Only a conservative contaminant and two potential contaminant source locations (CS1 and CS2) are considered. CS1 and CS2 are in layer 1 and layer 2, respectively.

Table 4.3 presents the information of this study area. Table 4.4 also presents the locations and flux magnitudes of the actual contaminant sources. There were five initial arbitrary monitoring wells. The locations of these monitoring wells are presented in Table 4.5. The total time of simulation was separated into five different stress periods (ST1 to ST5). The duration of each of the ST1 to ST4 was two years and the duration of ST5 was 12 years. It was assumed that potential contaminant sources were active only in the ST1 to ST4. It was specified that the contamination was detected just two years after the contaminant sources had stopped their activity. It was also specified that the five monitoring wells were monitored over the last 10 years at limited times. The breakthrough curves at initial arbitrary monitoring wells used for source identification are presented in Figure 4.4.

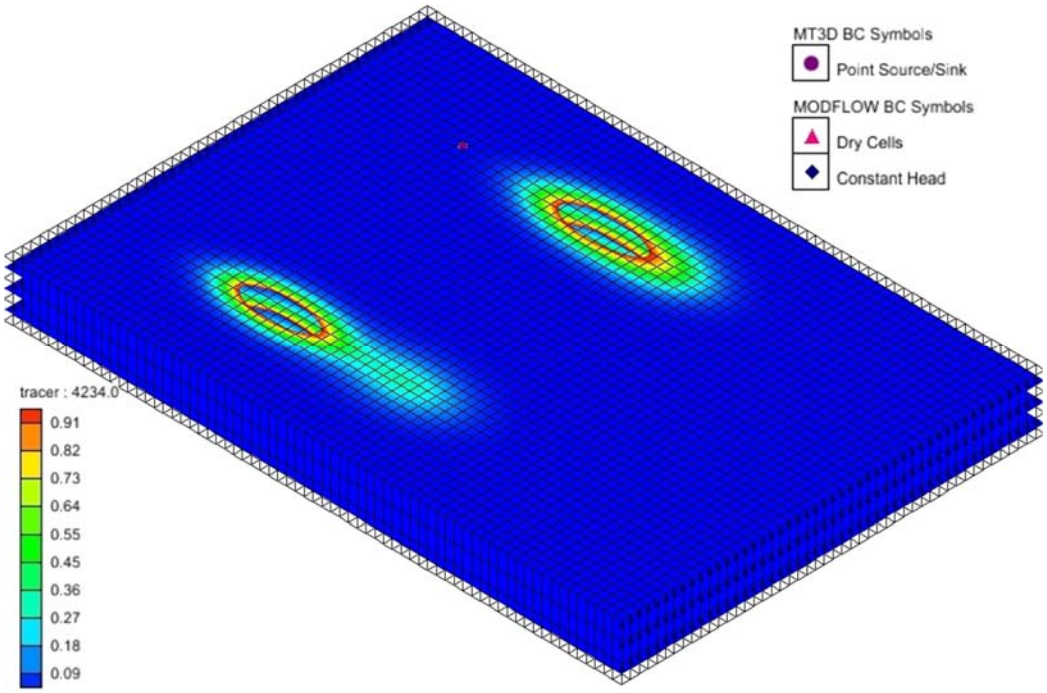



Figure 4.3 Illustrative study area representing typical concentration plumes 4234 days after start of first source activity (concentration values g/l)

Table 4.3 Hydrogeologic parameter values and the dimensions (in metre (m)) of the study area

| Parameter | Unit | Value |
|---|---|---|
| Maximum length | m | 2100 |
| Maximum width | m | 1500 |
| Saturated thickness, b | m | 30 |
| Grid spacing in X and Y-directions | m | 30 |
| Grid spacing in Z-direction | m | 10 |
| Vertical anisotropy | Dimensionless | 5 |
| Hydraulic gradient | Dimensionless | 0.00238 |
| Porosity | Dimensionless | 0.3 |
| Longitudinal Dispersivity | m | 15 |
| Transverse Dispersivity | m | 3 |
| Initial Contaminant Flux | g/s | 0-10 |

Table 4.4 Locations and flux magnitudes of actual contaminant sources

| Potential contaminant source location (row, column, layer) | Contaminant source fluxes (g/s) | | | | |
|---|---|---|---|---|---|
| | ST1 | ST2 | ST3 | ST4 | ST5 |
| CS1 (12, 15, 1) | 6.3 | 4.6 | 9.0 | 5.6 | 0.0 |
| CS2 (38, 9, 2) | 6.7 | 9.3 | 6.1 | 7.3 | 0.0 |

Table 4.5 Locations of monitoring wells

| Monitoring wells | Row | Column | Layer |
|---|---|---|---|
| 1 | 12 | 21 | 1 |
| 2 | 12 | 35 | 1 |
| 3 | 26 | 28 | 1 |
| 4 | 38 | 16 | 1 |
| 5 | 38 | 29 | 1 |

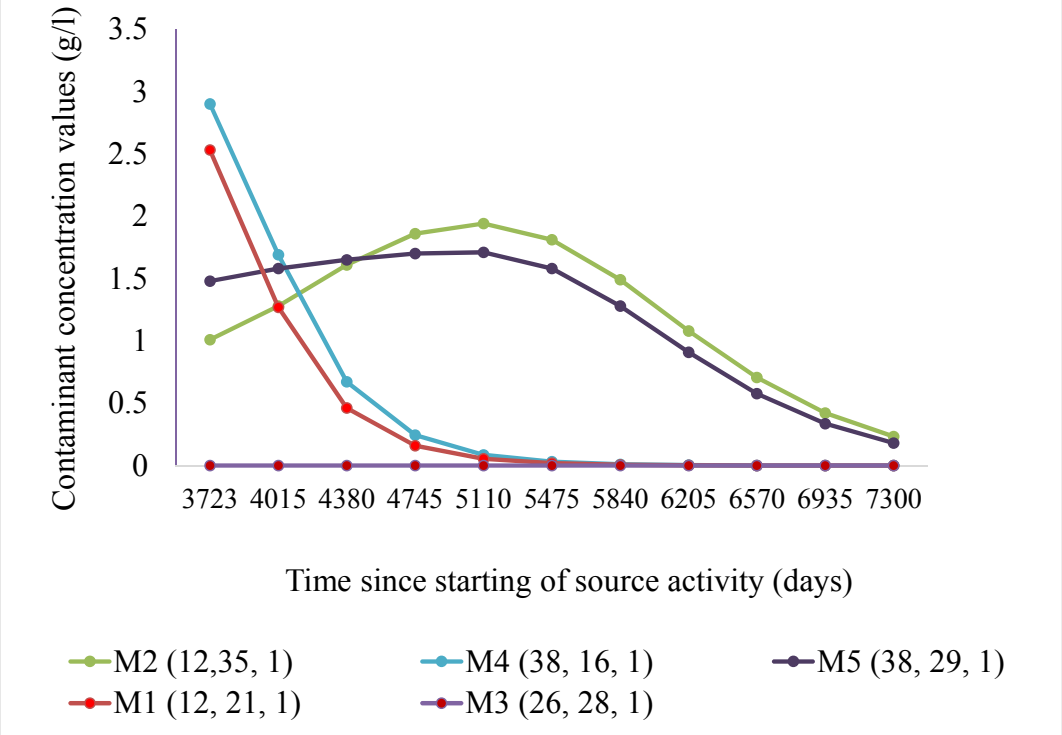



Figure 4.4 Breakthrough curves at initial arbitrary monitoring wells used for source identification

### 4.4.2. Performance Evaluation Results

The following steps were followed to train, test and evaluate the developed models for source identification:

1. Sampling plan: LHS was utilized to generate two groups of 500, and 1000 initial sample sets with two potential contaminant sources with contaminant fluxes in the range of 0-10 g/s.
2. Implementing the simulation models: Simulation models of groundwater flow (MODFLOW) and transport (MT3DMS) (within GMS 7) were solved for two randomly generated groups of source fluxes. The simulation results provided the contaminant concentration values at the five initial arbitrary monitoring wells resulting from these contaminated sources as specified.
3. Developing the surrogate models: The MARS algorithm was utilized to develop MARS based surrogate model. The developed surrogate model represented the relationship between the aquifer stresses in the form of contaminant injection and the resulting impacts in terms of contaminant concentration values at specified monitoring wells at specific times. The randomly generated potential source fluxes and their corresponding contaminant concentration magnitudes at specified monitoring wells at the specified time were utilized as the inputs for training the surrogate models.

For developing the MARS-based SMO, the randomly generated source fluxes at all potential contaminant sources at specific times were considered to be the predictors of the MARS prediction models. Also, the simulated contaminant concentration values at specific times and locations were assumed to be the target variables of the MARS models. Then, the MARS algorithm was utilized to extract BFs of the MARS models based on the non-linear relationships between the predictors and the target variables. In the next

step, all the constructed MARS models for all the target variables were integrated in MATLAB.

4. Evaluation of the developed surrogate model: A group of 100 randomly generated sample sets of potential contaminant source fluxes and corresponding simulated measured contaminant concentrations were used to evaluate the performance of the developed model once it had been adequately trained.

The performance of the developed surrogate model was assessed for source identification by using testing data. The evaluation results are presented in Table 4.6. As can be seen, by using the recorded information at initial arbitrary monitoring wells due to the sparse and missing contaminant concentration data, the accuracy of source identification results is not entirely satisfactory. Therefore, it was assumed that by using information from the designed monitoring wells, the accuracy of source identification results could be improved.

Table 4.6 Performance evaluation results obtained for testing data by using MARS-based surrogate model in terms of RMSE

| ID | Five initial arbitrary monitoring wells |
|---|---|
| | RMSE |
| MARS-based surrogate model | 0.9 |

5. Applying the monitoring network design procedure: To improve source identification results, the RF, TN and CART algorithms, which are powerful prediction techniques, were utilized to rank the potential monitoring wells based on their importance and contribution to the source identification process. The results are summarised and presented in Table 4.7.

Table 4.7 Ranked potential monitoring wells according to their expected influence on source identification by using the RF, CART and TN algorithms

| Monitoring wells | Rank | Cumulative weight |
|---|---|---|
| M10 (12, 34) | 1 | 98 |
| M12 (12, 40) | 2 | 97 |
| M17 (38, 24) | 3 | 88 |
| M11 (12, 38) | 4 | 87 |
| M18 (38, 28) | 5 | 84 |
| M8 (12, 28) | 6 | 83 |
| M5 (38, 29) | 7 | 82 |
| M9 (12, 30) | 8 | 79 |
| M29 (38, 38) | 9 | 76 |
| M20 (12, 43) | 10 | 71 |
| M34 (39, 37) | 11 | 71 |
| M33 (36, 42) | 12 | 69 |
| M19 (38, 32) | 13 | 68 |
| M7 (12, 26) | 14 | 66 |
| M30 (38, 41) | 15 | 63 |
| M28 (38, 35) | 16 | 59 |
| M16 (38, 21) | 17 | 57 |
| M2 (12,35) | 18 | 57 |
| M31 (38, 44) | 19 | 54 |
| M21 (12, 46) | 20 | 45 |
| M15 (38, 19) | 21 | 41 |
| M26 (13, 46) | 22 | 41 |
| M35 (39, 45) | 23 | 32 |
| M24 (10, 53) | 24 | 31 |
| M4 (38, 16) | 25 | 27 |
| M22 (12, 49) | 26 | 25 |
| M1 (12, 21) | 27 | 24 |
| M32 (38, 47) | 28 | 23 |
| M13 (38, 10) | 29 | 21 |
| M14 (38, 17) | 30 | 21 |
| M27 (11, 51) | 31 | 21 |
| M23 (12, 52) | 32 | 14 |
| M25 (13, 54) | 33 | 5 |
| M3 (26, 28) | 34 | 3 |

The top five monitoring wells among the 34 potential monitoring wells were selected for further performance evaluations. These five monitoring wells were selected to have the same numbers of locations as the initial arbitrary monitoring wells.

6. Updating the MARS-based surrogate model: The simulated contaminant concentration values at these five selected monitoring wells were utilized to construct new surrogate models. The contaminant concentration values for these selected monitoring wells were assumed to be available at the same times as the five initial arbitrary monitoring wells. The breakthrough curves at the selected monitoring wells used for source identification are presented in Figure 4.5.

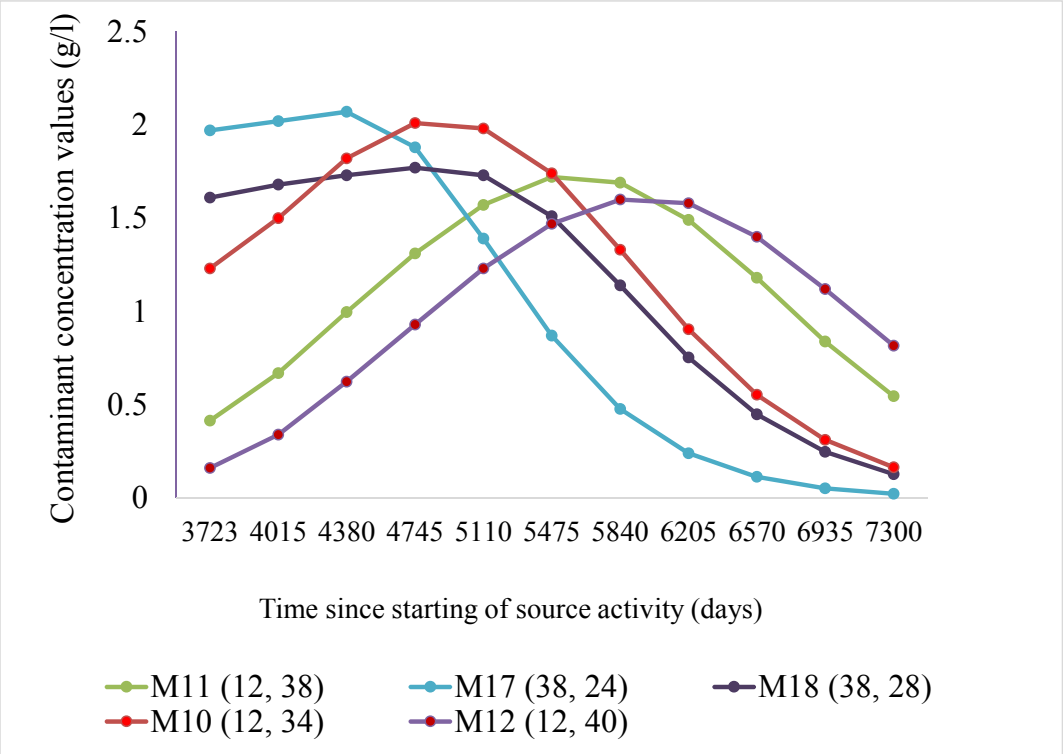



Figure 4.5 Breakthrough curves at selected monitoring wells used for source identification

The performance of the developed new surrogate models was assessed by using the same testing data (100 sample sets). The results for testing data are listed and compared in Table 4.8. The results demonstrate significant improvements by using information from the selected monitoring wells. For example, the accuracy of the MARS-based surrogate models results by using information from the selected monitoring wells for testing data improved by 0.7 in terms of RMSE.

Table 4.8 Performance evaluation results of the developed surrogate models for testing data in terms of RMSE

| ID | Five initial arbitrary monitoring wells | Five selected monitoring wells |
|---|---|---|
| | RMSE | RMSE |
| MARS-based surrogate model | 0.9 | 0.2 |

7. Integrating the developed surrogate model with an optimization model: The developed MARS-based surrogate model was linked to a GA-based optimisation model for source identification. In this MARS-based SMO, the boundaries of the surrogate model were addressed as one of the optimisation model constraints. The main objective function of the source identification problem (equation (4-1)) was also defined at this stage as the objective function of the optimisation model.
8. Unknown contaminant source identification: The corresponding simulated contaminant concentration values of the actual contaminant source fluxes were utilized for source identification as an inverse problem. The results of the MARS-based SMO are presented in Figure 4.6. The source identification results from using MARS-based SMO when using the information from initial arbitrary and selected monitoring wells in terms of NAEE were equal to 8.1 and 5.3%, respectively. This shows the utility of the designed monitoring network in improving the source identification results at least in this limited illustrative problem.

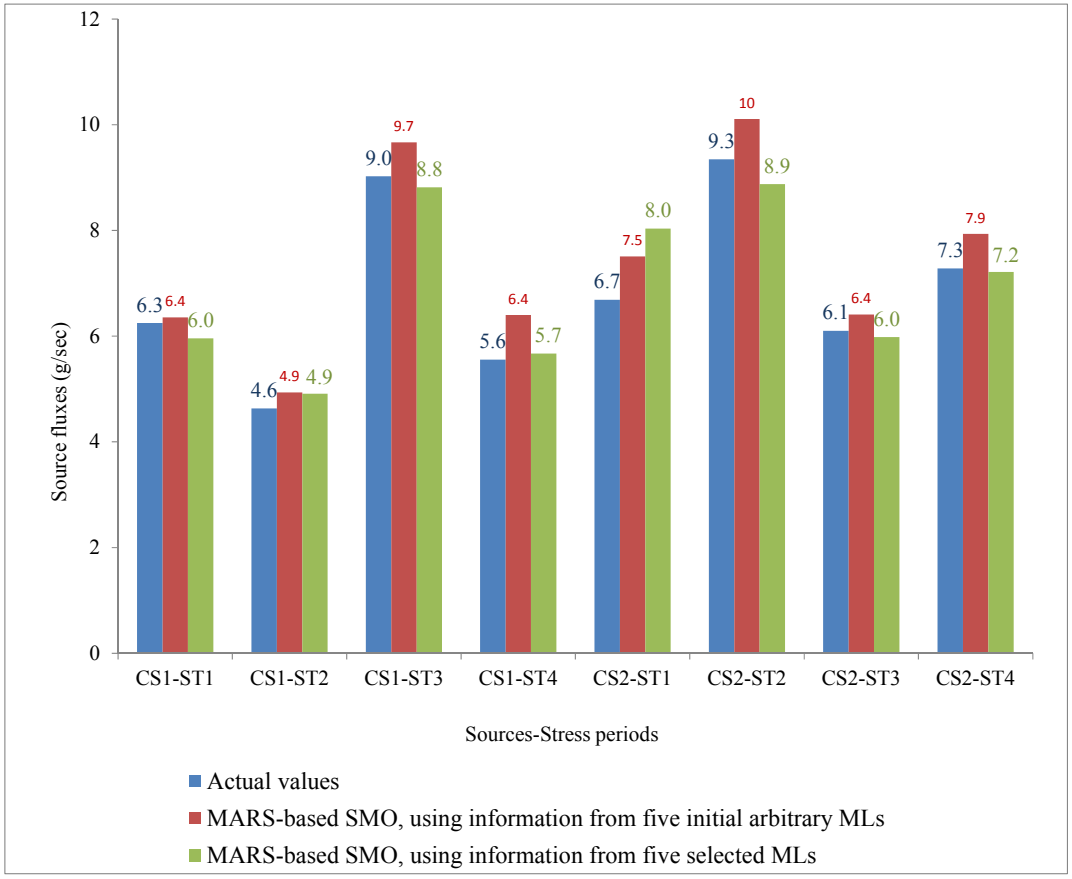



Figure 4.6 Comparison of the obtained results for source identification by utilizing MARS-based SMO and using the information from initial arbitrary and selected monitoring wells with actual data.

Performance of the developed methodology was also assessed for source identification when hydraulic conductivity values were not available in the entire study area. When using the first assumption, the hydraulic conductivity field for the whole study area was generated by assuming that the means of hydraulic conductivity in each of the three layers (layer 1, 2 and 3) were 20, 17, and 21 m/day and the standard deviations were 0.1, 0.08, and 0.12, respectively. In the second assumption, the hydraulic conductivity measurements were available only at 20 locations. The distances between any two locations along the maximum length and minimum length of the study area were 300 and 450 metres, respectively. Therefore, to generate hydraulic conductivity values at other locations, the IDW methodology was utilized as the interpolation method, due to its

efficiency and simplicity (Boman et al., 1995). Figure 4.7 represents the generated hydraulic conductivity field for layer 1 using the IWD interpolation method.

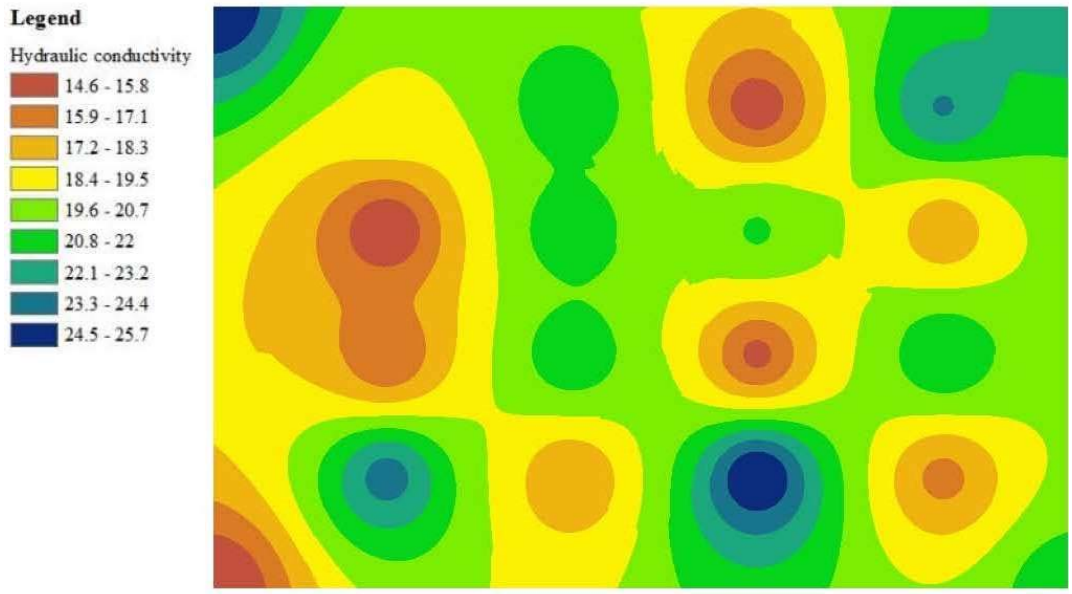



Figure 4.7 Generated hydraulic conductivity for layer 1

Hydraulic conductivity measurements based on the second assumption were utilized to simulate the groundwater flow and transport simulation models. The randomly generated source fluxes were then utilized in the updated groundwater flow and transport simulation models to generate corresponding contaminant concentration values. The contaminant concentration values at the initial monitoring wells at the specified time were utilized to construct the MARS-based SMO. The source identification results in terms of NAEE were equal to 19.9%.

## 4.5. CONCLUSION

The MARS algorithm was utilized to develop MARS-based SMO. The performance of the developed surrogate model was evaluated in a heterogeneous, multi-layered aquifer. The contaminant concentration values of this contaminated aquifer were assumed

missing for a long time interval. The performance of the developed methodology was also assessed by considering two scenarios representing two assumptions. First, the hydrogeologic parameters, i.e., hydraulic conductivity, were assumed to be known. Second, hydraulic conductivity values were uncertain and it was assumed that measurement values were known only at 20 locations. The performance evaluation results indicate that the MARS-based surrogate model could approximate simulation models of groundwater flow and transport adequately. These results also show that the developed MARS-based SMO could characterize unknown groundwater contaminant sources in terms of contaminant source locations, magnitude and release history.

The source identification results obtained by using data from initial arbitrary monitoring wells were not entirely satisfactory. However, a comparison of the source estimates and the actual source characteristics show a good match. Therefore, for improving the accuracy of the solution results, the monitoring network design procedure was applied. The RF, TN and CART algorithms were utilized to identify the most important monitoring wells among the potential locations for source identification. In designing the monitoring network design procedure, two objectives were considered. These objectives were: 1. Maximise the accuracy of source identification results; and 2. Limit the number of monitoring wells. Information from the designed monitoring network was utilized to develop new surrogate models. The source identification results for testing data show improvements by using the information from the designed monitoring network (Table 4.8). However, using data from the designed monitoring network could be more effective if data from more potential monitoring wells were used in the designing monitoring network process. The evaluation results show potential applicability of the developed procedures for contaminant source identification and monitoring network design.

The evaluation results presented in this chapter and Chapter 3 are based on very limited scenarios and therefore restricted in scope. Further performance evaluations are required to fully establish the applicability of the developed methodologies. Therefore, the next chapter discusses the evaluation results of the developed surrogate models in an experimental contaminant aquifer site.

# 5. Verification of the Developed Procedures for Source Identification by using data from an Experimental Aquifer Site

## 5.1. Introduction

Some contents of this chapter have been released in the following journal paper:

- Hazrati-Yadkoori, S., & Datta, B. (2017). Evaluation of Unknown Groundwater Contaminant Sources identification Efficiency under Hydrogeologic Uncertainty in an Experimental Aquifer Site by Utilizing Surrogate Models. Journal of Water Resource and Protection, 9, 22. doi:10.4236/jwarp.2017.913101

In this chapter, the application of the developed procedures for source identification to an experimental contaminated aquifer site is discussed. This experimental site was within the heterogeneous sand aquifer, located at the Botany Basin, New South Wales, Australia. The hydrogeologic characteristics of this experimental site were investigated through several tests (Beck, 2000). As a result, limited numbers of hydraulic conductivity and contaminant concentration measurement values were available. The measured contaminant concentration values and hydraulic conductivity values were not error free. Therefore, the main goal of this chapter was evaluating the performance of the developed surrogate models in an experimental contaminated aquifer site with data that were not error free.

In this chapter, first, the developed methodologies presented in Chapter 3 are briefly discussed. Next, the contaminated experimental site and its history are presented. Then, the results of using the developed surrogate models in the contaminated experimental site are discussed. Finally, the main conclusions of this study are explained.

## 5.2.Methodology

Source identification is an important but difficult step in effective groundwater management. The difficulties arise mainly due to the time of contaminant detection which usually happens long after the start of contaminant source(s) activities. As a result, usually limited information is available which also can be erroneous. Therefore, successful contaminant source identification and subsequently remediation process need the use of an efficient methodology.

Different surrogate models for comparison purpose by utilizing Self-Organising Map (SOM), Multivariate Adaptive Regression Splines (MARS), and Gaussian Process Regression (GPR) algorithms were developed. These surrogate models can approximate the complex groundwater flow and solute transport processes in contaminated aquifer sites. Simulated responses of the aquifer to randomly specified contamination stresses as simulated by using three-dimensional numerical simulations models were used for initial training of these surrogate models. The important feature of these developed surrogate models is that unlike previous methods, this source identification methodologies can be applied independently of any linked optimisation model solution.

### 5.2.1. Surrogate Models

Generally, implementation of simulation models for real-world cases is complex and extensively time-consuming. Therefore, to decrease the high computational cost of the complex simulation models, these computationally intensive simulation models have been replaced by response surface methodologies. It is supposed that by accurately constructing these models, the behaviour of more sophisticated simulation models can be approximately emulated with much reduced computational time (Gorissen et al., 2010).

As mentioned earlier in this chapter, for source identification, SOM, MARS, and GPR algorithms were utilized to construct the surrogate models (Figure 5.1). These models

mimic the behaviour of simulation models of the groundwater flow and transport, MODFLOW and MT3DMS, respectively. Also, the developed surrogate models were applied to identify unknown contaminant sources in terms of contaminant source locations, magnitudes and activity times. The main steps involved in developing a surrogate model for source identification are illustrated in Figure 5.1. These steps are also explained as follows:

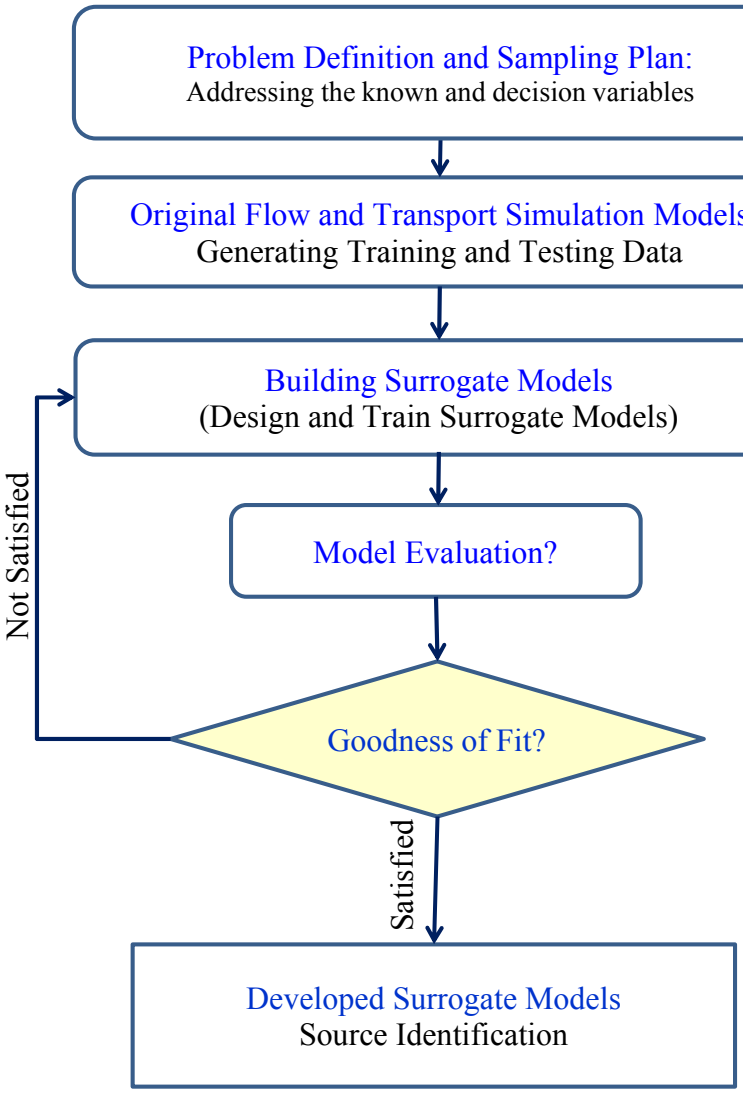



Figure 5.1. Flow chart of the main steps of developing surrogate models for source identification

1. Problem definition and sampling plan: This is a crucial step and has essential effects on the accuracy of results. First, the problem and the most important variables of the system which are highly dependent on the complexity of origin system are defined. These variables are constituted of known variables and decision variables. Then, for generating qualified sampling points for training and testing surrogate models a suitable random generating methodology needs to be selected and utilized. The sampling size suggested is 15-20 times the dimensions of the problem (Wang et al., 2014).
2. Implementing the simulation models: Simulation models of groundwater flow and transport for the contaminated aquifer site need to be solved. These models are solved to randomly generated source fluxes at the previous stage. As a result, the contaminant concentration values are obtained as the solution of these numerical simulation models.
3. Building surrogate models: At least one important question should be addressed, the tool(s) which are to be utilized for constructing the surrogate model(s) (Queipo et al., 2005). The design or architecture of the surrogate model also can be addressed in this step.
4. Model evaluation: Evaluating the performance of the developed surrogate models by using a new sample dataset which independent of the training data. The model results can be utilized to change the surrogate model type or its architecture.
5. Source identification/step 3: If the goodness of fit is achieved, the solution is obtained and stop. Otherwise, go to step 3.

### 5.2.2. Simulation Models

MODFLOW (Harbaugh, 2005) and MT3DMS (Zheng & Wang, 1999) were numerical simulation codes of groundwater flow and transport utilized in this study. The governing

equations of MODFLOW and MT3DMS are presented as equations 3-1 and 3-2, respectively. MODFLOW is a finite-difference based groundwater flow model. This model is utilized for numerical flow simulations (Harbaugh, 2005).

The MT3DMS is the numerical mass transport simulation model. MT3DMS can simulate the advection, dispersion, and chemical reaction processes of groundwater contaminants (Zheng & Wang, 1999).

## 5.3. Application of the Developed Procedures for Source Identification

### 5.3.1. Study Area

The performance of the developed methodology was assessed by using information from a natural gradient tracer test carried out at an experimental site. This experimental site is known as East Lake Experimental Site (ELE site) in Botany Basin, New South Wales, Australia (Jankowski & Beck, 2010). The Botany basin is south of the Sydney CBD and has been utilized as a water supply for Sydney since European colonisation (Beck, 2000). Figure 5.2 illustrates Botany Sands aquifer in Australia. In this region, commercial and industrial developments started in the early 20th century in the northern parts of Botany Bay. As a result of industrial and residential developments, the consumption of groundwater has been increased significantly especially for industrial purposes. The main infrastructures in this region are an airport, oil storage, refinery, and storage facilities and several other chemicals, industrial, and commercial manufacturing plants mainly in the northern parts (Beck, 2000). As the results of these extensive uses of the region's land, Botany Sands aquifer has had a long history of contamination. In this region, also because of the excessive pumping in some parts of the basin, the groundwater level has declined and this caused some alterations in the flow regime and spreading of some types of contaminants (Beck, 2000).

Despite the existence of some locally confined parts in this aquifer under clay, pet lenses and bands, the Botany Sands aquifer is mainly unconfined. The aquifer thickness varies from less than a few metres to more than 75 metres and its average thickness is estimated to be 15 metres (Beck, 2000). The groundwater level in the Botany Sands aquifer varies from 0-9 metres below the ground level. The main source of Botany Sands aquifer recharge is rainfall infiltration. The average annual rainfall of Botany basin within the recorded data varies from 750 mm to 1350 mm (Beck, 2000).

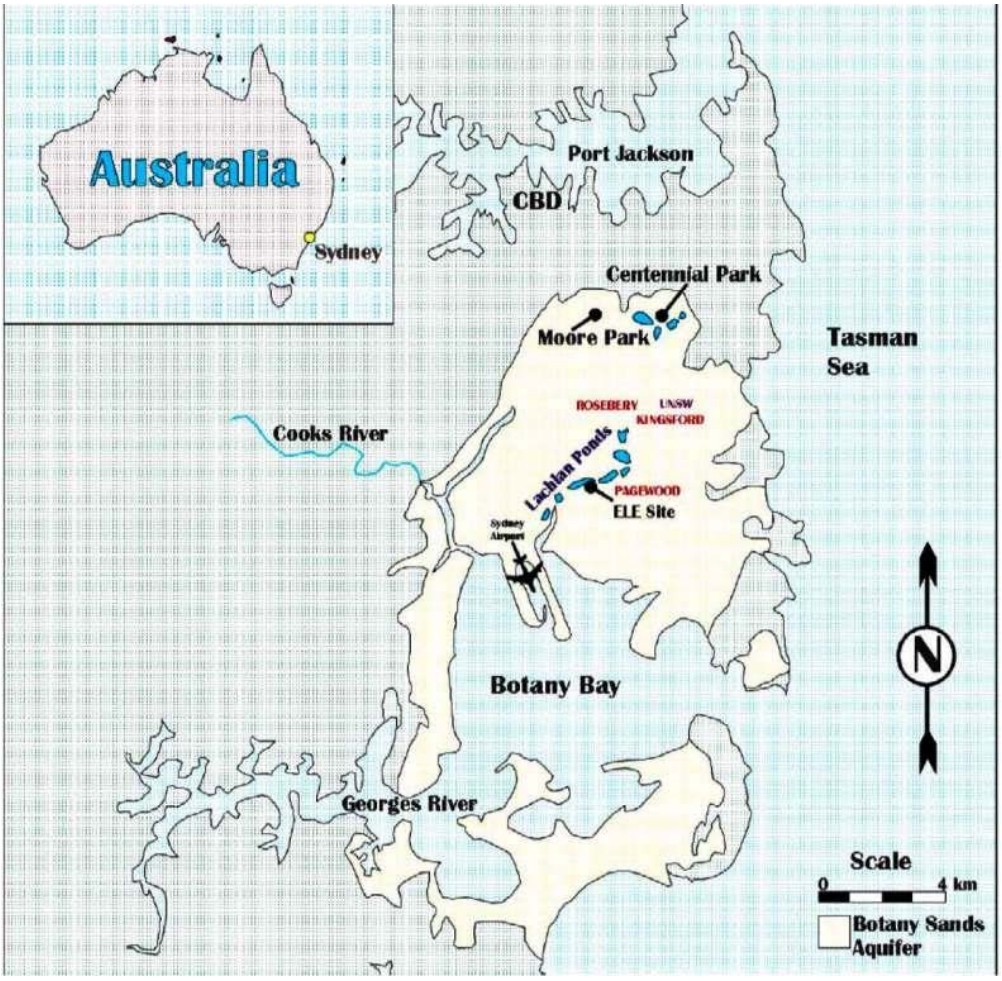



Figure 5.2. The East Lake Experimental Site location (ELE site) at the Botany Sands aquifer (Beck, 2000)

The Botany Sands aquifer– due to its important role as the source of industrial and agricultural water– has been investigated through different studies since the 1940s (Beck, 2000). Groundwater contamination in this aquifer also has been reported since that time. High concentrations of bromide, chloride, nitrate, sulfate, sodium, and calcium have been reported in different studies (Beck, 2000). Recently, various contamination sources such as leakage from sewer lines, landfills, underground storage tanks, urban run-off, and industrial and residential land use have been identified. Since the identification of contamination in this aquifer, different management and remediation strategies have been applied to control and decrease the negative effects of these contaminants (Beck, 2000).

### 5.3.2. Site Description, Eastlake Experimental Site

The ELE site was founded in 1992 for research studies at the University of New South Wales (UNSW) Groundwater Centre (Beck, 2000). Figure 5.3 shows this site in the area which is adjacent to the Lachlan Ponds. Figure 5.4 shows the most important features of the ELE site. This site is located in the upper part of the Botany Sands aquifer next to Pond 5 of Lachlan Ponds in an area about $80m^2$ (Amir Abdollahian, 2016; Beck, 2000). Although this aquifer is homogeneous and isotropic on a macroscopic scale, it is heterogeneous and anisotropic on a microscopic scale (Jankowski & Beck, 2010).

According to the results of previous geological investigations, the experimental site consists of five sedimentological distinct layers (Figure 5.5): 1. Medium sand with silt/clay content of up to 5%; 2. Waterloo Rock; 3. Organic silty sand; 4. Peat material; and 5. Silty/clay sand unit (Jankowski & Beck, 2010). The differences in the grain-size distribution of different soil types at the ELE site and deposition environments suggest some variations in hydraulic conductivity distributions of the ELE site (Amir Abdollahian, 2016).


Figure 5.3. The ELE site adjacent to the Lachlan Ponds (Beck, 2000)

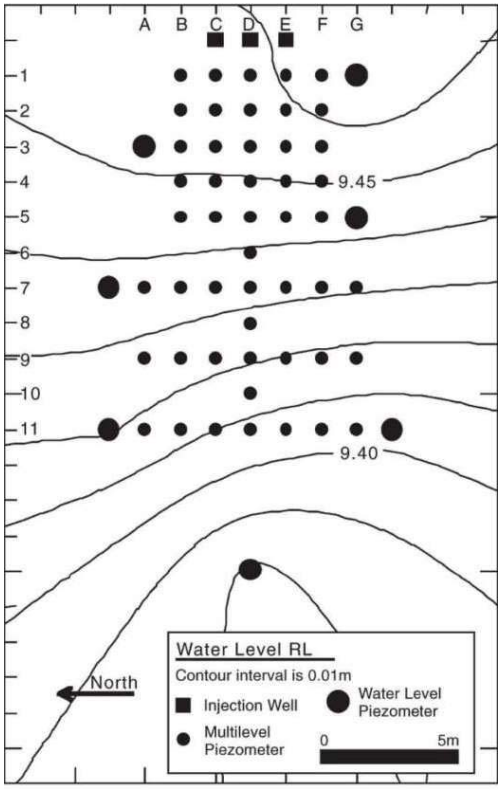



Figure 5.4. Layout of ELE site showing injection well locations, multilevel piezometers and water level piezometers (Jankowski & Beck, 2010)

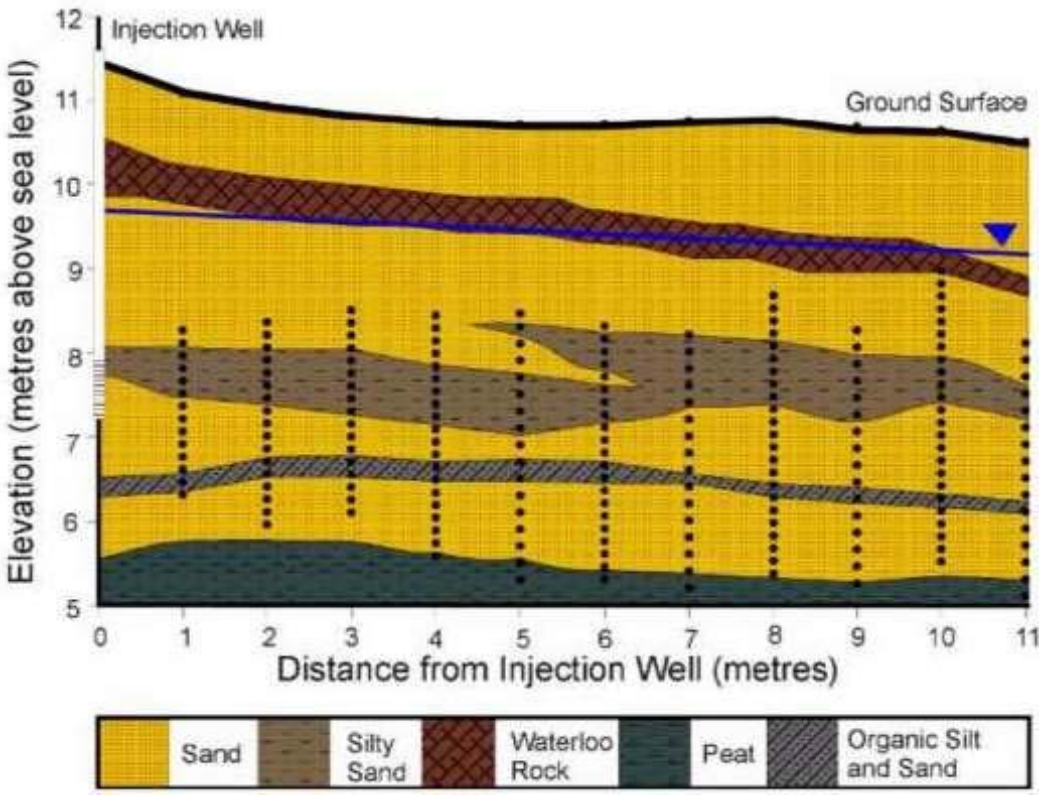



Figure 5.5 Geological cross-section of the ELE site along line D (Beck, 2000)

### 5.3.3. Tracer Test and Movement of a Conservative Element

In the tests carried out in the ELE site in July 1996, the injected tracer solutions included conservative and reactive inorganic elements such as bromide, calcium, lead, and potassium. Three injection wells, C, D, and E, were utilized in this test. These wells are illustrated in Figure 5.4. The tracer test was conducted by preparing 300 litres of a solution that included boron, bromide, chloride, and lithium as conservative tracers and six reactive solutes (Beck, 2000). The concentrations of conservative tracers needed to be three to four times higher than background concentrations to be properly monitored. To analyse the background chemical concentrations of tested elements, 88 groundwater samples were collected. The analysis results indicated that all of the tested elements' concentrations were below the analytical detection limit (Beck, 2000). The detection limit concentration for bromide was 1.8 mg/l (Beck, 2000). Bromide was considered as a

conservative contaminant. The concentration of bromide in the test was 186 mg/l. The containers of tracer solution were injected over 30 minutes, from 13:00 to 13:30 on 2nd July 1996. During the tracer injection, the flow rates of wells were kept low enough to avoid significant increases in the hydraulic heads at the injection wells (Beck, 2000).

The first samples of contaminant concentrations were collected two days after the injection, on 4th July 1996. Gathering samplings were repeated by nine more sessions 4, 6, 8, 12, 16, 20, 24, 28 and 32 days after injection. Monitoring transport of the tracers plume movements demonstrated that bromide and the other conservative element transports were mainly controlled by the variability of the aquifer's hydraulic conductivity (Beck, 2000). According to the previous studies at the ELE Site, for bromide, monitoring values until 16 days after the injection showed no noticeable chemical transport processes to affect the natural tracer behaviours (Beck, 2000). Advection and dispersion were the dominant physical processes of the bromide tracer transport during the monitoring time.

### 5.3.4. Simulation Models

The simulation models of groundwater flow and solute transport of the ELE site based on the information obtained from Beck (2000); Jankowski and Beck (2010) were developed. The ELE site extended from 6 metres above sea level to the groundwater level and could be divided into four distinct layers. The thickness of the top layer which extends from the top of the silty sand layer to the groundwater level is 1.5 metres. This layer is comprised mainly of sand. The second layer has 0.4-metre depth and it is mainly comprised of silty sand. The third layer with injection wells located in it has 0.6-metre depth. This layer is mainly comprised of sand. The thickness of the bottom layer is 1 metre and it is situated on the top of peat layer (Amir Abdollahian, 2016).

A network of 49 piezometers was installed in a 7×11 metres area in this part of the aquifer on a 1 metre × 1-metre grid (Beck, 2000; Jankowski & Beck, 2010). These piezometers penetrated up to 6 metres into the underlying sediments to investigate geological and hydrogeological characteristics of this experimental site. The dimensions and characteristic values of the ELE site are presented in Table 5.1 (Amir Abdollahian, 2016). This information was obtained from the previous studies reports at this experimental site (Beck, 2000; Jankowski & Beck, 2010).

Table 5.1 Hydrogeological information of the experimental study area

| Parameter | Unit | Value |
|---|---|---|
| Maximum length | Metre (m) | 15.00 |
| Maximum width o | m | 13.00 |
| Thickness of study area | m | 3.50 |
| Grid spacing in x-direction | m | 1.00 |
| Grid spacing in y-direction | m | 1.00 |
| Porosity (layer1, layer2, layer3 and layer 4) | Dimensionless | (0.39, 0.41, 0.36 and 0.41) |
| Longitudinal dispersivity (all layers) | m | 0.03 |
| Ratio: H/L dispersivity | Dimensionless | 0.10 |
| Specific storage (all layers) | 1/m | 0.20 |
| Specific Yield (all layers) | Dimensionless | 0.20 |
| Recharge | m/day | 0.00 |
| Flow rate in injection wells | $m^3$/day | 4.40 |
| Initial bromide injection concentrations | mg/l | 0-300 |

The ELE site is an unconfined aquifer. The east and west boundaries of ELE site were considered as specified head boundaries, due to the location of this site on the side of the Pond 5 of Lachlan Ponds that provides hydraulic continuity with the pond (Figure 5.2). The north and south boundaries were considered to be variable heads. The initially specified head distributions were based on the specified contours in Figure 5.4. As mentioned earlier, rainfall is the main source of recharge for the Botany Sands aquifer.

The total time of simulation was divided into five different stress periods. The first stress period was the only active stress period and its duration was 30 minutes. The second to fourth stress periods were each of two days duration and the last stress period was of eight days duration. The monitored contaminant concentrations at nine monitoring locations and totalling to 10 values, and belonging to stress periods two to five were utilized for source identification as presented in Table 5.2.

Table 5.2 The monitoring locations and observed concentration values

| ID | Monitoring locations (i, j, k) | Stress Period | Contaminant concentration values (mg/l) |
|---|---|---|---|
| 1 | M1 (7,3,3) | | 12.20 |
| 2 | M2 (6,3,3) | 2 | 15.50 |
| 3 | M3 (5,3,3) | | 0.10 |
| 4 | M4 (8,3,3) | 3 | 9.00 |
| 5 | M2 (6,3,3) | | 19.00 |
| 6 | M5 (5,4,3) | 4 | 0.09 |
| 7 | M6 (6,5,3) | | 0.09 |
| 8 | M7 (8,4,3) | | 0.15 |
| 9 | M8 (6,4,3) | 5 | 13.30 |
| 10 | M9 (7,6,3) | | 0.11 |

*: (i, j, k) the nodes coordinates in X, Y and Z directions, respectively.

In addition to the three injection sources, one more potential contaminant location was considered as a possible contaminant source location to assess the performance of the developed procedures for source identification. The flow rate of this additional potential contaminant source was considered to be $1 m^3$/day to prevent a significant change of the flow system and hydraulic head distribution (Amir Abdollahian, 2016). The monitored contaminant concentration values (Table 5.2) were utilized in this study to recover injected bromide concentrations.

The hydraulic conductivity values for ELE site were estimated by applying a combination of constant head tests and falling head tests (Beck, 2000). A total of 522 hydraulic conductivity values along the three lines shown C, D and E were available. The distributions of hydraulic conductivity showed considerable variations from 1.8-50m/day. Sometimes these variations were observed in short distances (Beck, 2000; Jankowski & Beck, 2010). According to the results of the previous studies, the mean hydraulic conductivity value for Botany Sands aquifer was likely around 20m/day (Beck, 2000). The simulation of groundwater flow and transport of ELE site needs the hydraulic conductivity values be known throughout the entire study area. Therefore, due to unavailability of the hydraulic conductivity values at all discretisation nodes; 240 hydraulic conductivity values (some of these were multiple measurements within the same layer) were utilized to generate interpolated hydraulic conductivity values for all nodes of the study area. The Inverse Distance Weighting (IDW) methodology was utilized to interpolate hydraulic conductivity values for the entire study area because of its simplicity and efficiency (Boman et al., 1995). The 240 hydraulic conductivity values were utilized in three different iterations to interpolate hydraulic conductivity values through the whole study area. As mentioned earlier, in some cases, for a certain location different measured hydraulic conductivity values were available. Therefore, IDW was utilized to interpolate hydraulic conductivity values throughout the whole study area in three different iterations. The average values of these three iterations for all nodes of the study area were utilized as the inputs of simulation models. Figure 5.6 represents the generated hydraulic conductivity values for layer three of the ELE aquifer using IWD interpolation method.

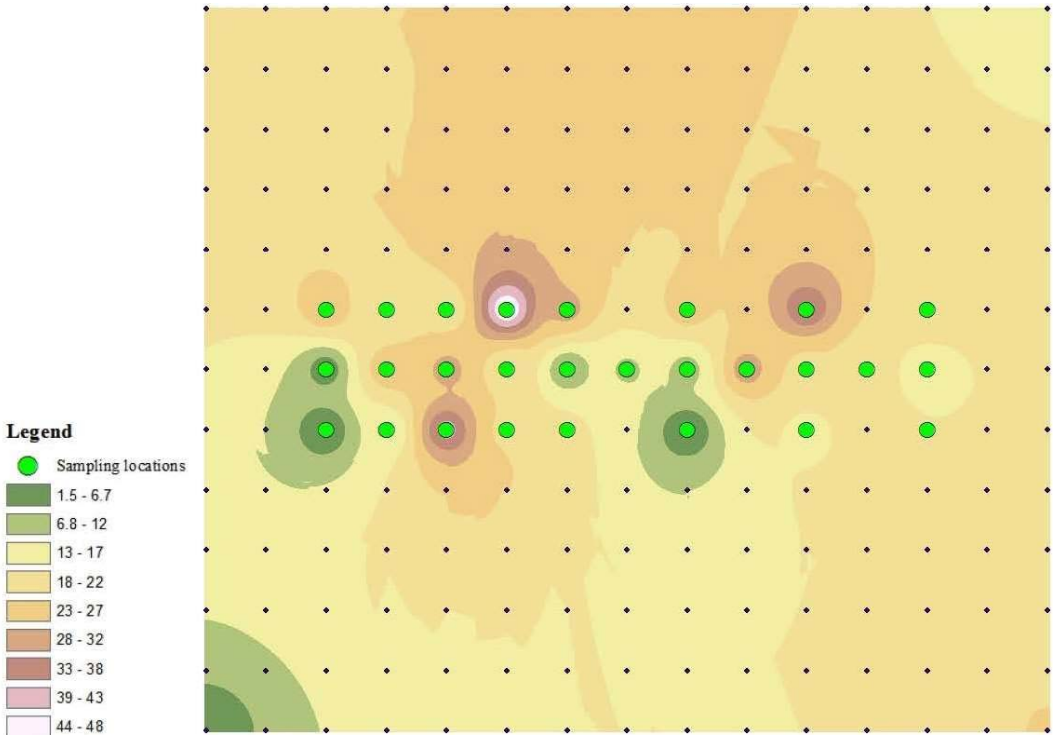



Figure 5.6 Generated hydraulic conductivity for layer three, iteration two; applying the IDW interpolation algorithm (m/day)

### 5.3.5. Performance Evaluation Results

In this section, first, the following steps for constructing surrogate models for source identification in this study are explained. Then, the evaluation results for the performance of the constructed surrogate models are discussed.

1. Problem definition and sampling plan: The problem, its main variables and the objective function of the problem were addressed. As previously mentioned in this chapter, four potential contaminant sources were considered in this study, sources C, D, E, and G. These four sources were included three injection wells (Figure 5.5) and one dummy source (source G) with (10, 2, and 3) coordinates along XYZ directions, respectively. Latin Hypercube Sampling (LHS) was utilized to randomly generate 1000 initial sample sets. The initial data consist of bromide injection concentration values at four potential contaminant sources. The

bromide injection concentrations were assumed to be in the range of 0-300 mg/l for all the potential contaminant sources.

2. Solving the numerical simulation models: Numerical simulation models of groundwater flow (MODFLOW) and transport (MT3DMS) (within GMS 7) were solved for randomly generated bromide injection concentration values at the previous step. The solutions contained the corresponding contaminant concentration magnitudes at selected monitoring locations at specific stress periods (Table 5.2).
3. Developing the surrogate models: SOM, MARS, and GPR algorithms were utilized to develop surrogate models.

Table 5.3 shows a typical set of inputs for training the surrogate models. This input set consists of five sample sets. Each set consists of randomly generated bromide injection concentration values at potential contaminant sources at first stress period (ST1) and corresponding contaminant concentration magnitudes at nine monitoring locations (ML1 to ML9) at four stress periods (ST2 to ST5). It was supposed that if the surrogate models were developed accurately, these models could properly approximate the simulation models of groundwater flow and transport.

Table 5.3 A typical input for training a surrogate model

| ID | Contaminant Sources | | | | Monitoring Locations (ML1-ML9) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 2 | 5 | 6 | 7 | 8 | 9 |
| | Bromide injection concentrations (mg/l) | | | | Contaminant concentrations (mg/l) | | | | | | | | | |
| | ST1 | | | | ST2 | | | ST3 | | ST4 | | | ST5 | |
| 1 | 290 | 251 | 8 | 146 | 13.3 | 36.0 | 5.7 | 0.3 | 55.0 | 2.7 | 0.5 | 0.0 | 15.8 | 0.0 |
| 2 | 163 | 216 | 245 | 157 | 18.3 | 14.9 | 3.7 | 5.4 | 26.1 | 1.2 | 0.1 | 0.2 | 12.4 | 0.0 |
| 3 | 289 | 0 | 5 | 59 | 0.1 | 24.9 | 3.5 | 0.3 | 42.3 | 0.5 | 0.2 | 0.0 | 15.6 | 0.0 |
| 4 | 16 | 159 | 102 | 269 | 13.2 | 1.5 | 0.4 | 0.2 | 3.5 | 0.1 | 0.0 | 0.1 | 0.3 | 0.0 |
| 5 | 55 | 298 | 52 | 84 | 16.8 | 6.7 | 0.0 | 1.6 | 9.2 | 0.0 | 0.1 | 0.1 | 1.5 | 0.0 |

Same sets of training data were used for developing the SOM, MARS, and GPR- based Surrogate Models (SOM, MARS, and GPR-based SMs). However, due to the different natures of the applied algorithms, for developing different surrogate models, different designs were utilized. In the SOM-based SMs, all the training data (Table 5.3) was utilized to develop the SOM-based SMs in a single run. Different SOM-based SMs representing different numbers of SOM map units were constructed. The developed SOM-based SMs without using an optimisation model were utilized for source identification as an inverse problem.

In the training process of GPR and MARS-based SMs, when the developed surrogate models were utilized directly for source identification, first, the predictors and target variables of the system need to be addressed. Since, in source identification problem, just observed contaminant concentrations data is available, unknown groundwater contaminant sources need to be characterized in an inverse mode. Therefore, in the training process of the MARS and GPR-based SMs, the contaminant concentration values of the training data at specific time and locations were addressed to be the predictors of the MARS/GPR prediction models. The randomly generated bromide injection concentrations at potential contaminant sources at specific times were considered to be the target variables of the MARS/GPR prediction models. Each MARS/GPR prediction model can only have one target variable. As a result, for each target variable, separate MARS/GPR model was developed. Then, after developing all the MARS/GPR prediction models, the constructed MARS/GPR prediction models were integrated to develop the MARS/GPR-based SMs. The developed surrogate models could approximate the simulation models of groundwater flow and transport and be applicable for source identification independently of any optimisation model. By providing the measured or simulated contaminant concentration values at specified locations and times, the

MARS/GPR-based SMs could be applied to identify the unknown groundwater contaminant sources in terms of location, magnitudes and time-release. After developing the SOM, MARS, and GPR-based SMs, the developed surrogate models were independently utilized for unknown groundwater contaminant source identification without using an explicit optimization model.

4. Validation of the surrogate models: the developed surrogate models were tested by using new sample sets. The bromide injection concentrations of these sample sets were randomly generated by applying the LHS method in the range of 0-300 mg/l. Then, the corresponding concentration values at monitoring locations were obtained by implementing the simulation models. The performance of the developed surrogate models was evaluated by using Normalised Absolute Error of Estimation (NAEE) as an error criterion. NAEE can be defined by equation 3-12.

To evaluate the capability and efficiency of the SOM, MARS, and GPR-based SMs to identify the unknown source characteristics, when the field concentration measurements resulting from specified bromide injection concentrations in the study area were specified, the surrogate models were utilized in an inverse mode. The simulated contaminant concentration values at specific locations and time of testing data were considered to be the known variables of the system. The developed surrogate models were utilized for source identification by using information regarding these known variables. Table 5.4 presents a typical input dataset with missing data for testing the surrogate models.

In the SOM-based SM case, when utilized in an inverse mode for source identification, the BMU command of the SOM algorithm (equation 3-3) which searches to find the most

similar vector of the SOM-based SM to the testing input data was utilized for source identification. The detailed information of the application of this surrogate model for source identification was discussed in Chapter 3.

Table 5.4 A typical input vector with missing data for testing the developed surrogate models

| Contaminant Sources | | | | Monitoring Locations (ML1-ML9) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 2 | 5 | 6 | 7 | 8 | 9 |
| Bromide injection concentrations (mg/l) | | | | Contaminant concentrations (mg/l) | | | | | | | | | |
| ST1 | | | | ST2 | | ST3 | | ST4 | | | | ST5 | |
| 1 | | | | 10.1 | 7.6 | 0.7 | 1.7 | 10.7 | 0.2 | 0.0 | 0.4 | 7.4 | 0.0 |
| 2 | | | | 2.8 | 5.7 | 0.0 | 0.4 | 11.1 | 0.0 | 0.1 | 0.0 | 3.4 | 0.0 |
| 3 | | | | 2.9 | 21.9 | 5.4 | 6.2 | 23.8 | 1.6 | 0.2 | 0.1 | 16.5 | 0.0 |
| 4 | | | | 13.1 | 21.7 | 0.1 | 3.3 | 29.8 | 0.0 | 0.2 | 0.1 | 18.0 | 0.0 |
| 5 | | | | 16.7 | 11.7 | 0.1 | 4.0 | 16.1 | 0.0 | 0.1 | 0.1 | 2.5 | 0.0 |

Moreover, SOM Map quality could be assessed by various methods. Quantisation Error (QE), a widely used criterion for evaluation of the SOM Maps was utilized. The QE gradually decreases by increasing map sizes. The earlier studies indicate that a suitable number of neurons have an essential role in the accuracy and performance of the SOM algorithm (Di Mauro et al., 2016). The “SOM Toolbox for Matlab 5" software was utilized in this study for constructing the SOM-based SMs (Vesanto et al., 2000).

Performance evaluations of the developed SOM-based SMs representing different numbers of SOM map units are illustrated in Figure 5.7. The obtained results demonstrate that the SOM-based SM with 120×120 map units had the least quantisation error while the surrogate model with 100×100 map units had the lowest error of estimation. Therefore, the developed SOM-based SM with 100×100 map units was considered as the selected surrogate model. This SOM-based SM was selected for its best accuracy of estimation. Required time for constructing a SOM-based SM was an important

consideration, as the computation time exponentially increased by increasing the number of SOM map units (Figure 5.8).

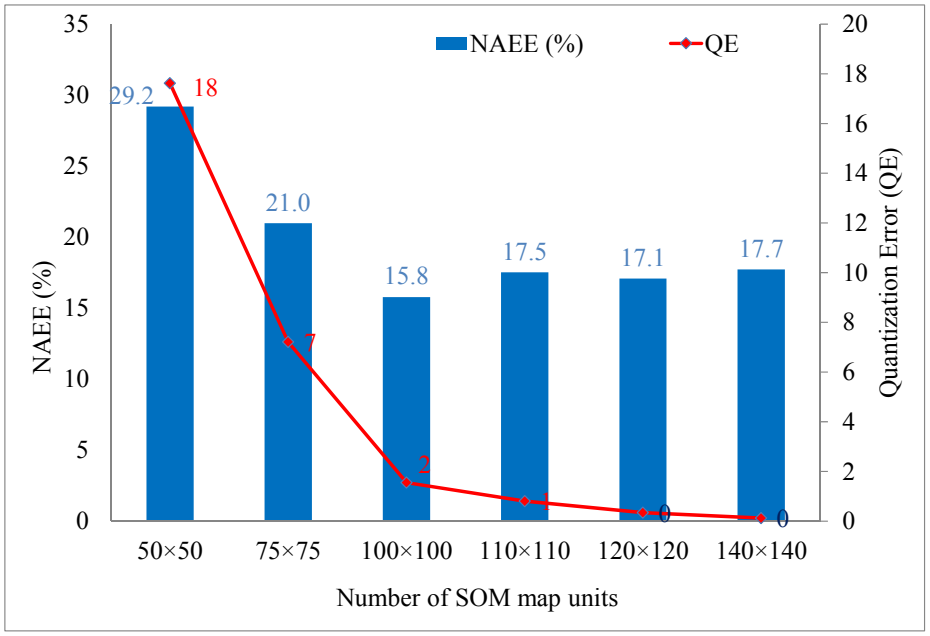



Figure 5.7 The performance evaluation results of the developed SOM-based SMs for various scenarios representing different numbers of SOM map units in terms of NAEE and QE values

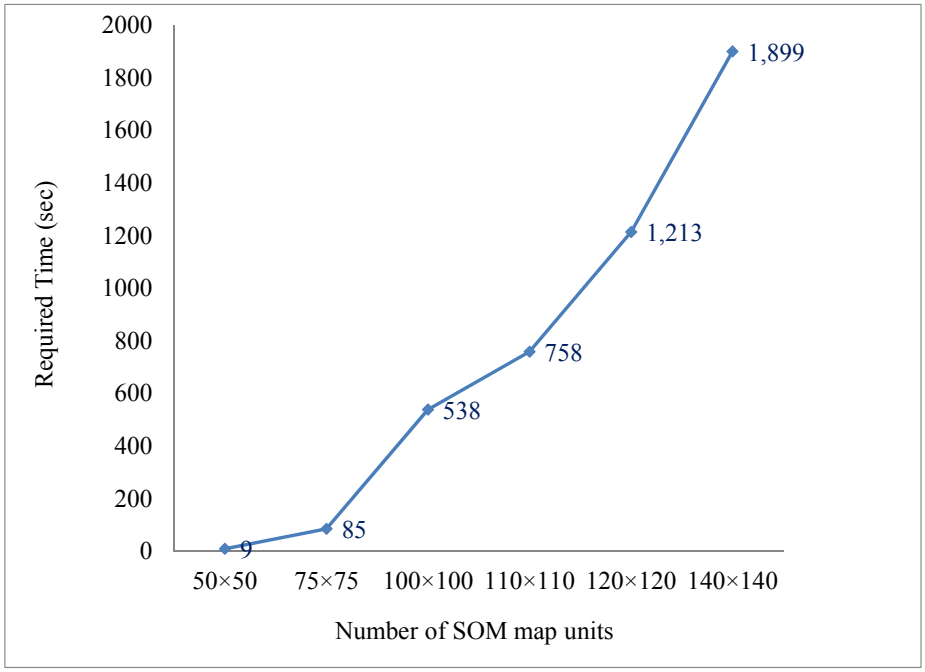



Figure 5.8 The required time for constructing different SOM-based SMs representing different numbers of SOM map units

The GPR and MARS-based SMs for source identification act as prediction models. These prediction models by using simulated contaminant concentration data at specific monitoring locations and times of testing data (Table 5.4) characterize unknown contaminant sources. However, the performances of the developed MARS and GPR-based SMs for source identification were also evaluated by using the same testing data. The performance evaluation results in terms of NAEE were equal to 15.8, 14.1 and 16.2% for the SOM, MARS, and GPR-based SMs, respectively. The results showed similar accuracy for the selected SOM-based SM compared with the performance results of the other two surrogate models. Despite on the average similar performance in terms of accuracy of these three surrogate models for source identification, there are some differences between the results of the developed surrogate models. One of the differences is in the accurately screening of the dummy sources by the SOM-based SM. The SOM-based SM in 98% of the cases accurately could screen the dummy sources against of 6 and 0% correct inferences by the MARS and GPR-based SMs. Actually, the approximation of the GPR and MARS-based SMs for dummy sources were not unsatisfactory. The MARS and GPR-based SMs could appropriately estimate the dummy sources (not actual sources) as very low magnitudes but not exactly as zero flux values. The obtained average NAEE for each source of all the developed surrogate models were compared and presented in Figure 5.9. Although, the accuracy of the developed MARS and GPR-based SMs is higher than the selected SOM-based SM (Figure 5.9); the capability of the SOM algorithm in clustering and subsequently in screening the dummy sources may make the SOM algorithm a potentially powerful tool for the unknown contaminant source identification problems.

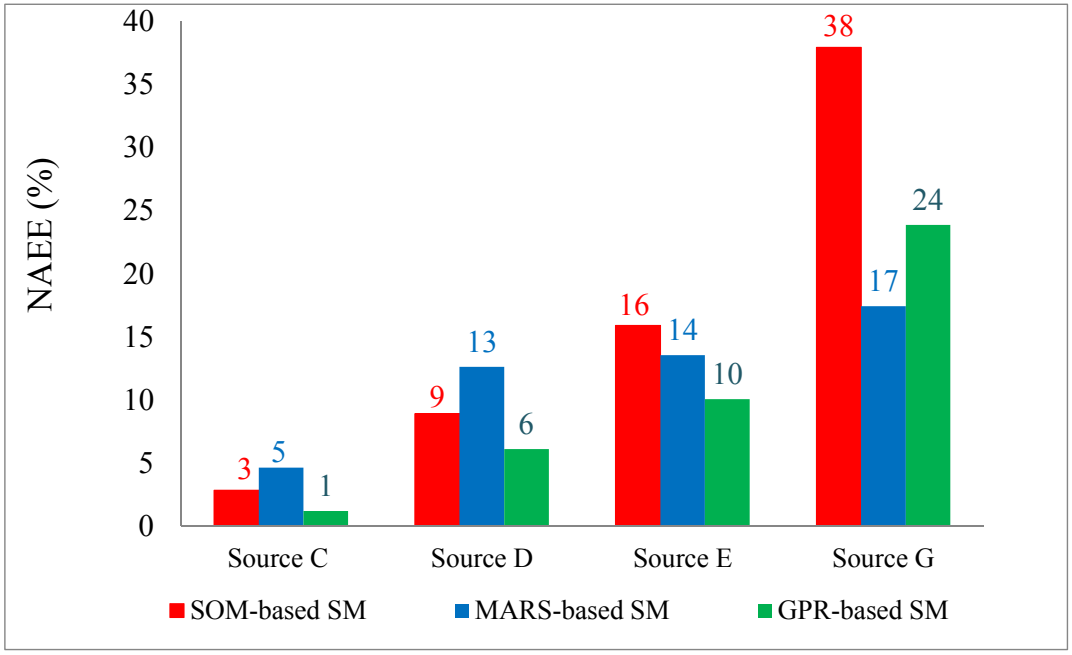



Figure 5.9 The obtained results of the constructed surrogate model for source identification using testing data in terms of NAEE

Based on the obtained performance evaluation results, application of the SOM algorithm for developing surrogate models in source identification problems has some advantages. These advantages are 1. simplifying the source identification problems by screening the dummy source and reducing the number of decision variables, and 2. estimating preliminary results for source identification which these results can be utilized for applying sequential sampling method (Chapter 3).

5. Source identification or recovering source injection history: The obtained results at evaluation stage demonstrate that the developed surrogate models could be utilized for source identification. Therefore, the developed MARS and GPR-based SMs, and the selected SOM-based SM were utilized for source identification. These developed surrogate models by using the measured concentrations (Table 5.2) were utilized to recover bromide injection history from ELE site. The obtained results in terms of NAEE were equal to 34.7, 24.9 and

24.6% for the MARS, SOM, and GPR-based SMs. The obtained results of the developed models for source identification are illustrated in Figure 5.10.

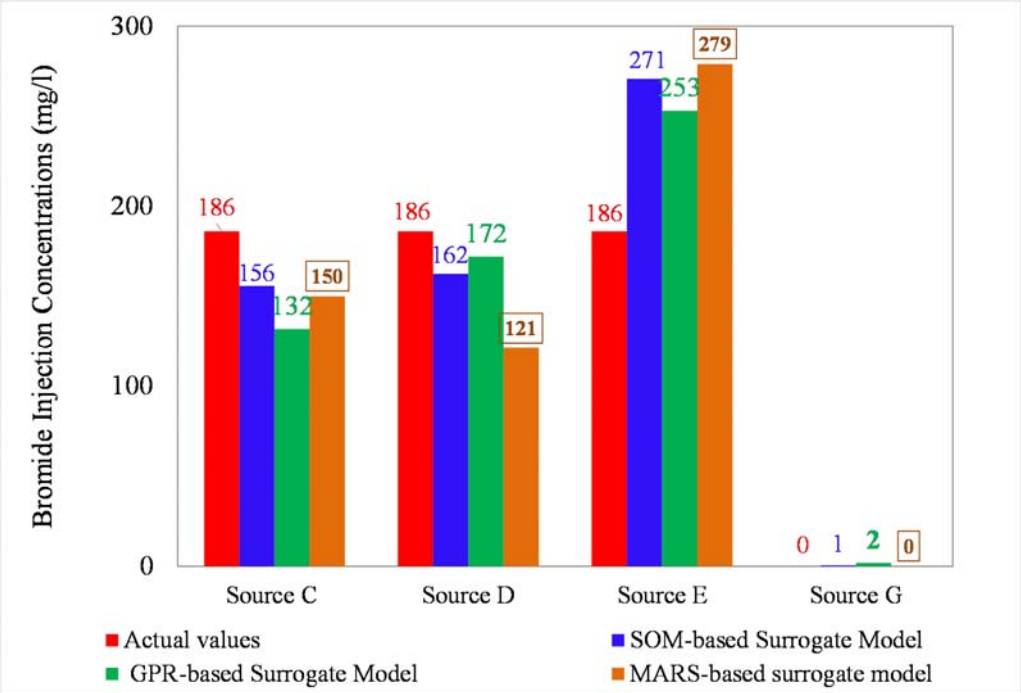



Figure 5.10 comparison of actual data with obtained results of selected SOM, MARS and GPR based SMs for source characterisation in terms of NAEE

MARS-based SMO was also developed for comparison purpose by using same sets of training data. The main steps involving in developing a MARS-based SMO was explained in detail in Chapter 4. For developing the MARS-based SMO, the randomly generated bromide injection concentrations at all potential contaminant sources at specific times were considered to be the predictors of the MARS prediction models. Also, the simulated contaminant concentration values at specific times and locations were assumed to be the target variables of the MARS models. Then, the MARS algorithm was utilized to develop MARS prediction models. In the next step, all the constructed MARS prediction models for all the target variables were integrated in MATLAB. Then, the developed MARS-based SM was linked to a Genetic Algorithm (GA)-based optimisation

model for source identification. The objective function of this optimisation model was defined by equation (4-1) (Jha & Datta, 2013; Mahar & Datta, 2001). This optimisation model minimises the difference between the simulated contaminant concentration values and corresponding observed contaminant concentration values. The MARS-based SMO was also applied to recover bromide injection history by using the measured contaminant concentrations (Table 5.2). The solution results are presented in Figure 5.11.

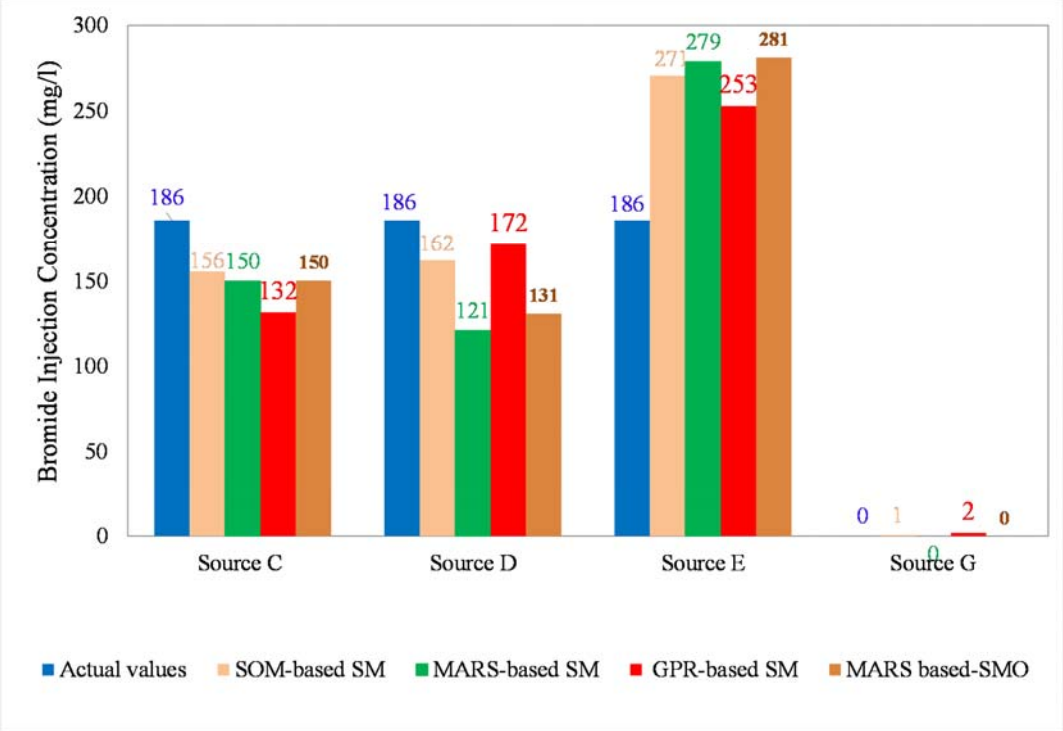



Figure 5.11 Comparison of actual data with obtained results of of the developed surrogate models for source identification in terms of NAEE

### 5.4.Conclusion

This chapter presents the performance evaluations of the constructed surrogate models in identifying unknown groundwater contaminant sources in an experimental site. The contaminated ELE site was a heterogeneous aquifer site with errors in measured contaminant concentration values. Limited performance evaluations of the developed

methodology were conducted to test the efficiency of the developed methodologies in source identification. The SOM, GPR, and MARS algorithms were utilized to construct the surrogate models for source identification. Various scenarios correspond to different surrogate models with various numbers of SOM map units were developed. The MARS and GPR based SMs using same training data were also developed. Main conclusions that could be drawn from these performance evaluation results can be listed as:

1. The SOM, MARS, and GPR based SMs were potentially efficient to approximate the groundwater flow and transport simulation processes in a multilayer heterogeneous experimental contaminated aquifer site.
2. The performance assessment results demonstrate potential applicability of the SOM, MARS, and GPR algorithms as the surrogate model types in an inverse mode, for source identification problems with erroneous contaminant concentration data (Figure 5.10).
3. In source identification problems, SOM algorithm capability in clustering multidimensional input data leads the SOM-based SM to screen dummy sources, i.e., not actual sources but included as potential sources precisely.
4. The developed surrogate models may provide a feasible approach for source identification in terms of location, magnitude, and release history, without the necessity of using a linked simulation-optimisation model.

MARS-based SMO was also developed for comparison purpose. The obtained solution results of the MARS-based SMO and MARS-based SM were very similar. The SOM- and GPR-based SM results show more accuracy in source identification. However, because of the high probability of using erroneous concentration data, it cannot be concluded that SOM-and GPR-based SMs perform more accurate than the MARS-based

SM and SMO. Especially, the SOM-based SMs source identification solution results usually do not show more accuracy than the MARS-based SMs (Chapter 3).

However, these performance evaluation results are limited to specific cases and further evaluations are necessary to establish the applicability of the utilized algorithm and the developed surrogate models.

In the next chapter, the application of the developed methodologies to a contaminated field in Australia is discussed.

# 6. Source Identification by Using Surrogate Models based Optimisation in Conjunction with Monitoring Network Design Using Field Data

## 6.1. Introduction

This chapter presents the application of the developed procedures for source identification to a real contaminated aquifer site. Self-Organising Maps (SOM) and Multivariate Adaptive Regression Splines (MARS) algorithms were utilized to develop surrogate models. In this chapter, unknown groundwater contaminant source characteristics in a real-world contaminated aquifer site were identified by using SOM-based Surrogate Model (SOM-based SM) and MARS-based Surrogate Model (MARS-based SM) linked to an optimisation model. The performance evaluations of the developed surrogate models in conjunction with the designed monitoring network procedure are also presented in this chapter.

However, first, the developed procedures for designing monitoring network and surrogate models for source identification are briefly discussed. The developed methodologies for monitoring network design and surrogate models are presented in detail in Chapters 4 and 3, respectively. Second, the contaminated aquifer site and related issues are briefly explained. Some details of this contaminated aquifer site are not included in this chapter due to confidentiality requirements. Then, the application of the developed monitoring network design procedure and surrogate models to the contaminated aquifer site is presented and discussed. Finally, main conclusions are summarised.

## 6.2. Methodology

### 6.2.1. Surrogate Models

Unknown groundwater contaminant source identification problem usually needs to be addressed by three main questions. These questions investigate contaminant source

location(s), magnitudes and release history. For successful groundwater management, developing an efficient methodology to answer these three main questions is necessary. The most frequently applied methodology for source identification is linked simulation-optimization approach. This approach consists of numerical simulation models and optimization models, with the linked simulation model embedded or implicitly embedded within the optimization model (Mahar & Datta, 2000). The main drawback of this approach is that its applications in real-world cases are computationally very intensive. To overcome this drawback, simulation models are replaced by surrogate models to develop Surrogate Models based Optimization (SMO). In the SMO, the optimization model instead of linking to a complex and time-consuming simulation model is linked to a simpler and faster surrogate model. This surrogate model can efficiently decrease the computational time once the surrogate models are developed after training and testing.

The main aim of this study was to develop an efficient approach for source identification. According to the obtained results of the assessment of the performance of the constructed surrogate models for source identification which are explained in Chapters 3 and 4, the capabilities of SOM in clustering lead SOM-based SMs to screen the active source(s) among all potential contaminant sources. In other words, a SOM-based SM could provide at least the answer of one of the three main questions of the source identification problem (contaminant source locations). Moreover, based on the presented evaluation results for performance of the developed surrogate models in Chapters 3 and 5 by using testing data, the MARS-based SM could be more accurate than the SOM-based SM in estimating contaminant source fluxes. This accuracy is because of the MARS algorithm's capability in interpolating and approximating multidimensional data. Therefore, the SOM and MARS algorithms were utilized to develop surrogate models and MARS-based SMO for source identification at a contaminated aquifer site in NSW, Australia. In the developed

methodologies, SOM-and MARS-based SMs which are generally simpler and faster compared to few other surrogate models due to model development algorithms used, are utilized to approximate the complex flow and transport process in a contaminated aquifer. In the developed MARS-based SMO, a Genetic Algorithm (GA) was utilized as the optimisation algorithm.

The developed monitoring network design procedure in Chapter 4 was also utilized to identify the monitoring locations and measurements which contributed comparatively more to source identification. Information obtained from the designed monitoring network was utilized in the developed surrogate models for source identification to improve the source identification results.

## 6.3. Integration of the Developed Source Identification Methodologies with the Designed Monitoring Network Approach

Designing a monitoring network is one of the essential steps of contaminant source identification and subsequently remediation. Using the obtained data from the designed monitoring network could efficiently improve source identification results. Further, collecting the contaminant concentration values for a long period would be more affordable. Usually, contaminant concentration values initially available at a contaminated site are limited and sparse. In these cases, designing monitoring networks and using concentration measurement information obtained from the designed monitoring network can significantly improve the source identification results.

The developed monitoring network design procedure which is presented in Chapter 4 was applied to improve source identification results. Random Forests (RF), Classification and Regression Trees (CART), and Tree Net (TN) tools were utilized as the designing monitoring network tools in this study. These tools are robust data mining techniques in regression and classification. These tools use the training datasets which consisted of

predictors and target variables to construct prediction models. These tools, utilized in the process of constructing prediction models, could determine the degree of importance and influence of the predictors in predicting the target variables. These capabilities of the RF, TN and CART tools were utilized to detect the most important monitoring wells among all the potential monitoring wells.

The performance of the developed monitoring network design procedure for source identification was also evaluated for data sets set aside for testing at arbitrary and selected monitoring locations. The developed surrogate models were updated by using new information from the designed monitoring network and then utilized for evaluating source identification efficiency for the testing data set.

## 6.4. Contaminated Aquifer Site

The contaminated aquifer site utilized in this study is a part of the upper Macquarie groundwater management area, NSW, Australia (Prakash & Datta, 2015). The exact location of the contaminated aquifer site is not constituted for confidentiality reasons. The detective contaminant in this contaminated aquifer site was BTEX which refers to the chemicals benzene, toluene, ethylbenzene, and xylene. For the first time, due to the complaints of BTEX vapour from buildings' basements, investigations of the contaminant aquifer were conducted. The first time that BTEX was recorded in this area is not certain. However, the contamination area was approximated roughly to be more than $1km^2$ (Prakash & Datta, 2015).

For monitoring the contaminant concentrations in the contaminated area, 74 monitoring wells were installed from October 2006 to July 2011 (Prakash & Datta, 2015) and 19 of these 74 monitoring wells were utilized as injection wells to inject neutraliser. According to the investigation results in the area, the contaminant source location probably was a leaking underground storage tank at a fuel station (Prakash & Datta, 2015). The starting

time of the leakage and time-release period were not identified by investigations (Prakash & Datta, 2015).

## 6.5. Study Area

As mentioned earlier in this chapter, information of a contaminated aquifer site in NSW, Australia was utilized for assessing the performance of the developed procedures. The Macquarie River is located on the western boundary of this contaminated aquifer site. For the purpose of simulation, due to lack of any specific geological formation at the other three boundaries of this contaminated aquifer, a larger area for simulation purposes was considered in this study (Prakash & Datta, 2015). This area measuring 2.1871km by 2.4256km included all hydrogeological conditions affecting the contaminated study area. To prevent any confusion, the contaminated area is identified as the "contaminated site" and the considered area is identified as the "simulated area". Figure 6.1 illustrates the simulated area and the contaminated site (Prakash & Datta, 2015). The elevation at the simulated study area varies from 292-251m on the north-eastern side (Prakash & Datta, 2015). Figure 6.1 shows the simulated study area.

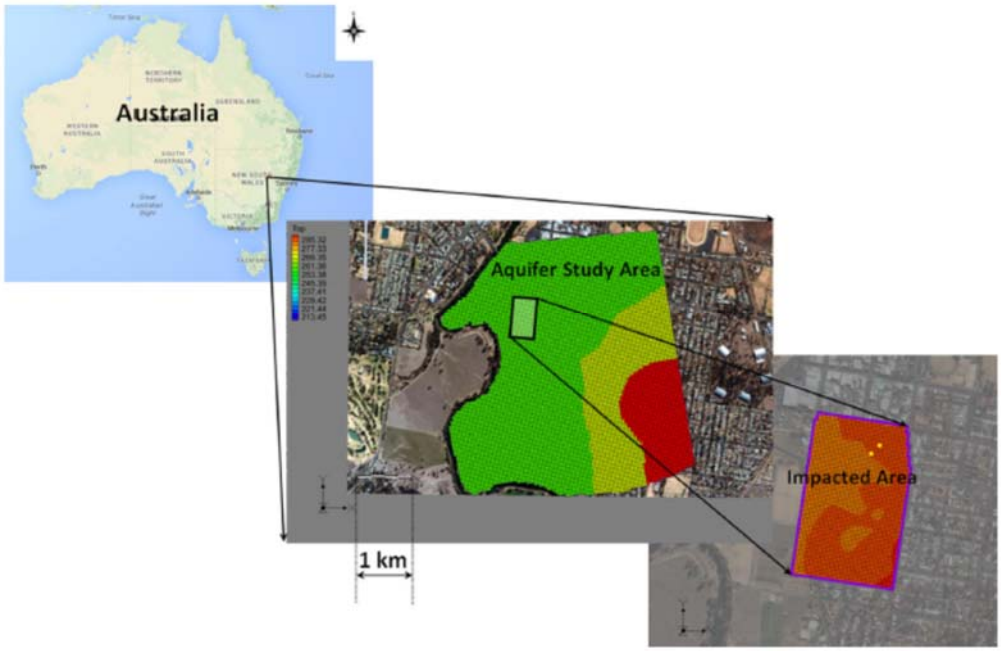



Figure 6.1 Plan view of the study area and the contaminated area (Prakash & Datta, 2015)

As previously mentioned, in this contaminated area, contaminant first was detected in the basement of buildings as vapour (Prakash & Datta, 2015). Then, an extensive investigation was started at the contaminated site from October 2006 to July 2011. As a result, the BTEX concentration data for this contaminated aquifer were collected at about 55 wells approximately every three months. The installed wells were located more around the preliminary identified potential contaminant sources. In this period, the contaminant concentration values were not recorded regularly at all the wells. In other words, only at a few wells were the data available at all the mentioned times. However, the maximum recorded BTEX magnitudes at this contaminated site during the mentioned period was 320mg/l.

Rainfall and river were the main recharge sources of the simulated study area (Prakash & Datta, 2015). Long-term average annual rainfall of the study area was 583mm. Evapotranspiration and groundwater extraction was the main loss sources of the study area. Groundwater was extracted from the study area, through several wells for domestic

and irrigation purposes. The extraction rates over the years were not constant. The evapotranspiration in the dry season can reach to 260 mm/month.

## 6.6. Simulation Models

The calibrated groundwater flow simulation model by Prakash and Datta (2015) was utilized. In the MODFLOW (within GMS7), the layer property flow package was utilized to model the groundwater flow. Table 6.1 presents important information of the simulated area. The simulated study area based on the available bore-hole logs, can be divided into three distinct layers (Prakash & Datta, 2015). The top layer and middle layer mainly were comprised of tertiary alluvium and quaternary alluvium, respectively. The third layer was comprised of impermeable bedrock. The thickness of these layers varied from one point to another. Due to the sparsity of boreholes across the study area (Prakash & Datta, 2015), the thickness of layers at the other points needed to be interpolated.

Table 6.1 Hydrogeologic characteristics of the contaminated study area (Prakash, 2014)

| Parameter | Unit | Value |
|---|---|---|
| Maximum length | metre | 2187.1 |
| Maximum width | metre | 2425.6 |
| Saturated thickness, b | metre | Variable |
| Number of Layers | - | 3 |
| Grid spacing in x-direction | metre | 21.87 |
| Grid spacing in y-direction | metre | 21.08 |
| Grid spacing in z-direction | metre | Variable |
| Kxx (layer 1, layer 2, layer 3) | metre/day | 12.37, 16.24, 0.001 |
| Kyy (all layers) | metre/day | 0.2 |
| Porosity (all layers) | - | 0.27 |
| Longitudinal Dispersivity | metre | 12 |
| Transverse Dispersivity | metre | 6 |
| Horizontal anisotropy | - | 1.5 |
| Specific Yield (all layers) | | 0.1 |
| Specific Storage | $(metre)^{-1}$ | 0.000006 |

Due to the presence of the river at the west of the simulated area, a specific head boundary was considered at this boundary. Based on the previous studies in the region, hydraulic heads at the other boundaries were estimated (Prakash & Datta, 2015).

In the calibrated model, groundwater flow of the simulated study area was modelled from 1st January 1995 to the end of December 2012. The total time of simulation was divided into 18 different stress periods all with one-year duration. The activity duration of the sources was assumed to be 10 years, started in 1999. The total time of contaminant source(s) activities were divided into 10 equal stress periods (ST1 to ST10) of one year each (Prakash & Datta, 2015). It was also assumed that the contaminant source concentrations were constant over each stress period.

According to the investigation results, the contaminated source(s) might be a leaking underground storage tank at a fuel station. Due to the uncertainty about the exact location of the contaminant source, in this study, two potential contaminant sources were considered and their coordinates are presented in Table 6.2. The contaminant fluxes from each of the potential contaminant sources are presented as $PCS_i$-$T_j$, where i and j indicate the number of contaminant sources and the stress periods, respectively. Therefore, in this study, total 20 unknown contaminant source concentration values were considered to be unknown variables.

Table 6.2. The grid locations of potential contaminant sources

| ID | Potential contaminant sources | Grid locations of (k, i, j) |
|---|---|---|
| 1 | PCS1 | (17, 29, 1) |
| 2 | PCS2 | (16, 24, 1) |

Rainfall was assumed to be uniform throughout the simulated study area. Rainfall was also assumed to be constant over each stress period. However, due to residential conditions of the simulated area which is a suburb, just 10% of the rainfall was considered to be infiltration recharge (Prakash & Datta, 2015). The extraction rates from the wells

used in the flow simulation model varied from one well to the other. The details of the extraction rates that used in the flow model can be found in (Prakash & Datta, 2015). However, the extraction rates were assumed to be constant over each stress period.

The groundwater flow model of the simulated study area was calibrated by using recorded hydraulic head data at 31 monitoring wells which spread throughout the contaminated area (Prakash & Datta, 2015). The calibration criteria were that the difference between the observed and simulated hydraulic head values remain within 1-metre intervals with a confidence level of 90% (Prakash & Datta, 2015).

For modelling the transport and fate of the BTEX, MT3DMS was utilized. The MT3DMS (Zheng & Wang, 1999) uses the flow field generated by the MODFLOW (Harbaugh, 2005) to predict the movement of contamination over the different stress periods in the contaminated aquifer. In this transport model, BTEX was assumed to be a conservative contaminant. The contaminant plume boundary was assumed to be contained within the boundary of the contaminated area (Prakash & Datta, 2015). The initial concentration values of this contaminant were also assumed to be zero in the transport simulation model. The measured contaminant concentrations from January 2009 to April 2010 were utilized as the observed concentration values of the contaminated study area for identifying unknown contaminant sources.

## 6.7. Performance Evaluations of the Developed Methodologies

### 6.7.1. Application of the Developed Monitoring Network Design Procedure

In this study, first, the developed monitoring network design procedure was used to identify the monitoring locations that contributed most to the source identification process. Then, surrogate models were developed based on the obtained information from the designed monitoring wells. In the contaminated aquifer site, 578 measured contaminant concentrations were available at 55 monitoring wells. At 20 of these 55

monitoring wells, more contaminant concentration values are available at the period of January 2009 to April 2010 compare to the other monitoring wells. So, the information of these 20 monitoring locations was utilized to select the monitoring locations that had most contributions to source identification.

TR, RF, and CART tools were utilized to rank the available monitoring locations by their contributions to the source identification process. The following steps were followed to use RF, TR, and CART to develop prediction models and select the most important monitoring locations among all potential monitoring locations. Figure 6.2 also presents the flow chart of designing a monitoring network procedure for source identification.

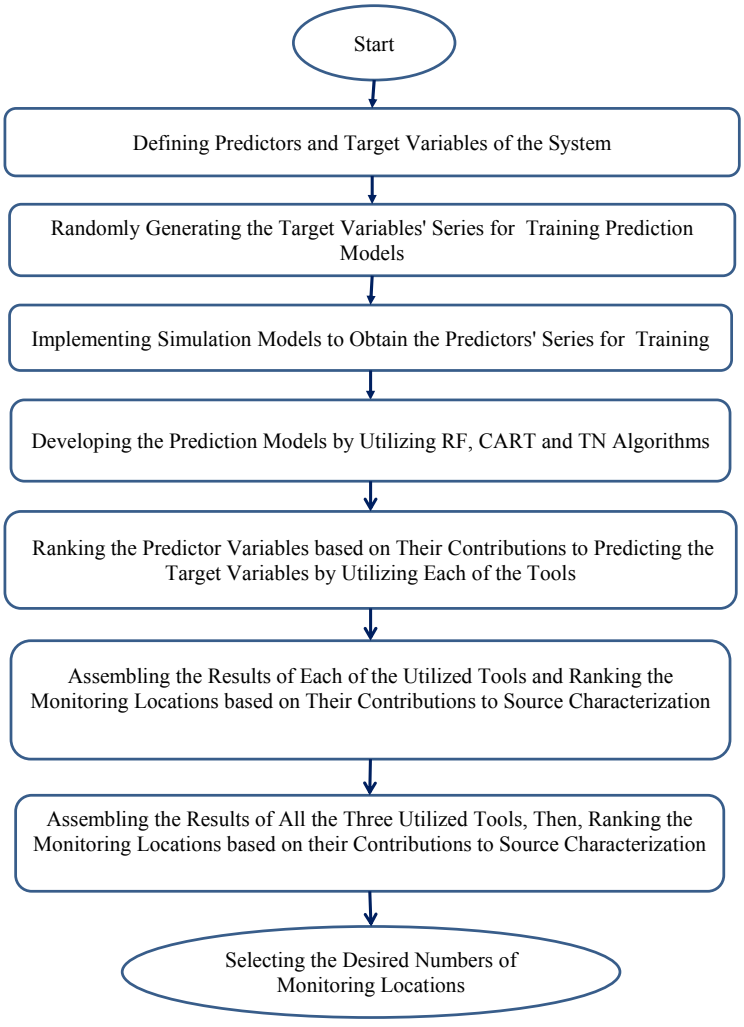



Figure 6.2 Schematic chart of the developed monitoring network design methodology using RF, TN and CART tools

1. Defining the target variables and predictors of the system: The predictors and target variables of prediction models were addressed in this step. Contaminant concentration values at all potential monitoring wells at specific times were considered to be predictors. Contaminant fluxes at two potential contaminant source locations at specific times (10 stress periods) were considered to be the target variables.
2. Generating training data for developing the prediction models: 1500 initial sample sets of contaminant source fluxes were generated by applying Latin Hypercube Sampling (LHS). These 1500 sample sets consisted of three groups of 500 initial sample sets. In the two groups, the sample sets were generated by considering that one of the potential contaminant sources might be inactive. The possible contaminant source fluxes were generated in the range of 0-100 g/s.
3. Implementing simulation models: Groundwater flow (MODFLOW) and transport (MT3DMS) (within GMS 7) simulation models, were implemented to obtain corresponding contaminant concentrations at all potential monitoring wells at specified times. The simulated contaminant concentration values at 1$^{st}$ January, April, July, and October 2009, and 1$^{st}$ January 2010 were utilized to train the prediction models.
4. Developing the prediction models: As mentioned earlier, RF, TN, and CART tools were utilized to develop prediction models in this study. In each prediction model, by using each of the utilized tools, one target variable needs to be addressed. So, for each utilized tool, 20 prediction models were developed. Totally, by using RF, TN, and CART tools, 60 prediction models were constructed.

Table 6.3 shows a typical input for a prediction model by using each of the RF, CART and TN tools. In this table, the simulated contaminant concentration values at two monitoring wells (MW1 and MW2) at four different times were assumed to be the predictors of prediction models. The contaminant source fluxes at specific locations and times were considered to be target variables. In this typical input, just 10 sample sets were considered as the training data.

5. Ranking the predictors based on their contributions to predicting the target variables: The RF, TN and CART tools by developing each prediction model also assigned a weight to all the predictors which demonstrate the level of their contributions in predicting the target variable. For each utilized algorithm, all the assigned weights to each predictor (total 20 weights) in the process of predicting all the target variables (20 target variables) were assembled. After assembling weights for all the predictors, the predictors were ranked by their weights from largest to smallest. As mentioned earlier, these weights demonstrate the level of predictors' contributions to predicting the target variables. Because variation ranges of the assigned weights for different tools were different, the assembling results of each tool were again re-weighted by weights that variated in the same range from 1 to n, in which n represents the total number of potential monitoring wells. The monitoring wells with more contributions to source identification were assigned by larger weighs. For example, the monitoring well with the most contributions in source identification was assigned by 20.

Table 6.3. Typical input vectors using in the RF, TN and CART prediction models

| ID | Target Variable | Predictors | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Contaminant source fluxes (g/s) | Contaminant concentration values (μg/l) | | | | | | | |
| | PCS2 | MW1 | | | | MW2 | | | |
| | T10 | Time after the start of first source activity | | | | | | | |
| | | 39931 | 40108 | 40199 | 40479 | 39835 | 39931 | 40290 | 40749 |
| 1 | 14.1 | 1834 | 1895 | 1934 | 2023 | 4816 | 4884 | 5253 | 5513 |
| 2 | 49.6 | 1781 | 1825 | 1848 | 1887 | 4479 | 4499 | 4600 | 4793 |
| 3 | 4.2 | 1689 | 1725 | 1751 | 1816 | 4758 | 4759 | 4792 | 4677 |
| 4 | 30.4 | 1281 | 1300 | 1318 | 1373 | 4197 | 4221 | 4450 | 4782 |
| 5 | 30.9 | 1497 | 1534 | 1562 | 1640 | 3538 | 3490 | 3363 | 3437 |
| 6 | 95.6 | 997 | 1037 | 1060 | 1118 | 3442 | 3529 | 3974 | 4559 |
| 7 | 21.4 | 1567 | 1617 | 1651 | 1728 | 4068 | 4103 | 4371 | 4562 |
| 8 | 47.1 | 1550 | 1662 | 1735 | 1913 | 3523 | 3595 | 3981 | 4448 |
| 9 | 43.4 | 1375 | 1397 | 1413 | 1466 | 4655 | 4674 | 4659 | 4598 |
| 10 | 81.4 | 2033 | 2058 | 2064 | 2045 | 4249 | 4305 | 4442 | 4486 |

6. Assembling the results of three utilized tools: The ranking results of all utilized tools for all the monitoring locations were assembled. The assembled results for all the potential monitoring locations based on their potential influence on source identification are presented in Table 6.4.

Table 6.4. Ranked potential monitoring locations based on their contributions to source identification by using RF, CART and TN

| Monitoring locations | | Cumulative weight | Rank |
|---|---|---|---|
| MW23 | (1,25,32) | 60 | 20 |
| MW24 | (1,30,25) | 59 | 19 |
| MW25 | (1,24,20) | 59 | 18 |
| MW15 | (1,24,28) | 52 | 17 |
| MW13 | (1,24,22) | 50 | 16 |
| MW17 | (1,18,24) | 47 | 15 |
| MW14 | (1,24,25) | 39 | 14 |
| MW16 | (1,21,23) | 38 | 13 |
| MW11 | (1,22,28) | 36 | 12 |
| MW20 | (1,20,29) | 36 | 11 |
| MW18 | (1,21,27) | 34 | 10 |
| MW21 | (1,13,24) | 34 | 9 |
| MW05 | (1,22,25) | 30 | 8 |
| MW08a | (1,20,26) | 20 | 7 |
| MW02 | (1,20,25) | 19 | 6 |
| MW04 | (1,21,25) | 18 | 5 |
| MW18 | (1,21,27) | 18 | 4 |
| RW19 | (1,29,25) | 16 | 3 |
| MW22 | (1,17,32) | 14 | 2 |
| MW19 | (1,17,26) | 11 | 1 |

### 6.7.2. Developing Surrogate Models and SMO for Source Identification

As mentioned earlier, the SOM algorithm was used to develop SOM-based SMs for source identification. The MARS algorithm was also utilized to develop MARS-based SMO for source identification. The following steps were followed to develop SOM-based SMs and MARS-based SMO for source identification:

1. Problem definition and sampling plan: Selection of the main variables of the system were addressed in this step.
2. Preparing training data: To develop surrogate models, 1500 randomly generated sample sets of contaminant source fluxes were utilized for training surrogate models. The sample sets consisted of randomly generated contaminant source fluxes at two potential contaminant sources at specific times and corresponding simulated contaminant concentrations at specific monitoring wells at specific times.
3. Developing the surrogate models: The SOM and MARS algorithms were utilized to develop surrogate models. The process of developing a SOM-based SM for source identification is explained in detail in Chapter 3. The main steps to develop a MARS-based SM which can be linked to an optimization model for source identification are also discussed in Chapter 4. As mentioned in Chapter 3 and Chapter 5, the main advantage of the developed SOM-based SMs is their capability in screening dummy sources among all potential contaminant sources. In this chapter, the main purpose of developing SOM-based SMs was to evaluate the performance of these surrogate models in screening dummy source(s) in a real contaminated aquifer site by using sparse and limited data.

However, 1500 randomly generated contaminant source fluxes and obtained corresponding contaminant concentration values at three arbitrary monitoring wells, MW17 (1, 18, 24), MW19 (1, 17, 26), and MW21 (1, 13, 24), were utilized to develop the surrogate models. Table 6.5 shows a typical training data for developing surrogate models. This training data consists of 10 sample sets. Each set consists of randomly generated contaminant source fluxes at two potential contaminant sources at three

specific times (T1 to T3) and corresponding contaminant concentration values at three monitoring wells (MW1 to MW3) at two specific times (T10 and T11).

Table 6.5. Typical input data for training a surrogate model

| ID | PCS1 | | | PCS2 | | | MW1 | | MW2 | | MW3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Contaminant source fluxes (g/s) | | | | | | Contaminant concentration values (µg/l) | | | | | |
| | T1 | T2 | T3 | T1 | T2 | T3 | T10 | T11 | T10 | T11 | T10 | T11 |
| 1 | 83.17 | 59.51 | 61.52 | 17.14 | 72.13 | 24.21 | 5293 | 5604 | 11570 | 6278 | 5770 | 4857 |
| 2 | 90.61 | 55.38 | 30.40 | 74.41 | 31.07 | 74.88 | 4802 | 4807 | 13340 | 7283 | 5671 | 4949 |
| 3 | 11.42 | 90.04 | 99.65 | 49.04 | 63.23 | 88.21 | 4923 | 4999 | 6915 | 4250 | 4754 | 4229 |
| 4 | 91.70 | 28.91 | 18.99 | 66.33 | 15.58 | 85.59 | 4176 | 4445 | 12990 | 6714 | 5316 | 4389 |
| 5 | 63.92 | 34.14 | 62.71 | 97.05 | 47.17 | 60.96 | 3810 | 3886 | 8228 | 4802 | 4394 | 3599 |
| 6 | 8.77 | 13.53 | 58.09 | 31.89 | 1.01 | 67.70 | 3580 | 3968 | 18730 | 9491 | 5427 | 4265 |
| 7 | 29.21 | 95.90 | 33.88 | 7.60 | 22.60 | 91.67 | 4443 | 4679 | 10820 | 5322 | 3562 | 2301 |
| 8 | 56.00 | 65.51 | 13.31 | 80.71 | 26.75 | 44.93 | 4412 | 4918 | 12130 | 6884 | 5969 | 4733 |
| 9 | 96.16 | 48.46 | 2.18 | 37.19 | 52.94 | 29.87 | 4602 | 4623 | 10600 | 6219 | 5364 | 4757 |
| 10 | 14.99 | 56.00 | 16.39 | 19.40 | 37.04 | 30.23 | 4907 | 5570 | 7608 | 4772 | 5931 | 4875 |

4. Evaluation of the developed surrogate models for testing data: New contaminant source fluxes were randomly generated using LHS. Then, MODFLOW and MT3DMS (within GMS 7) were utilized to obtain corresponding contaminant concentrations at specific monitoring locations at specified times. These obtained (simulated) contaminant concentrations were utilized to test the performance of the developed surrogate models for source identification in an inverse mode. Root Mean Square Error (RMSE) (equation (4-3)) and Normalised Absolute Error of Estimation (NAEE) (equation (3-9)) were utilized to quantify the performance evaluation of the developed surrogate models for testing data.

The performance of the developed MARS-based SM by using information from the three arbitrary monitoring locations was evaluated for testing data. The average performance

evaluation results of the developed MARS-based SM for testing data, in terms of RMSE equal to 42.5. The results were not entirely satisfactory, so for improving source identification results, the results of the designed monitoring network were utilized to develop new surrogate models.

5. Developing new surrogate models by using results of the designed monitoring network: To improve the accuracy of source identification results and evaluate the performance of the developed monitoring network procedure, new SOM-and MARS-based SMs were developed. These models were developed by using information from the top three, six and nine ranked monitoring locations in Table 6.4.
6. Evaluation of the new surrogate models: The results of the new developed MARS-based SMs using information from three arbitrary monitoring wells, top three, six and nine ranked monitoring wells are presented in Table 6.6.

Table 6.6. The performance evaluation results of the developed surrogate models of testing data in terms of RMS and NAEE

| ID | MARS-based SMs | |
|---|---|---|
| | NAEE (%) | RMSE |
| Three arbitrary monitoring wells | 0.9 | 42.5 |
| Three selected monitoring wells | 0.3 | 7.2 |

Comparison of obtained results of the developed MARS-based SMs for testing data indicates that the accuracy of results improved in terms of RMSE by using information from the top-ranked monitoring wells versus the three arbitrary monitoring wells.

The obtained results of two different developed SOM-based SMs using the information from three arbitrary and selected monitoring wells for testing data indicate that the developed SOM-based SMs could screen the dummy source(s) accurately among all the potential contaminant sources. For example, the developed SOM-based SMs in all the

cases accurately screened the dummy source(s). Therefore, the application of the SOM-based SMs for screening the potential dummy source(s) would be acceptable.

7. Developing the MARS-based SMO: The developed MARS-based SMs by using information from three arbitrary and selected monitoring locations were linked to an optimisation model. A GA optimisation model was utilized to solve the objective function of source identification problem (4-1).
8. Source identification: The developed SOM-based SMs and MARS-based SMO(s) were utilized for source identification. The recorded contaminant concentration data at the initial arbitrary monitoring wells and the selected monitoring wells were used for source identification. Breakthrough curves at the monitoring locations which were used for source identification are presented in Figure 6.3.

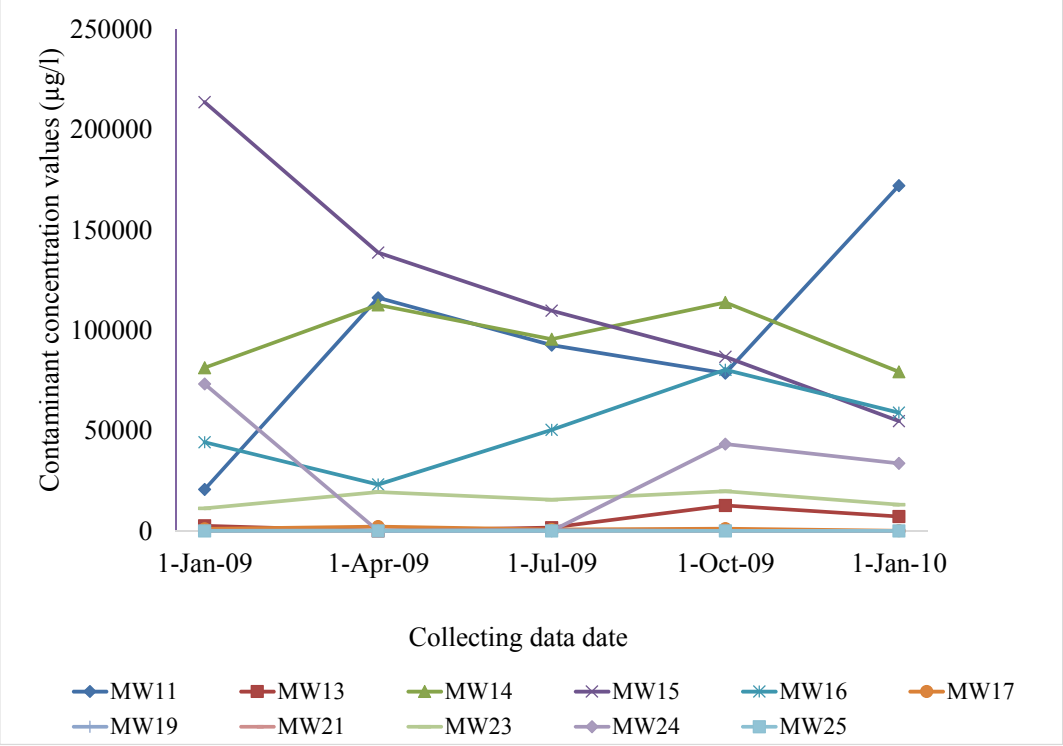



Figure 6.3 Breakthrough curves of specific monitoring wells at specific times utilized for source identification in the contaminated aquifer

The obtained source identification results for different SOM-based SMs representing different numbers of monitoring wells (3MW and 9MW) by using the observed contaminant concentration values are presented in Figure 6.4. The obtained results of these three developed SOM-based SMs demonstrate that PCS1 was an active source and PCS2 was a dummy source.

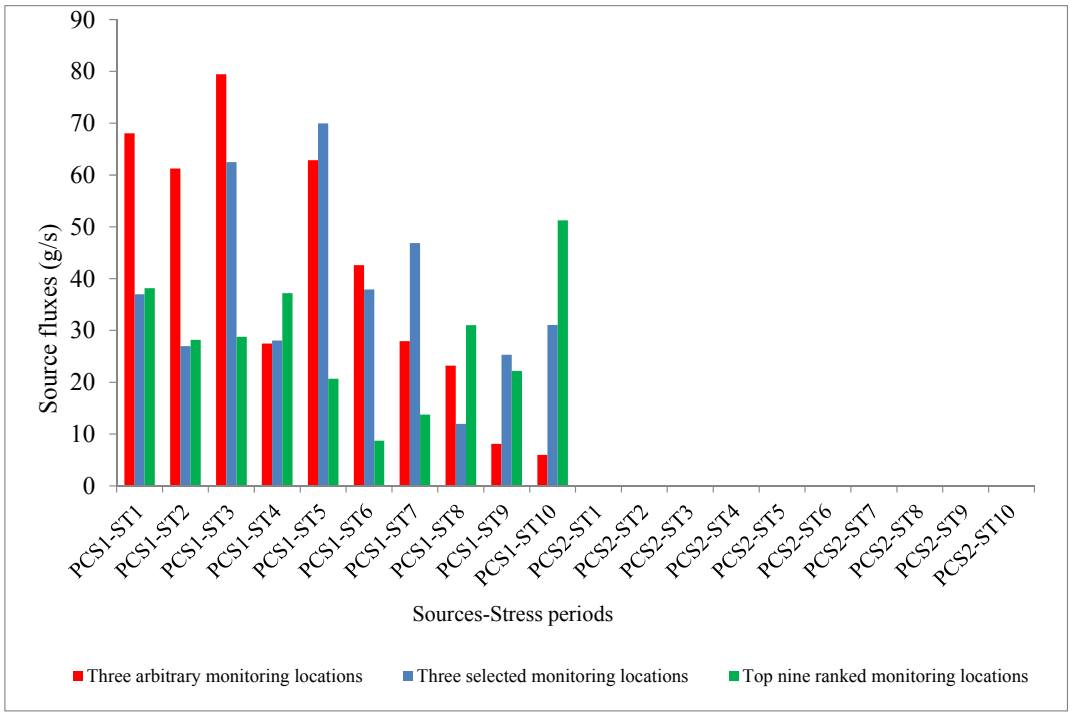



Figure 6.4 The preliminary obtained results of the developed SOM-based SMs representing different number of monitoring locations for source identification by using the observed contaminant concentration data

The performance evaluation results of the developed SOM-based SMs demonstrated that despite the capability of these types of surrogate models for screening the dummy sources, the accuracy of the developed SOM-based SMs for estimating source fluxes in terms of RMSE and NAEE were not entirely satisfactory. However, the evaluation results of the MARS-based SM for source identification of the testing data demonstrated the potential applicability of this model for source identification (Table 6.6). Therefore, the

developed MARS-based SMOs were utilized for source identification. The developed MARS-based SMOs were applied for source identification by using the observed contaminant concentration values. The obtained source identification solution results are presented in Table 6.7.

Table 6.7. The obtained source identification solutions of the developed MARS-based SMOs

| ID | PCS1(17,29,1) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
| Three arbitrary MWs | 2.8 | 0.2 | 0.5 | 6.6 | 12.0 | 6.1 | 2.9 | 1.9 | 0.2 | 1.3 |
| Top-ranked three MWs | 13.4 | 31.1 | 7.5 | 1.3 | 1.5 | 7.1 | 11.2 | 5.5 | 6.5 | 24.4 |
| Top -ranked six MWs | 17.0 | 23.5 | 4.2 | 1.3 | 6.6 | 6.9 | 20.5 | 4.7 | 0.0 | 23.1 |
| Top-ranked nine MWs | 11.9 | 28.6 | 3.6 | 1.3 | 0.5 | 12.4 | 26.6 | 4.6 | 0.0 | 14.7 |
| ID | PCS2(16,24,1) | | | | | | | | | |
| | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
| Three arbitrary MWs | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Top-ranked three MWs | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Top-ranked six MWs | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Top-ranked nine MWs | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## 6.8. Conclusion

In this chapter, the performance of the developed methodologies in a real contaminated aquifer site was evaluated. The SOM-based SMs and MARS-based SMO were developed and solved by using the same training data. The developed SOM-based SMs showed the

capability of these types of surrogate models in screening the non-active sources among all the potential contaminant sources (Figures 6.4). The SOM-based SMs could accurately screen the non-active sources among all potential contaminant sources even by using limited observed contaminant concentration data at limited times (Figure 6.4). For example, the developed SOM-based SMs could screen the dummy source(s) by using information from the three initial arbitrary monitoring wells at five times properly. The capability of the SOM algorithm in classification may make the SOM-algorithm potentially a powerful tool in the identification of unknown contaminant sources. However, the SOM-based SMs also have their disadvantages. Their performance evaluation results indicate that the accuracy of SOM-based SMs in the estimation of contaminant source fluxes in terms of NAEE was not entirely satisfactory.

The performance of the developed monitoring network design procedure in improving the accuracy of source identification results was also tested. In this contaminated aquifer, 20 monitoring locations were considered to be the potential monitoring locations. The contaminant concentration values at these monitoring wells were recorded at limited times (Figure 6.3). The obtained results of the constructed surrogate models were compared using information from the arbitrary monitoring locations and the selected monitoring locations for source identification of testing data (Table 6.6). The comparison of results demonstrated the potential applicability of this procedure for designing monitoring network. The accuracy of source identification results improved by using information from the designed monitoring locations in developing the surrogate models. The developed SOM-based SMs and MARS-based SMOs by using observed contaminant concentration data were also utilized for source identification. The performance evaluation results demonstrate potential applicability of these models for source identification in a contaminated aquifer site (Figures 6.4 and Table 6.7).

In the next chapter, a summary of the developed methodologies and their application to different contaminated sites is presented. Main conclusions and limitations of the developed methodologies in this study are also discussed in the next chapter.

# 7. Summary and Conclusions

## 7.1. Introduction

In this chapter, the developed methodologies are summarized. Then, the conclusions of this study are presented, including limitations of the developed procedures for source identification and monitoring network design based on the performance evaluation results. Finally, a possible scope for future research is highlighted.

## 7.2. Summary

Three different surrogate models for comparison purposes were developed for source identification. Self-Organising Maps (SOM), Gaussian Process Regression (GPR) and Multivariate Adaptive Regression Splines (MARS) algorithms were utilized to design different surrogate models. The developed models accurately could mimic the complex processes of groundwater flow and transport in contaminated aquifer sites. These surrogate models could also characterize contaminant sources in terms of contaminant source locations, magnitudes and release history. The important feature of these developed surrogate models is that unlike the previous methods, this source identification methodology can be applied independently of any linked optimisation model solution. In other words, the developed surrogate models are capable of directly identifying unknown groundwater contaminant sources.

However, in this study for comparison purposes, Surrogate Models-based Optimisation (SMO) models were also developed. MARS and Genetic Algorithm (GA) were utilized as the surrogate model and optimisation model types, respectively in the developed SMOs.

The SOM algorithm was selected as the surrogate model type because of its capabilities in classifying nonlinear multidimensional input data. The GPR and MARS algorithms

were also utilized as the other types of the surrogate model because they can reveal the relationships of high dimensional input data. Therefore, different developed surrogate models using these three algorithms were utilized for source identification in different contaminated aquifer site.

To improve the source identification results, the possibility of applying adaptive strategies such as sequential sampling method based on the preliminary solution results was also utilized and evaluated (Chapter 3). A monitoring network design procedure was also utilized (Chapters 4 and 6). The information from the designed monitoring network was utilized to develop new surrogate models. The source identification results of these new surrogate models were compared with the source identification results of the surrogate models developed by using information from arbitrary monitoring locations.

Information from four different study areas was utilized in this study to assess the performance of the developed procedures:

1. In the first case, the performance of the developed SOM, MARS, and GPR-based Surrogate Models (SOM, MARS, and GPR-based SMs) which independently could be utilized for source identification was evaluated. These surrogate models were utilized to characterize unknown groundwater contaminant sources, for an ideal scenario of error-free concentration data, as well as scenarios with different degrees of erroneous concentration measurements data.
2. In the second case, MARS algorithm was utilized to develop MARS-based SMO. In this SMO, a GA based optimisation model was utilized. The MARS-based SMs was utilized for source identification in a heterogeneous multi-layered illustrative contaminated aquifer site. In this study area, the performance of the developed MARS-based SMO was evaluated by using deterministic hydraulic conductivity values, and uncertain hydraulic conductivity values. To improve the source

identification results, a monitoring network design procedure was also utilized. Three different data mining tools were utilized to develop the monitoring network design procedure. These tools were Random Forests (RF), Classification and Regression Trees (CART), and Tree Net (TN). The main reason for selecting these tools was their capabilities in recognising the most important components of prediction models. The source identification results by using data from the designed and arbitrary monitoring locations were compared in this contaminated aquifer site.

3. In the third case, the performance of SOM, MARS, and GPR-based SMs was evaluated for source identification in an experimental contaminated aquifer site within the heterogeneous sand aquifer in Australia. MARS-based SMO was also developed for comparison purpose. In this study area, the measured contaminant source fluxes and hydraulic conductivity values were not error free.
4. In the last case, the SOM-based SMs and MAR-based SMO were developed and utilized to identify unknown contaminant sources in a contaminated aquifer site in Australia. In this case, limited and sparse data were available. In this study area, the developed monitoring network design procedure was also utilized to select the monitoring locations that have the most expected contributions to source identification.

## 7.3. Conclusions

The main conclusions from the limited performance evaluation results that can be derived are:

1. The developed SOM, MARS, and GPR-based SMs could accurately mimic the complex flow and transport processes in contaminated aquifers.

2. The developed surrogate models could independently provide a procedure for source identification without the necessity of using a linked simulation-optimisation model.
3. The capabilities of the SOM algorithm in clustering and finding missing values of multidimensional input data may make the SOM algorithm a potentially useful tool for the unknown contaminant source identification problems. For example, SOM algorithm capability in clustering leads the SOM-based SMs to screen the dummy sources, i.e., not actual sources but included as potential sources precisely.
4. By using SOM algorithm as a surrogate model type for source identification and providing the answer of one of the main questions of source identification problems, identifying the contaminant source locations, fewer decision variables would be needed to develop the surrogate models.
5. Selecting proper numbers of variables at the initial sampling stage of developing the surrogate models plays an important role in the solution results. Therefore, designing a monitoring network could improve source identification results.
6. Applying the designed monitoring network procedure in conjunction with source identification methodology improved source identification results.
7. Since the optimal number of the SOM map units relates to the memory of the PC utilized and the initial sample sizes, the number of SOM map units is an important issue in the SOM-based SMs.
8. The initial sample size for training the surrogate models has an important role in the accuracy of solution results. The other important issue about the initial sample size is that the training data should properly cover the whole possible ranges of potential contaminant concentration values.

9. The performance evaluation results of the developed surrogate models for different hypothetical and real field data demonstrate the potential applicability of the developed methodologies in source identification.
10. The potential to easily implement the developed surrogate models for source identification is one of the advantages of the developed surrogate models. However, the required steps need to be followed, which are unlike the Simulation-optimization approach based on solving a linked optimization model.
11. The performance evaluation results of the designed monitoring network also showed the potential applicability of the developed monitoring network design procedure. This procedure, by using the capabilities of three data mining tools in identifying the most relevant concentration measurements for source identification, could design efficient monitoring networks without being computationally intensive.

However, the developed methodologies also have some limitations as follows:

1. The main limitation of the developed SOM-based SMs is that the evaluation results showed comparatively large errors in terms of the specific error criteria utilized. However, a comparison of the source estimates and the actual values shows a better match when these values are directly compared instead of using the error statistics.
2. Achieving computational efficiency is one of the goals of this study. It is estimated that computing costs and time associated with repeated runs of the simulation models within the optimization algorithm were reduced by applying surrogate model techniques. However, by considering all the computational times

need for designing experiments, preparing training data and developing surrogate models, computational efficiency was not achieved to the extent desired.

3. Developing MARS-based SMOs by using another advanced optimization algorithms such as Adaptive Simulated Annealing (ASA) may improve source identification results.
4. The contaminants were considered to be conservative so the developed surrogate models need to be extended to consider non-conservative contaminations.
5. One limitation of the proposed use of surrogate models in inverse mode for source identification is that the time for initial activity of the sources relative to the first recorded concentration measurement needs to be known, although the SMO based models are capable of solving the source identification problem even if the source activity initiation times are not known.
6. The presented performance evaluation results for the developed procedures for source identification are limited in scope. Therefore, to establish the applicability of the developed procedures, more rigorous performance evaluations need to be utilized.
7. The presented performance evaluation results for the developed methodology for monitoring network design are based on very limited scenarios and therefore restricted in scope. Further performance evaluations are required to fully establish the applicability of the developed methodology.

## 7.4. Recommendations for Future Research

An alternative application of the SOM algorithm was introduced in this study. By developing the SOM-based SMs, the capabilities of this algorithm in classification were utilized to screen non-active sources in source identification problems. The application

of the SOM-based SMs can be tested and possibly extended for other complex source identification problems or other similar problems.

The application of the developed surrogate models could be evaluated for non-conservative contaminations occurring in many contaminated aquifer sites.

## References:


Ala, N. K., & Domenico, P. A. (1992). Inverse Analytical Techniques Applied to Coincident Contaminant Distributions at Otis-Air-Force-Base, Massachusetts. *Ground Water, 30*(2), 212-218. doi:DOI 10.1111/j.1745-6584.1992.tb01793.x

Amauri, H., Júnior, S., Barreto, G. A., & Corona, F. (2015). Regional models: A new approach for nonlinear system identification via clustering of the self-organizing map. *Neurocomputing, 147*, 16. doi:10.1016/j.neucom.2013.11.046

Amir Abdollahian, M. (2016). *Development of Integrated Methodologies for Optimal Monitoring and Source Characterization in Contaminated Groundwater Systems Under Uncertainty.* (PhD), James Cook University.

Amirabdollahian, M., & Datta, B. (2013). Identification of Contaminant Source Characteristics and Monitoring Network Design in Groundwater Aquifers: An Overview. *Journal of Environmental Protection, 04*(05), 16. doi:10.4236/jep.2013.45A004

Amirabdollahian, M., & Datta, B. (2014). Identification of Pollutant Source Characteristics Under Uncertainty in Contaminated Water Resources Systems Using Adaptive Simulated Anealing and Fuzzy Logic. *Int. J. of GEOMATE, 6*, 757-762.

Amirabdollahian, M., & Datta, B. (2015). Reliability Evaluation of Groundwater Contamination Source Characterization under Uncertain Flow Field. *International Journal of Environmental Science and Development, 6*(7), 512-518. doi:10.7763/ijesd.2015.v6.647

Aral, M. M., Guan, J. B., & Maslia, M. L. (2001). Identification of contaminant source location and release history in aquifers. *Journal of Hydrologic Engineering, 6*(3), 225-234. doi:Doi 10.1061/(Asce)1084-0699(2001)6:3(225)

Atmadja, J., & Bagtzoglou, A. (2001). State of the Art Report on Mathematical Methods for Groundwater Pollution Source Identification. *Environmental Forensics, 2*(3), 205-214. doi:10.1006/enfo.2001.0055

Atmadja, J., & Bagtzoglou, A. C. (2001). Pollution source identification in heterogeneous porous media. *WATER RESOURCES RESEARCH, 37*(8), 2113-2125. doi:Doi 10.1029/2001wr000223

Ayvaz, M. T. (2010). A hybrid simulation–optimization approach for solving the areal groundwater pollution source identification problems. *Journal of Hydrology, 538*, 16. doi:10.1016/j.jhydrol.2016.04.008

Bagtzoglou, A. C., & Atmadja, J. (2005). Mathematical Methods for Hydrologic Inversion: The Case of Pollution Source Identification. *Handb Environ Chem, 3*, 65-96. doi:10.1007/b11442

Bagtzoglou, A. C., Dougherty, D. E., & Tompson, A. F. B. (1992). Application of Particle Methods to Reliable Identification of Groundwater Pollution Sources. *Water Resources Management, 6*, 9.

Bagtzoglou, A. C., Tompson, A. F. B., & Dougherty, D. E. (1991). Probabilistic simulation for reliable solute source identification in heterogenous poroua media. from Springer-Verlag Berlin Heidelberg

Barbariol, F., Marcello Falcieri, F., Scotton, C., Benetazzo, A., Carniel, S., & Sclavo, M. (2016). Wave extreme characterization using self-organizing maps. *Ocean Sci, 12*, 13. doi:10.5194/os-12-403-2016

Bashi-Azghadi, S. N., Kerachian, R., Bazargan-Lari, M. R., & Solouki, K. (2010). Characterizing an unknown pollution source in groundwater resources systems using PSVM and PNN. *Expert Systems with Applications, 37*, 8.

Beck, p. H. (2000). *Transport of conservative and reactive inorganic elements in the saturated part of a hetrogeneous sand aquifer, Botany Basin, Sydney, Australia.* (PhD), University of New South Wales, Sydney, Australia.

Belyaev, M., Burnaev, E., Kapushev, E., Panov, M., Prikhodko, P., Vetrov, D., & Yarotsky, D. (2016). GTApprox: surrogate modeling for industrial design.

Bhattacharjya, R. K., & Datta, B. (2005). Optimal Management of Coastal Aquifers Using Linked Simulation Optimization Approach. *Water Resources Management, 19*, 26. doi:10.1007/s11269-005-3180-9

Boman, G. K., Molz, F. J., & Guven, O. (1995). An Evaluation of Interpolation Methodologies for Generating Three-Dimensional Hydraulic Property Distribution from Measured Data. *Ground Water, 33*, 12.

Borah, T., & Bhattacharjya, R. K. (2014). Development of Unknown Pollution Source Identification Models Using GMS ANN–Based Simulation Optimization Methodology. *Journal of Hazardous, Toxic, and Radioactive Waste*, 12. doi:10.1061/(ASCE)HZ.2153-5515.0000242

Bullinaria, J. A. (2004a). Self Organizing Maps: Algorithms and Applications (Introduction to Neural Networks: lecture 17).

Bullinaria, J. A. (2004b). Self Organizing Maps: Fundamentals (introduction to Neural Networks: lecture 16).

Bullinaria, J. A. (2014). Self Organizing maps: Properties and applications (Neural Computation: lecture 17).

Chadalavada, S., & Datta, B. (2008). Dynamic optimal monitoring network design for transient transport of pollutants in groundwater aquifers. *Water Resources Management, 22*(6), 20. doi:10.1007/s11269-007-9184-x

Chadalavada, S., Datta, B., & Naidu, R. (2011a). Optimisation approach for pollution source identification in groundwater: an overview. *Int. J. Environment and Waste Management, 8*, 22.

Chadalavada, S., Datta, B., & Naidu, R. (2011b). Uncertainty based optimal monitoring network design for a chlorinated hydrocarbon contaminated site. *Environ Monit Assess, 173*, 12. doi:10.1007/s10661-010-1435-2

Chadalavada, S., Datta, B., & Naidu, R. (2012). Optimal Identification of Groundwater Pollution Sources Using Feedback Monitoring Information: A Case Study. *Environmental Forensics, 13*(2), 14. doi:10.1080/15275922.2012.676147

Chalasani, R., & Principe, J. C. (2015). Self-organizing maps with information theoretic learning. *Neurocomputing, 147*, 12. doi:10.1016/j.neucom.2013.12.059

Cleveland, T. G., & Yeh, W. W.-G. (1991). Optimal configuration and scheduling of Ground-Water tracer test. *Journal of Water Resource Planning and Management, 117*, 37-51.

Datta, B. (2002). Discussion of ‘‘Identification of Contaminant Source Location and Release History in Aquifers’’ by Mustafa M. Aral, Jiabao Guan, and Morris L. Maslia. *J. Hydrol. Eng., 7*(5), 3.

Datta, B., Beegle, J. E., Kavvas, M. L., & Orlob, G. T. (1989). *Development of an expert system embedding pattern recognition techniques for pollution source identification*. Retrieved from

Datta, B., Chakrabarty, D., & Dhar, A. (2009). Optimal Dynamic Monitoring Network Design and Identification of Unknown Groundwater Pollution Sources. *Water Resources Management, 23*(10), 2031-2049. doi:10.1007/s11269-008-9368-z

Datta, B., Chakrabarty, D., & Dhar, A. (2011). Identification of unknown groundwater pollution sources using classical optimization with linked simulation. *Journal of Hydro-environment Research, 5*(1), 25-36. doi:10.1016/j.jher.2010.08.004

Datta, B., & Dhiman, S. (1996). Chance-Constrained Optimal Monitoring Network Design for Pollutants in Ground Water. *JOURNAL OF WATER RESOURCES PLANNING AND MANAGEMENT, 122*(3), 8. doi:10.1061/(ASCE)0733-9496(1996)122:3(180)

Datta, B., & Kourakos, G. (2015). Preface: Optimization for groundwater characterization and management. *Hydrogeology Journal, 23*(6), 1043-1049. doi:10.1007/s10040-015-1297-3

Datta, B., Prakash, O., Campbell, S., & Escalada, G. (2013). Efficient Identification of Unknown Groundwater Pollution Sources Using Linked Simulation-Optimization Incorporating Monitoring Location Impact Factor and Frequency Factor. *Water Resources Management, 27*(14), 18. doi:10.1007/s11269-013-0451-8

Dhar, A., & Datta, B. (2007). Multiobjective design of dynamic monitoring networks for detection of groundwater pollution. *Journal of Water Resources Planning and Management-Asce, 133*(4), 10. doi:10.1061/(Asce)0733-9496(2007)133:4(329)

Dhar, A., & Datta, B. (2010). Logic-Based Design of Groundwater Monitoring Network for Redundancy Reduction. *Journal of Water Resources Planning and Management-Asce, 136*(1), 7. doi:10.1061/(Asce)0733-9496(2010)136:1(88)

Di Mauro, M., Maggioni, M. F., Grasso, M., & Colosimo, B. M. (2016). *Design performance analysis of a Self-Organizing Map for statistical monitoring of distribution-free data streams*. Paper presented at the CIRP CMS 2015.

Dokou, Z., & Pinder, G. F. (2009). Optimal search strategy for the definition of a DNAPL source. *Journal of Hydrology, 376*(3-4), 542-556. doi:10.1016/j.jhydrol.2009.07.062

Dragomir, O. E., Dragomir, F., & Radulescu, M. (2014). Matlab Application of Kohonen Self-organizing Map to Classify Consumers' Load Profiles. *Procedia Computer Science, 31*, 6. doi:10.1016/j.procs.2014.05.292

EPA., U. (2005). RoadMap to long-term Monitoring Optimization. *US EPA*(542-R-05-003).

Forrester, A. I. J., & Keane, A. J. (2009). Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences, 45*(1-3), 50-79. doi:10.1016/j.paerosci.2008.11.001

Freeze, R. A. (1975). A Stochastic-Conceptual Analysis of One-Dimensional Groundwater Flow in Nonuniform Homogeneous Media. *WATER RESOURCES RESEARCH, 11*, 17.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines *The Annuals of Statistics, 19*, 67.

Gaussian Process Regression. (2017).

Gong, W., & Duan, Q. (2017). An adaptive surrogate modeling-based sampling strategy for parameter optimization and distribution estimation... *Environmental Modelling & Software, 95*, 16. doi:10.1016/j.envsoft.2017.05.005

Gorelick, S. M., Evans, B., & Remson, I. (1983). Identifying Sources of Groundwater Pollution - an Optimization Approach. *WATER RESOURCES RESEARCH, 19*(3), 779-790. doi:DOI 10.1029/WR019i003p00779

Gorissen, D., Couckuyt, I., Demeester, P., Dhaene, T., & Crombecq, K. (2010). A Surrogate Modeling and Adaptive Sampling Toolbox for Computer Based Design. *Journal of Machine Learning Research, 11*, 5.

Harbaugh, A. W. (2005). *MODFLOW-2005, The U.S. Geological Survey Modular Ground-Water Model-the Ground-Water Flow Process*: U.S. Geological Survey Techniques and Methods 6–A16.

Harrington, N., & Cook, P. (2014). *Groundwater in Australia*. Retrieved from National Centre for Groundwater Research and Training, Australia:

Hazrati-Yadkoori, S., & Datta, B. (2017). Adaptive Surrogate Model Based Optimization (ASMBO) for Unknown Groundwater Contaminant Source Characterizations Using Self-Organizing Maps. *Journal of Water Resource and Protection, 9*, 23. doi:10.4236/jwarp.2017.92014

He, L., Huang, G. H., & Lu, H. W. (2009). A coupled simulation-optimization approach for groundwater remediation design under uncertainty: an application to a petroleum-contaminated site. *Environ Pollut, 157*(8-9), 2485-2492. doi:10.1016/j.envpol.2009.03.005

Jankowski, J., & Beck, p. (2010). Aquifer heterogeneity: Hydrogeological and hydrochemical properties of the Botany Sands aquifer and their impact on contaminant transport. *Australian Journal of Earth Sciences, 47(1)*, 20. doi:10.1046/j.1440-0952.2000.00768.x

Jha, M., & Datta, B. (2012). *Simulated annealing based simulation-optimization approach for identification of unknown contaminant sources in groundwater aquifers*. Paper presented at the Third International Conference on Challenges in Environmental Science & Engineering, The Sebel, Cairns, Queensland, Australia.

Jha, M., & Datta, B. (2013). Three-Dimensional Groundwater Contamination Source Identification Using Adaptive Simulated Annealing. *Journal of Hydrologic Engineering, 18*(3), 307-317. doi:10.1061/(Asce)He.1943-5584.0000624

Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization, 21*(4), 345-383. doi:Doi 10.1023/A:1012771025575

K. Esfahani, H., & Datta, B. (2016). Linked Optimal Reactive Contaminant Source Characterization in Contaminated Mine Sites: Case Study. *Journal of Water Resource Planning and Management, 142*, 14. doi:10.1061/(ASCE)WR.1943-5452.0000707

Kohonen, T., & Oja, E. (2001). *Self-Organizing Maps* T. S. Huang, T. Kohonen, & M. R. Schroeder (Eds.), doi:10.1007/978-3-642-56927-2

Kohonen, T., Oja, E., Simula, O., Visa, A., & Kangas, J. (1996). Engineering Applications of the Self-Organizing Map. *IEEE, 84(10)*, 1358-1384. doi:0018-9219(96)07176-9

Koziel, S., Ciaurri, D. E., & Leifsson, L. (2011). Chapter 3: Surrogate-Based Methods. *Springer-Verlag Berlin Heidelberg, 356*, 27.

Le Thi, H. A., & Nguyen, M. C. (2014). Self-organizing maps by difference of convex functions optimization*28*, 30. Retrieved from doi:10.1007/s10618-014-0369-7

Lee, Y. M., & Ellis, J. H. (1996). Comparison of algorithms for nonlinear integer optimization: Application to monitoring network design. *Journal of Environmental Engineering-Asce, 122*(6), 524-531. doi:Doi 10.1061/(Asce)0733-9372(1996)122:6(524)

Liu, C. X., & Ball, W. P. (1999). Application of inverse methods to contaminant source identification from aquitard diffusion profiles at Dover AFB, Delaware. *WATER RESOURCES RESEARCH, 35*(7), 1975-1985. doi:Doi 10.1029/1999wr900092

Loaiciga, H. A., Charbeneau, R. J., Everett, L. G., Fogg, G. E., Hobbs, B. F., & Rouhani, S. (1992). Review of Ground-water quality monitoring network design. *Journal of Hydraulic Engineering, 118*, 11-37.

Luo, Q., Wu, J., Yang, Y., Qian, J., & Wu, J. (2016). Multi-objective optimization of long-term groundwater monitoring network design using a probabilistic Pareto genetic algorithm under uncertainty. *Journal of Hydrology, 534*, 12.

Mahar, P. S., & Datta, B. (1997). Optimal monitoring network and ground-water-pollution sources identification. *Journal of Water Resource Planning and Management, 123* (4), 199.

Mahar, P. S., & Datta, B. (2000). Identification of pollution sources in transient groundwater systems. *Water Resources Management, 14*(3), 19. doi:Doi 10.1023/A:1026527901213

Mahar, P. S., & Datta, B. (2001). Optimal identification of ground-water pollution sources and parameter estimation. *Journal of Water Resources Planning and Management-Asce, 127*(1), 10. doi:Doi 10.1061/(Asce)0733-9496(2001)127:1(20)

Mahinthakumar, G. K., & Sayeed, M. (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. doi:Doi 10.1061/(Asce)0733-9496(2005)131:1(45)

Mahinthakumar, G. K., & Sayeed, M. (2006). Reconstructing groundwater source release histories using hybrid optimization approaches. *Environmental Forensics, 7*(1), 45-54. doi:10.1080/15275920500506774

Mugunthan, P., & Shoemaker, C. A. (2010). Time Varying Optimization for Monitoring Multiple Contaminants under Uncertain Hydrogeology. *Bioremediation Journal, 8*(3-4), 129-146. doi:10.1080/10889860490887509

Neupauer, R. M., Borchers, B., & Wilson, J. L. (2000). Comparison of inverse methods for reconstructing the release history of a groundwater contamination source. *WATER RESOURCES RESEARCH, 36*(9), 2469-2475. doi:Doi 10.1029/2000wr900176

Neupauer, R. M., & Wilson, J. L. (1999). Adjoint method for obtaining backward-in-time location and travel time probabilities of a conservative groundwater contaminant. *WATER RESOURCES RESEARCH, 35*(11), 3389-3398. doi:Doi 10.1029/1999wr900190

NGWA. (2016). Facts About Global Groundwater Usage. National Ground Water Association, U.S.A.

Norouzi, K., & Rakhshandehroo, G. R. (2011). A Self-Organizing Map based Hybrid Multi-Objective Optimum Of Water Distribution Networks. *IJST, Transactions of Civil and Enviromental Engineering, 35*, 105-119.

Pinder, G. F., Ross, J., & Dokou, Z. (2009). *Optimal search strategy for the definition of a DNAPL source*. Retrieved from <Go to ISI>://WOS:000271165600018

Prakash, O. (2014). *Optimal Monitoring Network Design and Identification of Unknown pollutant Sources in Polluted Aquifers.* (PhD), James Cook University, James Cook University.

Prakash, O., & Datta, B. (2013). Multiobjective Monitoring Network Design for Efficient Identification of Unknown Groundwater Pollution Sources Incorporating Genetic Programming–Based Monitoring. *Journal of Hydrologic Engineering, 19*(11), 04014025. doi:10.1061/(asce)he.1943-5584.0000952

Prakash, O., & Datta, B. (2014). Characterization of Groundwater Pollution Sources with Unknown Release Time History. *Journal of Water Resource and Protection, 6*, 14. doi:10.4236/jwarp.2014.64036

Prakash, O., & Datta, B. (2015). Optimal characterization of pollutant sources in contaminated aquifers by integrating sequential-monitoring-network design and source identification: methodology and an application in Australia. *Hydrogeology Journal, 23*(6), 1089-1107. doi:10.1007/s10040-015-1292-8

Queipo, N. V., Haftka, R. T., Shyy, W., Goel, T., Vaidyanathan, R., & Tucker, P. K. (2005). Surrogate-based analysis and optimization. *Progress in Aerospace Sciences, 41*(1), 1-28. doi:10.1016/j.paerosci.2005.02.001

Random Forest for Beginners. (2014) In S. Systems (Series Ed.), (pp. 71).

Razavi, S., Tolson, B. A., & Burn, D. H. (2012). Review of surrogate modeling in water resources. *WATER RESOURCES RESEARCH, 48*, 32. doi:10.1029/2011wr011527

Reed, P. M., & Minsker, B. S. (2004). Striking the Balance: Long-Term Groundwater Monitoring Design for Conflicting Objectives. *J. Water Resour. Plann. Manage, 130*, 10. doi:10.1061/~ASCE!0733-9496~2004!130:2~140!

Retherford, J. Q., & McDonald, M. (2010). *Estimation and Validation of Gaussian Process Surrogate Models for MEPDG-Based Sensitivity Analysis and Design Optimization*. Retrieved from Transportation Research Board Annual Meeting:

Salford Predictive Modeller 8 (2017).

Schulz, E., Speekenbrink, M., & Krause, A. (2016). A tutorial on Gaussian process regression with a focus on exploration-exploitation scenarios (Publication no. 10.1101/095190).

Sidauruk, P., Cheng, A. H. D., & Ouazar, D. (1998). Ground water contaminant source and transport parameter identification by correlation coefficient optimization. *Ground Water, 36*(2), 208-214. doi:DOI 10.1111/j.1745-6584.1998.tb01085.x

Simula, O., Vesanto, J., Alhoniemi, E., & Hollmen, J. (1999). Analysis and Modeling of Complex systems Using the Self-Organizing Map. 16.

Singh, R. M., & Datta, B. (2006). Identification of groundwater pollution sources using GA-based linked simulation optimization model. *Journal of Hydrologic Engineering, 11*(2), 9. doi:10.1061/(Asce)1084-0699(2006)11:2(101)

Singh, R. M., & Datta, B. (2007). Artificial neural network modeling for identification of unknown pollution sources in groundwater with partially missing concentration observation data. *Water Resources Management, 21*(3), 557-572. doi:10.1007/s11269-006-9029-z

Singh, R. M., Datta, B., & Jain, A. (2004). Identification of unknown groundwater pollution sources using artificial neural networks. *Journal of Water Resources Planning and Management-Asce, 130*(6), 9. doi:10.1061/(Asce)0733-9496(2004)130:6(506)

Skaggas, T. H., & Kabala, Z. J. (1994). Recoveringt the release history of a groundwater contaminant. *WATERR ESOURCREESS EAR*.

Skaggas, T. H., & Kabala, Z. J. (1998). Limitations in recovering the history of a groundwater contaminant plume. *Journal of Contaminant Hydrology*.

Snodgrassa, M. F., & Kitanidis, P. K. (1997). A geostatisticaal pproach to contaminant source identification. *WATER RESOURCES RESEARCH*.

SPM User Guide, Introducing CART. (2013).

SPM User Guide, Introducing Tree Net. (2013). In S. Systems (Ed.).

SPM User Guide, Introduction to MARS. (2013): Salford Systems.

SPM User Guide, Introduction to Random Forests. (2012), (pp. 42).

Sreekanth, J., & Datta, B. (2010). Multi-objective management of saltwater intrusion in coastal aquifers using genetic programming and modular neural network based surrogate models. *Journal of Hydrology, 393*, 12. doi:10.1016/j.jhydrol.2010.08.023

Sun, A. Y., Painter, S. L., & Wittmeyer, G. W. (2006a). A constrained robust least squares approach for contaminant release history identification. *WATER RESOURCES RESEARCH, 42*(4), 1-13. doi:Artn W04414

10.1029/2005wr004312
Sun, A. Y., Painter, S. L., & Wittmeyer, G. W. (2006b). A robust approach for iterative contaminant source location and release history recovery. *Journal of Contaminant Hydrology, 88*(3-4), 181-196. doi:10.1016/j.jconhyd.2006.06.006
Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications.* (Master ), Humboldt University.
Vatanen, T., Osmala, M., Raiko, T., Lagus, K., Sysi-Aho, M., Orešič, M., . . . Lähdesmäki, H. (2015). Self-organization and missing values in SOM and GTM. *Neurocomputing, 147*, 10. doi:10.1016/j.neucom.2014.02.061
Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). *SOM Toolbox for Matlab 5*. Retrieved from http://www.cis.hut.fi/projects/somtoolbox:
Wagner, B. J. (1992). Simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport
modelling. *Journal of Hydrology*.
Walter, J., Ritter, H., & Schulten, K. (1990). Non-Linear prediction with Self-Organizing Maps. I-589-594.
Wang, C., Duan, Q. Y., Gong, W., Ye, A. Z., Di, Z. H., & Miao, C. Y. (2014). An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. *Environmental Modelling & Software, 60*, 167-179. doi:10.1016/j.envsoft.2014.05.026
Whigham, P. A. (2005). Local Modelling by SOM partitioning and linear regression for Ecological Modelling. *Modsim 2005: International Congress on Modelling and Simulation: Advances and Applications for Management and Decision Making*, 1319-1325.
Woodbury, A., Sudicky, E., Ulrych, T. J., & Ludwig, R. (1998). Three-dimensional plume source reconstruction using minimum relative entropy inversion. *Journal of Contaminant Hydrology, 32*(1-2), 131-158. doi:Doi 10.1016/S0169-7722(97)00088-0
Wu, J., Zheng, C., & Chien, C. C. (2005). Cost-effective sampling network design for contaminant plume monitoring under general hydrogeological conditions. *J Contam Hydrol, 77*(1-2), 41-65. doi:10.1016/j.jconhyd.2004.11.006
Yeh, William W-G. (1986). Review of Parameter Identification Procedures in Groundwater Hydrology: The Inverse Problemm. *WATER RESOURCES RESEARCH, 22*(2), 14.
Yeh, H.-D., Chang, T.-H., & Lin, Y.-C. (2007). Groundwater contaminant source identification by a hybrid heuristic approach. *WATER RESOURCES RESEARCH, 43*(9), 1-16. doi:10.1029/2005wr004731
Zhang, W., & T.C. Goh, A. (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers, 7*, 8. doi:10.1016/j.gsf.2014.10.003
Zheng, C., & Wang, P. P. (1999). *MT3DMS: A Modular Three-Dimensional Multispecies Transport Model for Simulation of Advection, Dispersion, and Chemical Reactions of Contaminants in Groundwater Systems; Documentation and User's Guide* (pp. 220).