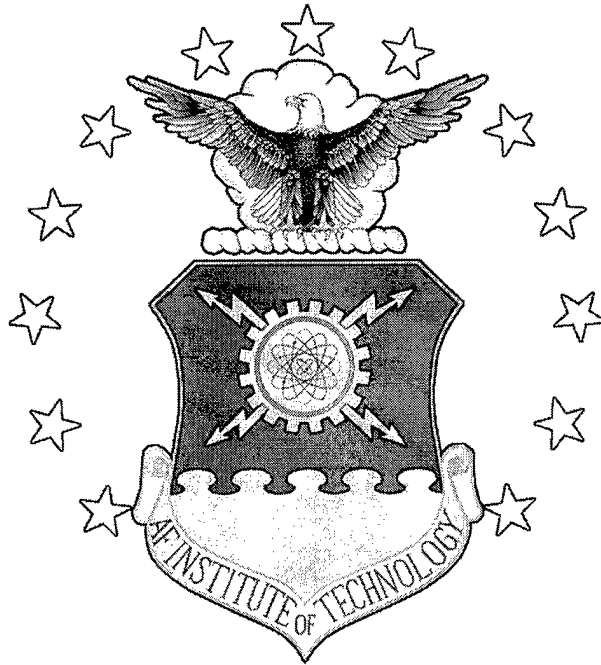




**DETERMINING THE VALUE OF GROUNDWATER
CONTAMINATION SOURCE REMOVAL: A METHODOLOGY**
THESIS

Benjamin C. Recker, Second Lieutenant, USAF

AFIT/GEE/ENV/01M-15

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

20010612 154

Disclaimer Statement

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense or U.S. Government.

AFIT/GEE/ENV/01M-15

DETERMINING THE VALUE OF GROUNDWATER CONTAMINATION SOURCE
REMOVAL: A METHODOLOGY

THESIS

Presented to the Faculty

Department of Systems and Engineering Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the
Degree of Master of Science in Engineering and Environmental Management

Benjamin C. Recker, B.S.

Second Lieutenant, USAF

March, 2001

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED.

DETERMINING THE VALUE OF GROUNDWATER CONTAMINATION SOURCE
REMOVAL: A METHODOLOGY

Benjamin C. Recker, B.S.
Second Lieutenant, USAF

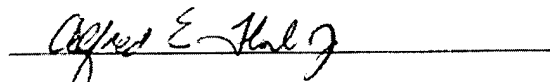
Approved:



Dr. Mark N. Goltz
Chair, Advisory Committee

5 Mar 2001

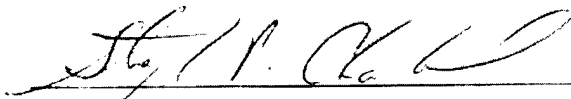
Date



Lieutenant Colonel Alfred E. Thal, Jr.
Member, Advisory Committee

5 MAR 01

Date



Captain Stephen P. Chambal
Member, Advisory Committee

05 MAR 01

Date

Acknowledgements

I owe a tremendous debt of gratitude to several people who were instrumental to the completion of this thesis. First of all, my thesis advisor, Dr. Mark Goltz, provided endless support, encouragement, feedback, and focus throughout the thesis process. Without his efforts, this thesis would not be what it is today. I am also extremely appreciative of the efforts of my other committee members, Lt Col Thal and Capt Chambal. Their guidance, advice, encouragement and expertise were an invaluable asset to me throughout the completion of this thesis.

I must also thank my beautiful wife, whose love, support, and understanding, were instrumental in seeing me through the good times and bad of my AFIT experience. You were always there when I needed you. You will never know how important you were in helping me reach this destination. I must also thank our son, who was always there to put a smile on my face and remind me what pure joy can mean. I must also thank the other members of my family who supported me over the past year. I greatly appreciate everything that you have done for me.

Benjamin C. Recker

Table of Contents

Section	Page
Acknowledgements.....	iv
List of Figures.....	vii
List of Tables.....	viii
Abstract	ix
1.0 Introduction	1
1.1 Background	1
1.2 Remediation Strategies.....	2
1.3 Problem Statement	6
1.4 Research Objective.....	7
1.5 General Research Approach.....	7
1.6 Scope and Limitations of Research.....	8
2.0 Literature Review.....	10
2.1 Overview	10
2.2 Current Work in Source Removal Valuation	10
2.3 Site Characterization	12
2.3.1 Properties of the Aquifer	13
2.3.2 Properties of the Contaminant	15
2.3.3 Remedial Objectives.....	17
2.4 Analytical Techniques for Reducing Field Data.....	18
2.4.1 Cluster Analysis.....	19
2.4.1.1 Distance-Type Measures	21
2.4.1.2 Matching-Type Measures.....	22
2.4.1.3 Distance-Type to Matching-Type Conversion.....	25
2.4.1.4 Using Measures to Cluster Objects	27
2.4.1.4.1 Hierarchical Techniques	27
2.4.1.4.2 Partitioning Methods.....	33
2.4.1.5 Stopping Rules	35
2.4.2 Discriminant Analysis	37
2.4.3 Classification Using Decision Trees	40
2.4.4 Case Studies of Application of Analytical Techniques to Group Objects	41
2.4.4.1 Cluster Analysis Example: Cluster Analysis of Environmental Data which are not Interval Scaled but Categorical	42
2.4.4.2 Decision Tree Classification: Classifying Environmental Pollutants 1: Structure-Activity Relationships for Prediction of Aquatic Toxicity	45

Section	Page
3.0 Methodology	47
3.1 Overview	47
3.2 Data Management	47
3.2.1 Data Collection/Parameter Selection.....	47
3.2.2 Data Gap Management	49
3.3 Grouping the Data	52
3.3.1 Grouping Technique Selection	52
3.3.2 Similarity Coefficient Determination	53
3.3.3 Clustering Technique Selection.....	54
3.3.4 Specifics of Cluster Analysis for this Study.....	56
3.3.5 Cluster Analysis Example	61
3.4 Using Clustered Data to Create Lifecycle Cost versus Percent Source Removal Plots.....	69
4.0 Results	71
4.1 Overview	71
4.2 Cluster Analysis	71
4.2.1 Cluster 1.....	72
4.2.2 Cluster 2.....	74
4.3 Cost Versus Percent Removal	76
5.0 Conclusions	83
5.1 Summary	83
5.2 Lifecycle Cost versus Percent Source Removal Plots.....	83
5.3 Utility of Cluster Analysis	84
5.4 Limitations	86
5.5 Recommendations for Future Study.....	87
BIBLIOGRAPHY	91
Appendix A	A-1
Appendix B	B-1
Appendix C	C-1
Vita.....	V-1

List of Figures

Figure	Page
Figure 1.1 Kavanaugh and Goldstein Conceptualization (Kavanaugh and Goldstein, 1999).....	6
Figure 2.1. Estimated Cleanup Time Less Than 100 Years (Kavanaugh and Goldstein, 1999).....	11
Figure 2.2: Estimated Cleanup Times Greater Than 100 Years (Kavanaugh and Goldstein, 1999)	11
Figure 2.3: Between and within cluster variation (Dillon and Goldstein, 1984: 160)	20
Figure 2.4 Simple Dendrogram (Based on Krzanowski, 1988: 91)	28
Figure 2.6: Example Decision Tree.....	40
Figure 4.1: Cluster 1 Plot of Normalized Cost Versus Percent Source Removal for All Data Points.....	80
Figure 4.2: Cluster 2 Plot of Normalized Cost Versus Percent Source Removal for All Data Points.....	80
Figure 4.3: Cluster 1 Plot of Normalized Cost Versus Percent Source Removal (Normalized Cost Scale Reduced).....	81
Figure 4.4: Cluster 2 Plot of Normalized Cost Versus Percent Source Removal (Normalized Cost Scale Reduced).....	81

List of Tables

Table	Page
Table 2.1: Key Aquifer Parameters Affecting the Performance and Cost of a Subsurface Remediation Project	13
Table 2.2: Key Contaminant Parameters that Affect the Performance and Cost of a Subsurface Remediation Project.....	16
Table 2.3: Basis, Rationale and Applicability of Remedial Objectives	18
Table 2.4: Cluster Analysis Study Parameters	43
Table 2.5 Cluster Analysis Results	44
Table 3.1 Association Table Construction	57
Table 3.2: Cluster Analysis Example: Raw Data.....	62
Table 3.3: Cluster Analysis Example: Quantitative Parameter Standard Deviations	62
Table 3.4: Cluster Analysis Example: Standardized Data	63
Table 3.5: Cluster Analysis Example: Non-quantitative Attributes.....	63
Table 3.6: Cluster Analysis Example: Quantitative Distance Measures.....	64
Table 3.7: Cluster Analysis Example: Matching-type Similarity Measures for Quantitative Parameters.....	64
Table 3.8: Cluster Analysis Example: Matching-type Similarity Measures.....	66
Table 3.9: Cluster Analysis Example: Combined Similarity Matrix	66
Table 3.10: Clustering Analysis Example: Clustering Similarity Matrix	66
Table 3.12: Cluster Analysis Example: Similarity Coefficient Matrix for 5 Clusters	68
Table 3.13: Cluster Analysis Example: Clustering Coefficient for 1 Cluster.....	69
Table 4.1: Cluster 1 Projects and Parameter Values	72
Table 4.2: Cluster 1: Parameter Range, Average, Standard Deviation, and Frequency ...	73
Table 4.3: Cluster 2 Project and Parameter Values.....	75
Table 4.4: Cluster 2: Parameter Range, Average, Standard Deviation and Frequency	75
Table 4.5: Cluster 1 Mass Contaminant Treated, Lifecycle Cost, Normalized Cost and Percent Source Removal.....	77
Table 4.6: Cluster 2 Mass Contaminant Treated, Lifecycle Cost, Normalized Cost and Percent Source Removal.....	78

Abstract

Subsurface contamination by industrial chemicals is one of the most prevalent and costly environmental problems facing the United States government. This contamination problem must be managed to protect human health and the environment. Two basic strategies are used to deal with subsurface contamination: source removal and contaminant containment. While much cost and performance data are available for individual technologies associated with each of these strategies, there have been very few studies that have examined the benefits of implementing source removal technologies in order to reduce contaminant containment (and hopefully total) costs.

This study examines the tradeoff between extent of source removal and the lifecycle cost of a subsurface remediation project. It has been suggested that the lifecycle cost of a remediation project may be minimized at a certain percent source removal. This study attempts to validate this concept using real world data collected from 72 completed and on-going environmental remediation projects. Project data include total cost, extent of source removal, and site and contamination characteristics. Cluster analysis is used to group the diverse set of individual remediation projects under the assumption that projects within a cluster are similar enough to plot on a single lifecycle cost versus percent source removal plot. From the cluster analysis, two groups of 28 and 11 projects are used to develop lifecycle cost versus percent source removal curves. The resulting curves exhibit no apparent correlation between percent source removal and lifecycle cost. This study concludes with suggestions for future research that may shed light on the value of source removal towards reducing lifecycle cost of a subsurface contamination remediation project.

DETERMINING THE VALUE OF GROUNDWATER CONTAMINATION SOURCE REMOVAL: A METHODOLOGY

1.0 Introduction

1.1 Background

Groundwater contamination by industrial chemicals, one of the most prevalent environmental problems facing the United States government, is present at virtually every installation in the Air Force. The subsurface contamination must be managed to reduce the risk of adverse effects to human health and the environment as well as meet regulatory requirements. Cleaning up subsurface contamination has proven to be a very costly endeavor for the Air Force.

Through fiscal year (FY) 1999, Air Force installations have identified 4,530 sites that require some level of environmental response (Defense Environmental Restoration Program (DERP), 2000). Response actions, the measures taken to deal with the environmental contamination, range from long term monitoring of the site to construction of complex remediation technologies. Of the 4,530 sites requiring response actions, 2,377 sites have had the response action completed, meaning no further action is necessary (DERP, 2000). A total of 1,220 sites are still in the investigation stage, meaning information is being gathered to determine what level of environmental response action will be required for the site (DERP, 2000). To date, 94 site investigations are planned, but have not yet started (DERP, 2000).

Of the 1,220 sites that are currently in the investigation stage, 618 are going to require remediation technologies to cleanup the environmental contamination. The extent of remediation required for the remaining sites is presently unknown. Thus at least 618 decisions are waiting to be made concerning remediation of subsurface environmental contamination at Air Force installations.

One of the driving factors for any environmental remediation decision is cost. The following cost figures were taken from the DERP 1999 Annual Report. Through FY 1999, the Air Force has spent \$3.5 billion on environmental response actions, including site investigations. \$278 million has been obligated in the Air Force's FY 1999 budget, with another \$278 million planned for FY 2000. Almost \$300 million is being budgeted for FY 2001. The high cost of remediating sites in the Air Force poses a major problem for decision makers as they struggle to meet environmental remediation requirements as well as mission requirements under budget constraints. An analysis of remediation strategies could prove helpful to decision makers as they struggle to cost effectively manage contaminated sites.

1.2 Remediation Strategies

There are two basic strategies, source removal and contaminant containment, currently being employed, either separately or in conjunction with one another, to deal with subsurface contamination. Source removal is focused on removing contaminant mass at the source of contamination. One common method for source removal is excavation/treatment/disposal (colloquially known as "dig and dump"). This method consists of excavating a volume of contaminated soil from the subsurface, then treating and/or disposing of the excavated material in a safe manner. Another source removal

methodology is direct pumping, which is used to remove pools of contaminant. This treatment technology is limited in its application to Non-Aqueous Phase Liquids (NAPLs). Once located, the free product is pumped from the subsurface, reducing the mass of the contaminant in the subsurface. Direct pumping is primarily used for Light Non-Aqueous Phase Liquids (LNAPLs). LNAPLs have a density less than water resulting in the LNAPL pooling at the surface of the groundwater (at the water table). The LNAPL can be directly pumped to the surface for treatment and disposal. In unusual cases, direct pumping can also be used for Dense Non-Aqueous Phase Liquids (DNAPLs). DNAPLs have a density greater than that of water, causing them to migrate downward until an impermeable layer is reached, where they form a pool of contaminant. DNAPLs in the subsurface are extremely difficult to locate, but when and if located, the free product can be pumped to the surface for treatment and disposal. Another way to effect source removal is by flushing. Flushing involves injection of a material (e.g., surfactants, alcohols, steam, etc.) into the subsurface to mobilize the contaminant. The contaminant, which may be dissolved in the flushing solution or mobilized as a separate phase, is pumped to the surface for treatment and disposal. A fourth source removal technology is in situ destruction, typically done by addition of a strong oxidant to the subsurface to effect chemical oxidation of the contaminant. Both flushing and in situ destruction are considered innovative though they have become more commonly used in recent years.

The second strategy for managing subsurface contamination is to contain the plume of contaminated groundwater to prevent the contamination from migrating to potential human or ecological receptors. One of the most commonly used containment practices

for both large and small scale groundwater sites is pump and treat (Wang and Zheng, 1997). This process requires the extraction of contaminated groundwater from the subsurface and treating the contaminated water above ground. The treated water is then either injected back into the subsurface or discharged to surface water. Another technology that is becoming more commonly used for containment involves permeable reactive barriers. Permeable reactive barriers are placed in the subsurface so that contaminated groundwater must flow through a reactive media that destroys or contains the contaminant. A third containment process involves natural attenuation. In natural attenuation, containment is achieved without human intervention (although the technology may involve extensive monitoring). The mass, toxicity, mobility, volume or concentration of the groundwater contaminant is reduced by natural processes, such as biodegradation or chemical transformation.

Source removal technologies typically have high capital costs associated with their implementation, but require relatively little in the way of maintenance costs.

Containment technologies, on the other hand, have low capital costs, but operations and maintenance costs are high, since these technologies are most often applied over years or decades. When desired or required removal efficiencies are high, the costs for both strategies increase dramatically.

Removal efficiencies are a measure of performance of the remediation technology. The efficiency of the two remediation strategies must be measured differently. Source removal technologies are concerned with reducing the mass of contaminant in the subsurface and the subsequent effect the reduction has on the down gradient contaminant

concentrations in the groundwater. Removal efficiency for a source removal technology may therefore be best measured in terms of mass reduction. Containment technologies, on the other hand, are concerned with reducing the contaminant concentration to an acceptable level down gradient of the treatment technology. For the purposes of this thesis, remediation objectives will be defined in terms of achieving acceptable contaminant concentration at some point downgradient of a treatment technology. Removal efficiency for a containment technology may be defined as the contaminant concentration reduction achieved by the treatment technology divided by the contaminant concentration up gradient of the treatment technology.

Currently there is no source removal technology that can achieve total source removal. Any contaminant remaining in the source area may lead to dissolved contaminant in a plume downgradient of the source. As discussed earlier, the plume must be managed using a containment technology, to prevent the contaminant from reaching human or ecological receptors. There is obviously a trade-off between source removal and containment strategies. Ideally if total source removal is achieved (as noted earlier, currently unattainable) containment costs for a particular site and remediation objective will be minimized. On the other hand, if no source removal is effected, containment costs will be maximized. Total costs to manage a site include both source removal and containment costs throughout the life of a remediation project. Kavanaugh and Goldstein (1999) have recently hypothesized that for given site conditions, there is an optimal source removal efficiency that will minimize lifecycle cost of a remediation project (Figure 1.1).

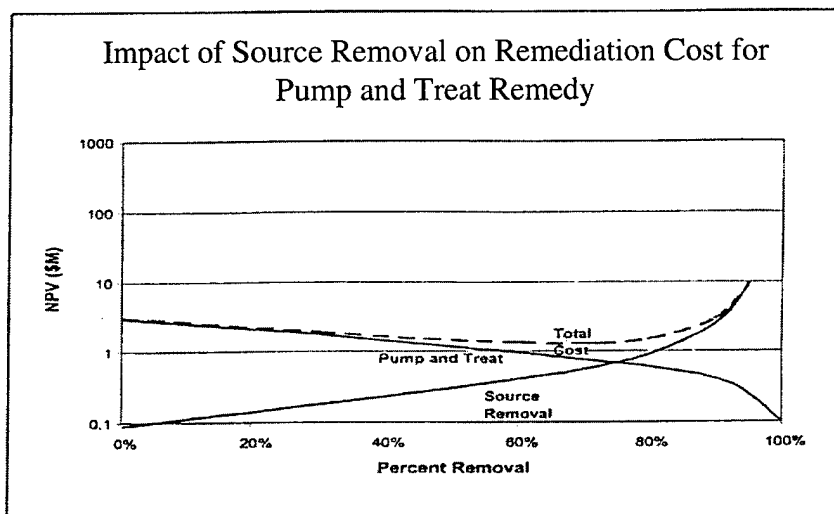


Figure 1.1 Kavanaugh and Goldstein Conceptualization (Kavanaugh and Goldstein, 1999)

It is important to note that this figure is hypothetical because little research has been done examining the tradeoffs between source removal and contaminant containment strategies. Current literature contains a wealth of cost and performance information for individual technologies (e.g. FRTR, 2000; US EPA, 1999; FRTR, 1998a-d). However, these studies provide no comparison of the tradeoffs between source removal and contaminant containment treatment strategies, as discussed above.

1.3 Problem Statement

Decision makers are faced with the challenge of managing subsurface contamination as cost effectively as possible while protecting human health and the environment.

Currently, source removal technologies such as flushing and in situ destruction are being developed which, at considerable capital cost, may allow site managers to remove a relatively large fraction of contaminant mass from source areas. Site managers must decide what level of source removal is most cost effective for a given site and remediation objective. Although there has been much cost and performance data published on the effectiveness of particular source removal and containment

technologies, there is little guidance available regarding the economic tradeoff between source removal and containment strategies. Site managers need such guidance, so they can make better decisions on the appropriate level of source removal to implement at the large number of sites currently being considered for remediation.

1.4 Research Objective

In this study, we will develop a methodology that can be used by site managers to determine the value of source removal in reducing containment costs when dealing with a subsurface contamination problem. The methodology can be applied at a site to determine what level of source removal, in combination with containment, can most cost effectively achieve remediation objectives. In this study, we will also attempt to validate the hypothesis that there is an optimal source removal fraction that will minimize lifecycle costs at a remediation site.

1.5 General Research Approach

1. Gather cost and performance data from a number of Air Force remediation projects. Sites, as well as technologies, will be diverse.
2. Conduct a literature review to determine methods that can be used to perform a meaningful analysis on such a diverse data set.
3. Analyze the data set with the goal of validating the hypothesis, set forth in Figure 1, that there is an optimal source removal fraction that minimizes life cycle cost.
4. Generalize the method so it can be used by site managers to make source removal decisions at a particular contaminated site.

1.6 Scope and Limitations of Research

Using data from Air Force remediation projects, this research will focus on developing a methodology for determining the value of source removal at a site. Once the methodology is developed, it will be utilized to validate the hypothesis set forth in Figure 1.1, that there is an optimum source removal fraction that will minimize lifecycle cost of a remediation project. With the extremely large variety of contaminants and site characteristics that are being dealt with at Air Force remediation sites, and the number of different remediation technologies that have been used, this research must be limited in scope. Four example charts, similar to Figure 1, will be constructed, using data from actual sites. Each example chart will use data from various remediation projects that can reasonably be graphed together due to project similarities. As an example, one chart may consider jet fuel contamination sites where there is floating jet fuel on top of the water table (the source) along with a fuel hydrocarbon plume. There are many Air Force sites where direct pumping has been used to remediate the floating source, while pump-and-treat methods were applied to contain the plume. Cost and performance data from these sites could be plotted on a single chart, for various fractions of source removal, in an attempt to validate the hypothesized relationship between source removal and lifecycle cost. Innovative technologies will not be considered in this study because the cost and performance data for these technologies are limited. Also many remediation efforts are currently ongoing, so final cost data may have to be extrapolated from available information.

As with all studies using real world data, we are constrained by data availability and reliability. The ideal situation would be that all the required data is available and reliable,

but that is not the case. We will always have data gaps that must be managed. Data gap management is necessary and appropriate; however, data gap management determines the best value from the given information and resources not the actual parameter value at the project.

2.0 Literature Review

2.1 Overview

In order to investigate the issue of determining the value of source removal, one must understand the state of the art in several areas. In this literature review, we will examine research relevant to formulating a methodology for determining the value of source removal in subsurface remediation. This literature review is composed of three sections. The first section will examine the current work in source removal valuation. The second section will describe the parameters that may be used to characterize and categorize subsurface remediation projects. The third section will examine analytical techniques that may be used to manipulate field data, to transform the data into a form that can be used to perform a meaningful analysis.

2.2 Current Work in Source Removal Valuation

While an obvious gap exists in the knowledge base, little has been done in the area of source removal valuation. Recently the Strategic Environmental Research and Development Program (SERDP) has issued a statement of need (SON) concerning the necessity for better understanding of the impacts of source zone treatment (SERDP, 2000). According to the SON, a knowledge gap exists in the understanding of the long-term impacts of source zone treatment (SERDP, 2000). The SON continues that “improved understanding of the benefits and risks of source zone treatment should result in more cost-effective remediation strategies at DoD sites” (SERDP, 2000). Kavanaugh and Goldstein (1999) presented a conceptual view of the tradeoffs between source removal and total life cycle cost of a remediation project.

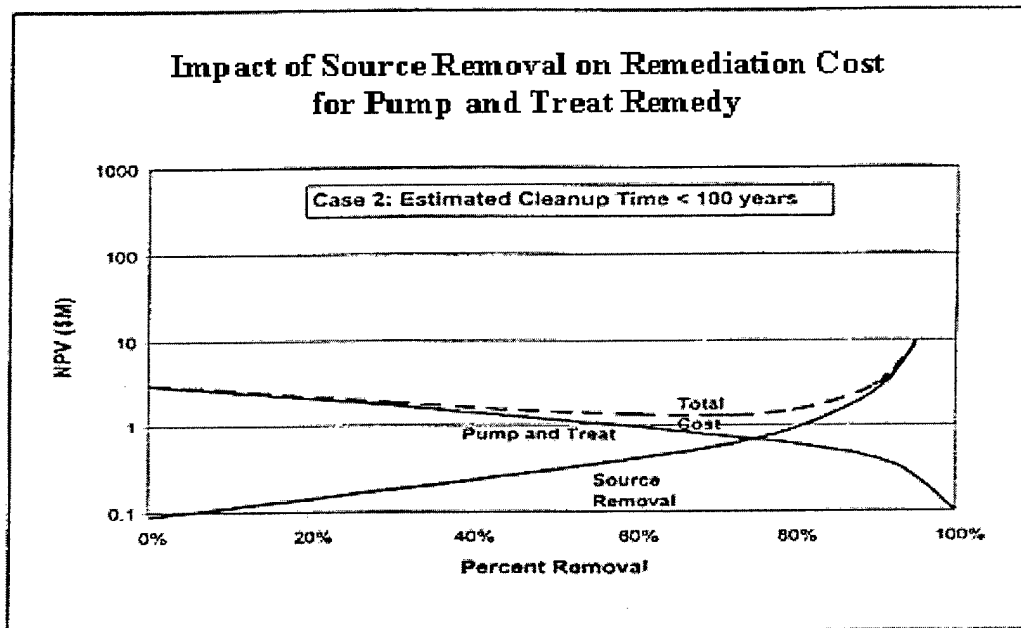


Figure 2.1. Estimated Cleanup Time Less Than 100 Years (Kavanaugh and Goldstein, 1999)

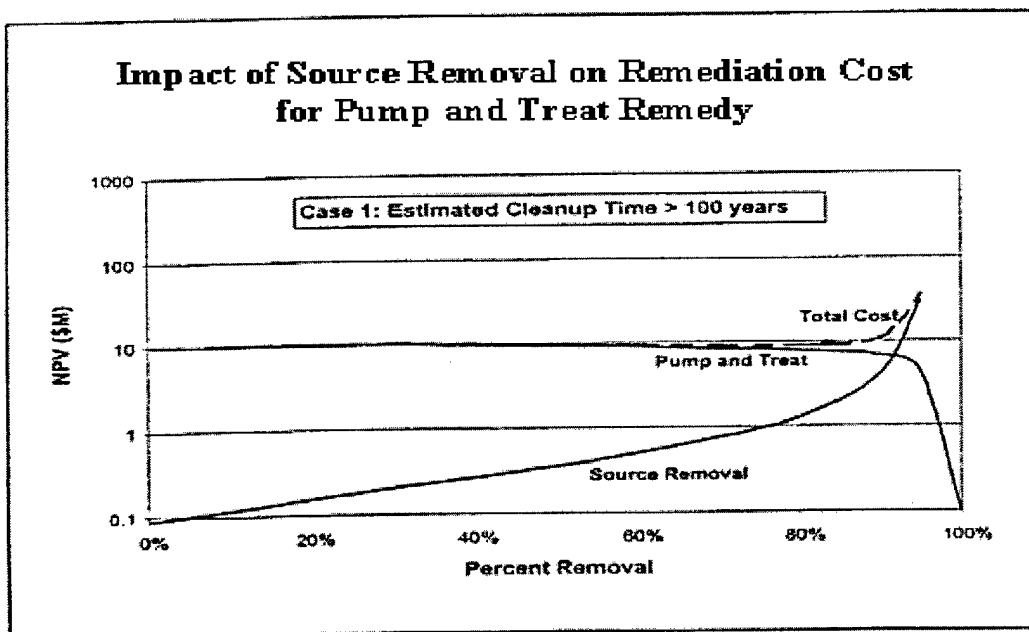


Figure 2.2: Estimated Cleanup Times Greater Than 100 Years (Kavanaugh and Goldstein, 1999)

Kavanaugh and Goldstein (1999) presented their conceptual view using two quantitatively different figures; the first (Fig 2.1) for relatively short cleanup times (less than 100 years) and the second (Fig 2.2) for relatively longer times (greater than 100) years. Percent removal on the x-axis of the figures represents the efficiency of the source removal technology and NPV on the y-axis is the lifecycle cost associated with the percent source removal. There are three curves on each figure. The source removal curve represents the cost of achieving the given percent of source removal. The pump and treat curve represents the cost of containing the plume that is present in the subsurface, given the percent removal of the source zone. It is important to note that a containment technology will always be utilized either alone or in conjunction with a source removal technology, as using current technologies, it is impossible to achieve 100% removal of the source. It should also be noted that zero percent removal on the figures represents that no source removal is effected. The total cost curve represents the cost of source removal and contaminant containment combined to show the total life cycle cost of the remediation effort. From Figures 2.1 and 2.2, total cost will have a minimum value associated with a certain percent removal of the source.

2.3 Site Characterization

As mentioned in Chapter 1, when managing subsurface contamination, many variables are relevant. There are a vast number of site characteristics that will affect the cost and performance of the remediation effort including contaminant and aquifer characteristics, as well as the goals of the remediation project (or remedial objectives). This portion of the literature review focuses on describing the characteristics of the aquifer, the contaminant and the remedial objectives, that may affect the cost and performance of the

remediation project. The characteristics described below can be used to define a subsurface remediation project and those projects with similar characteristics can be grouped. Using a group of similar projects will allow us to compare “apples to apples” in order to quantify the value of source removal for remediation projects.

2.3.1 Properties of the Aquifer

One of the most important things to consider when dealing with subsurface contamination is the properties of the subsurface itself. This discussion will focus on the properties of the subsurface that have proven to most affect the cost and performance of a remediation project. The Federal Remediation Technology Roundtable (FRTR) has identified the following properties as the suggested parameters to document for full-scale remediation projects (FRTR, 1997): soil classification, clay content or particle size distribution, hydraulic conductivity/water permeability, moisture content, air permeability, pH, porosity, transmissivity, total organic carbon, oil and grease, and Non-Aqueous Phase Liquids (FRTR, 1997). Table 2.1, taken from US EPA (1998e, Appendix A), FRTR (1997), and US EPA (1991) describes how each of these properties affect the cost and performance of remediation projects.

Table 2.1: Key Aquifer Parameters Affecting the Performance and Cost of a Subsurface Remediation Project

Parameter	Parameter Affect on Performance and Cost of Treatment Technology	Quantitative or Non-Quantitative
Soil Classification	Soil classification affects the relative ease of treating soil and groundwater.	Non-Quantitative, ASTM Classification
Clay content or particle size distribution	Clay and particle size distribution affect the flow of air and fluid through contaminated media.	Quantitative, Range from 0 to 1

Hydraulic conductivity/water Permeability	Hydraulic conductivity and water permeability affect the zone of influence of the extraction wells and therefore affect the number of wells needed for the remediation effort and the cost of operating the extraction wells.	Quantitative, Range >0
Moisture content	The moisture content of the matrix typically affects the performance, both directly and indirectly, of such in situ technologies as bioventing and soil vapor extraction and such ex situ technologies such as stabilization, incineration and thermal desorption. For example, air flow rates during operation of soil vapor extraction are affected by moisture content of the soil.	Quantitative Range from 0 to 1
Air permeability	This characteristic is important to in situ soil remediation technologies that involve venting or extraction. Air permeability affects the zone of influence of the extraction wells and therefore affects the number of extraction wells needed for the remediation effort and the cost of operating the extraction wells.	Quantitative, Range >0
pH	The pH of the matrix can affect the solubility of contaminants and biological activity. In addition, pH can affect the operation of treatment technologies. pH in the corrosive range (<2 and >12) can damage equipment and other special handling procedures.	Quantitative, Range from 1 to 14
Porosity	This characteristic is important to in situ technologies that rely upon use of a driving force to transfer contaminants into an aqueous or air-filled space. Porosity affects the driving force and therefore the performance achieved by the technologies.	Quantitative, Range from 0 to 1
Transmissivity	This characteristic is important for groundwater pump and treat or fluid cycling systems. Transmissivity affects the zone of influence in this type of remediation, thereby affecting the number of wells needed and the cost of operating the wells.	Quantitative, Range >0
Total organic carbon (TOC)	TOC affects the desorption of contaminants from soil and affects in situ soil remediation, soil washing, stabilization, and in situ groundwater bioremediation. TOC content may differ in uncontaminated and contaminated soil.	Quantitative, Range 0 to 1

Oil and Grease	Oil and grease affect the desorption of contaminants from soil.	Non-Quantitative Binary, Present/Not
Temperature	Temperature affects the biological processes within the subsurface especially the activity and kinetics.	Quantitative, Range >0 K
Oxygen Availability	Oxygen availability affects the biological processes through aerobic/anaerobic metabolism.	Quantitative, Range ≥ 0
Redox potential	Redox potential affect chemical speciation and mobilization.	Quantitative, Real Numbers
Heterogeneity/ layering	Heterogeneity/layering affects extraction or injection rates.	Non-Quantitative, Degree of layering
Depth to groundwater	Depth to groundwater affects many aspects of remediation effort. Along with area, determines the volume of contaminated material as well as affects cost when sampling, extracting, and excavating.	Quantitative, Range ≥ 0
Flow Velocity	Flow velocity affects the direction, location and extent of contamination.	Quantitative, Range ≥ 0

(Adapted from US EPA (1991), FRTR (1997), and US EPA (1998e))

The American Society for Testing and Materials also uses many of the parameters mentioned above as part of their standard subsurface characterization procedures, (ASTM, 1998). The parameters in Table 1 often play a major role in determining the cost and performance of a remediation effort at a site.

2.3.2 Properties of the Contaminant

Just as the subsurface has several properties that affect remediation technology performance and cost, the contaminant has several properties that impact the performance and cost of an environmental remediation effort. The following table lists contaminant parameters of interest as well as a brief description of the importance of the parameter.

Table 2.2: Key Contaminant Parameters that Affect the Performance and Cost of a Subsurface Remediation Project

Parameter	Parameter Affect on Performance and Cost of Treatment Technology	Quantitative or Non-Quantitative
Chemical Class/Contaminant Type	Chemical class affects the performance of a remediation system, some systems are more appropriate for a particular class of contaminant.	Non-Quantitative, Organic (alkanes; Volatile Organic Compounds (VOCs); semivolatile organic compounds (SVOCs); and polychlorinated biphenyls (PCBs)) Inorganic (metals and cyanides)
Henry's Constant	Henry's constant is the concentration of a chemical in the vapor phase over the concentration of chemical in aqueous phase (at equilibrium). This is important in remedy selection and operation time of the system.	Quantitative, Range ≥ 0
Vapor Pressure	Vapor pressure is the amount of contaminant in the vapor phase over a pool of pure contaminant. This is important in remedy selection and operation time of the system.	Quantitative, Range ≥ 0
Specific Gravity	Specific gravity is the ratio of the weight of a given substance to the weight of a reference substance with the same volume. For our purposes the reference liquid is water. A specific gravity above 1 means the chemical is heavy than water or a DNAPL. A specific gravity less than 1 means the chemical is lighter than water or a LNAPL. This effects the performance and cost of a system because LNAPLs and DNAPLs source zones will be found at different points within the aquifer.	Quantitative, Range > 0
Present as LNAPL	Present as LNAPL means the contamination is in pure phase and is lighter than water. LNAPLs can be a continuous source of groundwater contamination driving down performance and driving up cost and operational time.	Non-Quantitative. Binary Present/Not

Present as DNAPL	Present as DNAPL means the contamination is in pure phase and is lighter than water. DNAPLs can be a continuous source of groundwater contamination driving down performance and driving up cost and operational time. DNAPLs are very difficult to locate.	Non-Quantitative Binary Present/Not
Organic Carbon Partition Coefficient (K_{oc})	K_{oc} is an estimate of the adsorption tendencies of the chemical to organic material in the subsurface. It is assumed that when a chemical particle is adsorbed, it is not available for treatment. This parameter can affect the system operating time.	Quantitative, Range ≥ 0
Contaminant Concentration	Contaminant concentration is the amount of contaminant per unit of subsurface. Several methods for report concentration are available and can be based on mass or volume measurements. High contaminant concentrations mean more contaminant is present in the subsurface. Contaminant concentration can affect system operation time as well as size of treatment system.	Quantitative, Range ≥ 0

(Adapted from US EPA, 1991; FRTR, 1998; Sellers, 1999)

2.3.3 Remedial Objectives

Remedial objectives are the goals of the remediation project, which determine “how clean is clean”? In other words, the objectives specify the criteria for completion of a remediation project. Intuitively, remedial objectives will drive cost by establishing ultimate clean up levels. Obviously, a project that has a remedial objective of reducing dissolved contaminant concentrations to 5 ppm in groundwater will be much cheaper than a project that has to attain dissolved contaminant concentrations of 5 ppb. The primary objective of remediation projects is to protect human health and the environment (Sellers, 1999). This objective is usually defined by project-specific contaminant concentrations (in each subsurface zone) that have been deemed adequate to protect human health and

the environment (Sellars, 1999). The project specific objectives can have a number of bases. The following table lists these bases, their rationale, and when they might be applied.

Table 2.3: Basis, Rationale and Applicability of Remedial Objectives

Basis	Rationale	Applicability
Detection Limit	Impossible to quantify what cannot be measured so cleanup to point that is not detectable.	When risk-based cleanup level is below analytical detection limit.
Background Levels	Cleanup to level equal to concentration before contamination occurred.	Site-specific background levels can be determined. Use for sites with unusually high background concentration of contaminant.
Regulatory Standards	Regulatory standards are set forth by law or regulation so cleanup contaminant to meet or exceed standard for contaminant concentration.	Most widely used. Standards for soil, sediment, groundwater, and air.
Risk Assessment	Risk assessment is the analysis of potential risks to humans or the environment so cleanup contaminant to minimize risk or meet regulatory standard for acceptable risk.	Site-specific risk assessment identifies potential receptors and risk to these receptors, applicable if potential risk meets acceptable levels.
Protection of Groundwater	Cleanup to level so that contaminant leaching into groundwater is below acceptable concentration.	Set soil standards based on potential of contaminant leaching into groundwater.
Mass Removal	Reduce contaminant mass to protect groundwater or fulfill regulatory objectives.	Applicable to hot spots or other source areas.

(Adapted from Sellers, 1999)

2.4 Analytical Techniques for Reducing Field Data

Dealing with field data from environmental remediation projects can be challenging. Field data may be available, but comparing data between sites can be difficult because each site is different. The properties of the contamination, aquifer, and remediation

objectives can differ dramatically. When analyzing field data from these varied sites, it is necessary to reduce the data to a form that is useful in answering the problem statement. To plot data from widely varying sites on graphs such as those in Figures 2.1 and 2.2, it is necessary to apply data reduction techniques. The following two sections will discuss a pair of techniques, cluster and discriminant analysis, that may be used for data reduction. In addition, a third section will describe classification using decision trees. The goal of each of these techniques is to divide a given data set into manageable subgroups, in which the objects within groups are more similar than the objects between groups (Nouwen et al., 1997). This will allow us to compare apples to apples when constructing plots like Figure 2.1 and 2.2.

2.4.1 Cluster Analysis

The goal of cluster analysis is to arrive at clusters of objects that display small within-cluster variation relative to between-cluster variation. This basically means that the objects within one cluster are more similar to each other than the objects within another cluster (Dillon and Goldstein, 1984). In the context of this study, an object would be a particular remediation project and a cluster would be group of remediation projects that have very similar characteristics. Dillon and Goldstein present the following figure to illustrate the theory of cluster analysis.

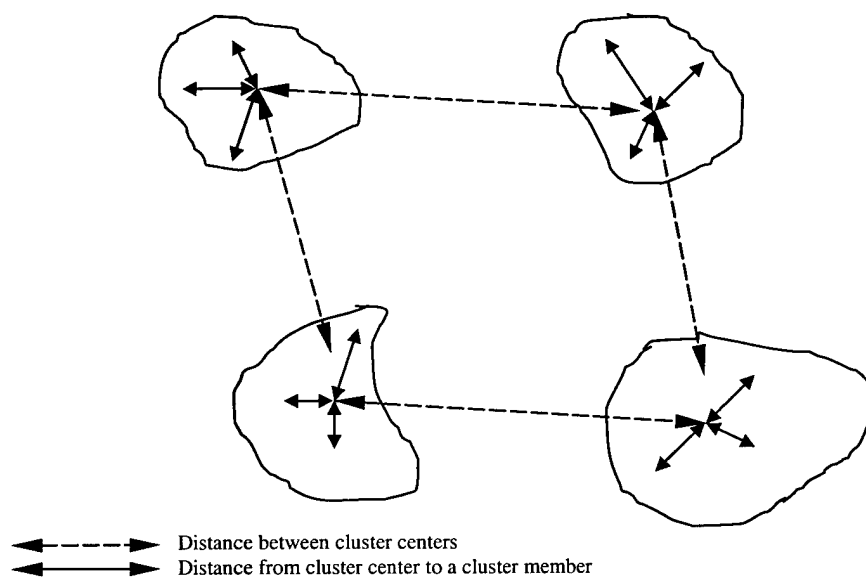


Figure 2.3: Between and within cluster variation (Dillon and Goldstein, 1984: 160)

An important parameter needed for cluster analysis is a measure of the similarities of objects. Two measures, distance-type and matching-type, can be used to measure the similarities (Dillon and Goldstein, 1984). Distance-type measures are for use with data having metric (quantifiable) properties while matching-type measures are for data having qualitative (non-quantifiable) properties (Dillon and Goldstein, 1984). When dealing with the objects of this study, environmental remediation projects, distance-type measures would be applicable for quantitative parameters like hydraulic conductivity and depth to groundwater table while matching-type measures would be appropriate for non-quantitative parameters like “type of contaminant” or “present as DNAPL”. The following two sections describe distance-type and matching-type measures.

2.4.1.1 Distance-Type Measures

Distance measures are used to determine the distance between vectors of observations from each object (Dillon and Goldstein, 1984). X_i equals $(X_{i1}, X_{i2}, X_{i3}, \dots, X_{ip})$ is the notation used to denote the measurements collected on the i th object or individual over the range of p variables (Dillon and Goldstein, 1984). In the context of this study, X_i would be the i th remediation project, where i goes from 1 to n (where n is the total number of projects), and p would be the number of metric (quantifiable) variables under consideration. Suppose two variables, hydraulic conductivity and depth to groundwater table, were being considered. Using this scenario, p would equal 2, so X_{i1} would be the hydraulic conductivity value for project i and X_{i2} would be the depth to groundwater table of project i . The distance can be calculated using the Minkowski metric, which is as follows (Dillon and Goldstein, 1984: 162):

$$d_{ij} = \left\{ \sum_{k=1}^p |X_{ik} - X_{jk}|^r \right\}^{1/r} \quad (2.1)$$

where d_{ij} equals the distance between objects i and j ,

X_{ik} represents the value of the k th parameter for the i th project ,

X_{jk} represents the value of the k th parameter for the j th project,

k represents the parameter descriptor $k = 1, 2, \dots, p$ for the i th and j th projects,

p is the number of parameters under consideration, and

r is a weighting factor chosen by the user where $r \geq 1$ (Anderberg, 1973).

By selecting different values of r , various metric distance functions can be obtained by the user (Anderberg, 1973). Setting r equal to one yields the city-block metric which is as follows (Dillon and Goldstein, 1984):

$$d_{ij} = \sum_{k=1}^p |X_{ik} - X_{jk}| \quad (2.2)$$

An r value equal to two yields the familiar Euclidean distance measure, represented by the following notation (Dillon and Goldstein, 1984):

$$d_{ij} = \left\{ \sum_{k=1}^p |X_{ik} - X_{jk}|^2 \right\}^{1/2} \quad (2.3)$$

One problem that arises from distance-type measures is scale invariance, meaning different distance-type measure values are obtained when units are changed (Dillon and Goldstein, 1984). One method of dealing with this problem is to standardize the data by dividing each variable by its standard deviation before computing the distance-type measure (Dillon and Goldstein, 1984). This procedure will preserve the relative distances and eliminate the problem of scale invariance (Dillon and Goldstein, 1984).

2.4.1.2 Matching-Type Measures

Matching-type measures are used when the data are nominally scaled, meaning the data are described in terms of classes which allows the user to place the object in one and only one set of mutually exclusive, collectively exhaustive classes (Dillon and Goldstein, 1984). This means that the object can be defined as possessing or not possessing a certain attribute (Dillon and Goldstein, 1984). Several methods exist for determining the similarity between objects using this method. The following example, adapted from Dillon and Goldstein (1984) but using two hypothetical environmental remediation projects, demonstrates the matching-type measures. Matching-type measures are for non-

quantifiable data. In this example, we will look at soil classification (more specifically, is the aquifer sandy?), LNAPL presence, DNAPL presence and type of contaminants present (VOC, SVOC, and/or metals). The following data are for two projects, A and B, where 1 represents the presence of an attribute and 0 represents the absence of the attribute (Dillon and Goldstein, 1984: 163).

Attribute Project	1 (Sandy)	2 (LNAPL)	3 (DNAPL)	4 (VOC)	5 (SVOC)	6 (Metals)
A	0	1	1	0	1	1
B	1	0	1	0	0	1

From the given data set, the following association table can be generated (Dillon and Goldstein, 1984: 164), where the (+,+) cell value is the total number of attributes that are present for both Projects A and B (Attributes 3 and 6). The (-,+) cell value is the total number of attributes that are present at Project B, but not at Project A (Attribute 1). The (+,-) cell value is the total number of attributes that are present at Project A but not at Project B (Attributes 2 and 5). Finally, the (-,-) cell value is the number of attributes that are not present at Project A or Project B (Attribute 5).

		Project A		
		+	-	
Project B	+	2	1	3
	-	2	1	3
		4	2	6

Based on the association table, the similarity between Projects A and B can be determined using a similarity coefficient. Several similarity coefficients are available depending on the use for which the coefficient is intended (Dillon and Goldstein, 1984).

The following list contains six of the most popular similarity coefficients (Dillon and Goldstein, 1984: 164):

$$(i) \frac{a+d}{a+b+c+d} \quad (2.4)$$

$$(ii) \frac{a}{a+b+c} \quad (2.5)$$

$$(iii) \frac{2a}{2a+b+c} \quad (2.6)$$

$$(iv) \frac{2(a+d)}{2(a+d)+b+c} \quad (2.7)$$

$$(v) \frac{a}{a+2(b+c)} \quad (2.8)$$

$$(vi) \frac{a}{a+b+c+d} \quad (2.9)$$

The values of a, b, c and d correspond to the cells of the two by two association table (Dillon and Goldstein, 1984) as follows:

	+	-
+	a	b
-	c	d

Dillon and Goldstein note the following ways the six similarity coefficients are different from each other:

- (1). how negative matches, that is, (-,-), are incorporated into the measure.
- (2). whether or not matched pairs of variables are equally weighted, or carry twice the weight of unmatched pairs. Note measures (iii) and (iv) double weight matched pairs.
- (3). whether or not unmatched pairs carry twice the weight of matched pairs. Note measure (v) double weights unmatched pairs.

(4). whether negative matches are excluded altogether. (Measures (ii), (iii) and (v))

For our example the six similarity coefficients will be calculated.

$$(i) \frac{a+d}{a+b+c+d} = \frac{2+1}{2+1+2+1} = \frac{3}{6} = \frac{1}{2} = .5$$

$$(ii) \frac{a}{a+b+c} = \frac{2}{2+1+2} = \frac{2}{5} = .4$$

$$(iii) \frac{2a}{2a+b+c} = \frac{2*2}{2*2+1+2} = \frac{4}{7} = .5714$$

$$(iv) \frac{2(a+d)}{2(a+d)+b+c} = \frac{2(2+1)}{2(2+1)+1+2} = \frac{6}{9} = \frac{2}{3} = .6667$$

$$(v) \frac{a}{a+2(b+c)} = \frac{2}{2+2(1+2)} = \frac{2}{8} = \frac{1}{4} = .25$$

$$(vi) \frac{a}{a+b+c+d} = \frac{2}{2+1+2+1} = \frac{2}{6} = \frac{1}{3} = .3333$$

Note that the similarity measures vary from .25 to .6667. The range can be attributed to the four differences described above.

A similarity coefficient of 1 indicates that both projects are identical for the attributes under consideration. A coefficient of 0 indicates that the attributes of both projects are totally dissimilar.

2.4.1.3 Distance-Type to Matching-Type Conversion

It is rather easy to transform distance-type similarity measures into matching-type similarity measures, but the reverse is not true (Dillon and Goldstein, 1984). Distance measures can be transformed to matching-type measures using $1/(1+d_{ij})$ where d_{ij} is the

distance-type measure between objects i and j (Dillon and Goldstein, 1984). An example of this can be seen using the quantifiable variable of hydraulic conductivity. Suppose two environmental remediation projects are under consideration and the hydraulic conductivity of each site is known. The distance-type measure, d_{ij} between the objects can be determined, as described above. If the hydraulic conductivity values for the two remediation projects are close to each other, the distance-type measure would be small (let us say .01), resulting in a matching-type measure of $1/(1+.01)$ or nearly 1 (nearly identical). If the hydraulic conductivity values for the two projects are drastically different, the distance-type measure would be very large (assume 50). This would result in a matching-type measure of $1/(1+50)$ or .02 (very dissimilar). This example illustrates the relative ease with which a distance-type measure can be transformed to a matching-type measure; however, matching-type measures can not be transformed to distance-type measures, because matching-type measures do not meet the necessary conditions of distance-type measures (Dillon and Goldstein, 1984).

Once the distance-type measure has been converted to a matching-type measure, the two will need to be combined to an overall similarity measure. According to Backer (1995), an overall coefficient can be determined using the following expression;

$$s(i, j) = \alpha s_n(i, j) + (1 - \alpha) s_b(i, j) \quad (2.10)$$

where $s(i,j)$ is the combined similarity coefficient, α is a weighting factor, s_n is the similarity coefficient from quantitative parameters, and s_b is the similarity coefficient from non-quantitative parameters. The weighting factor would be determined from available literature or expert opinion.

2.4.1.4 Using Measures to Cluster Objects

Once the similarity measure, either distance-type or matching-type, has been determined, a computational algorithm must be selected in order to cluster the data (Dillon and Goldstein, 1984). There are many computational algorithms available. This review will look at two of the most popular, hierarchical and partitioning methods (Dillon and Goldstein, 1984).

2.4.1.4.1 Hierarchical Techniques

Hierarchical techniques perform successive fusions or divisions of the data. Fusion is where each object begins as an individual cluster with larger clusters created by combining similar, smaller clusters (Dillon and Goldstein, 1984). Division is where all objects begin in one cluster and smaller clusters are formed by dividing the larger cluster (Dillon and Goldstein, 1984). Unending application of fusion hierarchical techniques will eventually lead to all objects in one cluster, and unending application of division hierarchical techniques will lead to all objects in their own cluster (Dillon and Goldstein, 1984). Thus, the question for hierarchical techniques is where to stop the clustering process (Dillon and Goldstein, 1984). The output of a hierarchical method is a dendrogram, which is a two-dimensional treelike diagram of the fusions or divisions at each successive level (Dillon and Goldstein, 1984). The following figure shows a simple dendrogram (Krzanowski, 1988: 91).

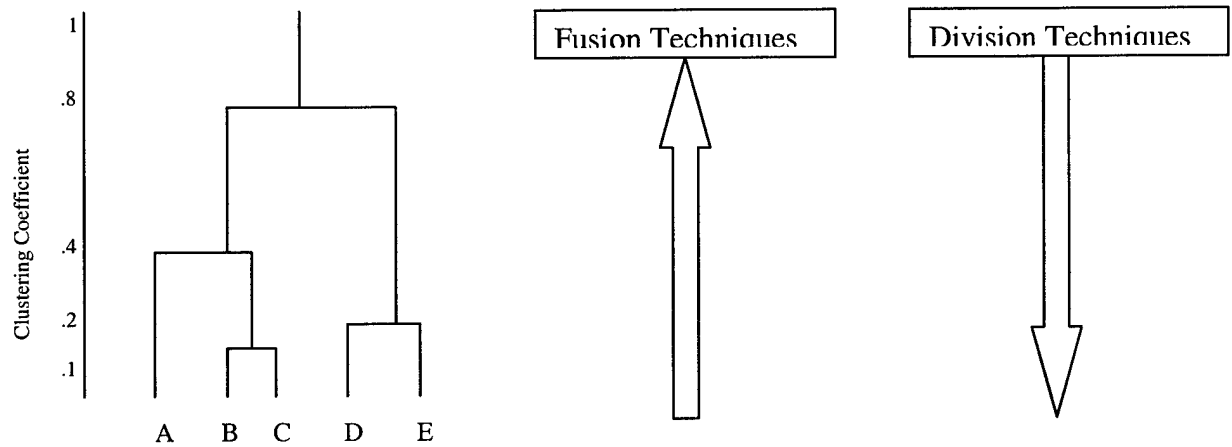


Figure 2.4 Simple Dendrogram (Based on Krzanowski, 1988: 91)

At the bottom of the dendrogram, all objects form individual groups; while at the top of the dendrogram, all objects fall into one group (Krzanowski, 1988). The arrows to the right show the starting and ending direction for both fusion and division hierarchical techniques. The vertical axis of the plot represents the clustering coefficient. A high clustering coefficient means less similarity within the group; while a smaller clustering coefficient, means more similarity within the group (Krzanowski, 1988). An example from above is that B and C are joined at a clustering coefficient of .1 which means they are very similar. This group is combined with A at a clustering coefficient of .4. The similarity between objects B and C is high and these objects are more similar to A than to D or E. It is important to note that Krzanowski assumes a small clustering coefficient means the projects are similar, while the previous discussion said that the higher the clustering (or similarity measure) the higher the degree of similarity between the projects. The clustering coefficients (or similarity measures) from Sections 2.4.1.1 and 2.4.1.2 can be transformed to coincide with the Krzanowski assumption by subtracting the similarity measure from 1.

With dendrogram and hierarchical methods defined, let us now describe how the dendrogram and the subsequent clusters are created using hierarchical methods. Divisive methods are not commonly used on large data sets because of the large number of computations that must be made (Krzanowski, 1988). Fusion techniques however are widely used because of the ease of computations (Krzanowski, 1988). Krzanowski outlines a four step algorithm for creating a hierarchy using fusion methods. The four steps are as follows:

Step 1: Define each individual as a group and store the similarity measure (either a distance measure or a similarity coefficient), between pairs of groups in a similarity measure matrix.

Step 2: Find the smallest element (using the Krzanowski assumption that 0 means projects are very similar and 1 means projects are very dissimilar) of the similarity matrix (making a random choice if necessary in the case of equality). Fuse the existing groups at this element and note the value of the similarity measure.

Step 3: Calculate the similarity measure between the new group and each of the existing groups. Replace the rows and columns of the similarity matrix of the fused groups with a single row and column of new similarity values, reducing the order of the matrix by one.

Step 4: Repeat steps 2 and 3 until left with just one group (Krzanowski, 1988).

As an example of this algorithm, suppose there are four remediation projects A, B, C, and D. The first step is to define each project as a group and determine the similarity measure between all pairings of the four groups (Step 1). Suppose that the similarity measure between Projects A and C is the smallest. Combine Projects A and C into one cluster

(Step 2). Proceed to step 3 which is the most difficult to compute. Rules must be established for calculating the similarity between two groups each of which may contain more than one member, that is with more than one object in the group (Step 3) (Krzanowski, 1988). Krzanowski outlines the six most frequently used methods for determining the similarity value for groups with more than one member. A description of the method, as well as the mathematical expression to calculate the new similarity measure, for each of the six methods is as follows: (For mathematical expressions, $D_{k,ij}$ is the similarity between any group k and the union of groups i and j . D_{ki} , D_{kj} and D_{ij} are the similarity measure (values), as discussed above, for groups k and i , k and j , and i and j respectively. The variable n_i is the number of objects in group i).

(1) Nearest Neighbor (single-link) method: the new similarity measure is the smallest value of the similarity measures between each element of the group,

$$D_{k,ij} = \frac{1}{2}(D_{ki} + D_{kj} - |D_{ki} - D_{kj}|) \quad (2.11)$$

(2) Furthest Neighbor (Complete-link) method: the new similarity measure is the largest of the similarity measures between each of the elements in the group,

$$D_{k,ij} = \frac{1}{2}(D_{ki} + D_{kj} + |D_{ki} - D_{kj}|) \quad (2.12)$$

(3) Group average method: the new similarity measure is the average of the similarity values between each of the elements within the group,

$$D_{k,ij} = \frac{n_i}{n_i + n_j} D_{ki} + \frac{n_j}{n_j + n_i} D_{kj} \quad (2.13)$$

(4) Median method: the new similarity measure is the distance between the medians of the groups,

$$D_{k,ij} = \frac{n_i}{n_i + n_j} D_{ki} + \frac{n_j}{n_j + n_i} D_{kj} - \frac{n_i n_j}{(n_i + n_j)^2} D_{ij} \quad (2.14)$$

(5) Centroid method: the new similarity measure is the squared Euclidean distance between the centroids of group i and group j where the elements of these groups are represented by points in space,

$$D_{k,ij} = \frac{1}{2} D_{ki} + \frac{1}{2} D_{kj} - \frac{1}{4} D_{ij} \quad (2.15)$$

(6) Minimum variance method: the new similarity measure is the between-group sum-of-squares for groups i and j .

$$D_{k,ij} = \{(n_i + n_k) D_{ki} + (n_j + n_k) D_{kj} - n_k D_{ij}\} / (n_i + n_j + n_k) \quad (2.16)$$

Continuing with the above example, a new similarity measure is calculated between the combined group of Projects A and C and the individual Projects B and D using one of the above methods, where $D_{k,ij}$ is the similarity measure between cluster A and C and individual groups B and D. For this example, $D_{B,AC}$ and $D_{D,AC}$ would be substituted for

$D_{k,ij}$. The similarity measure between sites B and D remains the same as first calculated. With the new similarity measures determined, the process repeats itself (Step 4) until one cluster is formed that contains all of the individual projects. For this example, we will assume the similarity measure between cluster A and C and group B is smaller than the similarity measure between cluster A and C and group D and the similarity measure between group B and group D. This is to say $D_{B,AC} < D_{D,AC}$ and $D_{B,AC} < D_{BD}$. With the new similarity measures compared, group B is added to cluster A and C because $D_{B,AC}$ is the smallest similarity value. The process repeats again, with group D eventually being clustered with A, C, and B to form one cluster that contains all objects.

There are advantages and disadvantages to each of the six methods. Furthest neighbor (complete-link), group average, and centroid methods all lead to spherical clusters exhibiting high internal affinity (Krzanowski, 1988). Single link methods lead to elongated clusters in which pairs of very dissimilar objects may occur because an object may join a cluster based on its relationship (similarity) to one object already assigned to the cluster, this could lead to heterogeneities within the cluster or the chaining effect (Krzanowski, 1988; Dillon and Goldstein, 1984). Only nearest and furthest neighbor techniques are invariant under monotonic transformation of the similarity measure, meaning only these two methods would give the same output if the logarithm of the similarity measure was used (Krzanowski, 1988). Nearest neighbor is also the easiest to perform making the analysis cost relatively low (Krzanowski, 1988). Nearest neighbor also works well if the data contain outliers (Krzanowski, 1988). Minimum-variance tends to force the data into clusters of equal spatial diameter, while the median method

weights the objects most recently added to the cluster more heavily than prior objects in the cluster (Krzanowski, 1988).

2.4.1.4.2 Partitioning Methods

Dillon and Goldstein (1984) describe partition methods as a cluster method that allows the movement of objects from one cluster to another if the initial cluster assignment was determined to be inaccurate. Partitioning methods usually assume that the final number of clusters is known or predetermined. Partitioning methods all look to maximize or minimize some criterion. Two partitioning clustering methods will be discussed, K-mean clustering methods and methods based on the Trace (Dillon and Goldstein, 1984).

2.4.1.4.2.1 K-means Clustering

K-means clustering can be described by the following example. Assume n remediation projects, with each project described by p parameters (Dillon and Goldstein, 1984). $X(i, j)$ is the value of the j th parameter for the i th project ($j=1,2,\dots,p$ and $i=1,2,\dots,n$) (Dillon and Goldstein, 1984). For this example, we will assume that distance-type measures were obtained and the Euclidean distance between objects can be calculated. $P(n, K)$ defines a partition such that each of the n projects is allocated to one of K clusters (Dillon and Goldstein, 1984). Note that there can be a number of possible ways that the n projects can be partitioned among K clusters. The goal is to determine the optimal distribution of projects among the clusters so that distances between projects in the same cluster are minimized.

To accomplish this, we define $X_{\text{bar}}(l, j)$ as the mean of the j th parameter in the l th cluster. We also define $n(l)$ as the number of projects in the l th cluster (Dillon and Goldstein, 1984). Using this notation, the following expression for the distance between the i th individual project and the l th cluster results (Dillon and Goldstein, 1984: 186):

$$D(i, l) = \left(\sum_{j=1}^p [X(i, j) - X_{\text{bar}}(l, j)]^2 \right)^{1/2} \quad (2.17)$$

The error component of the partition can be defined as (Dillon and Goldstein, 1984: 187):

$$E[P(n, K)] = \sum_{i=1}^n D[i, l(i)]^2 \quad (2.18)$$

where $l(i)$ is the cluster that contains the i th individual, and $D[i, l(i)]$ is the Euclidean distance between object i and the cluster mean of the cluster that contains object i . The procedure for clustering the data set so that the distances between objects in all the clusters are minimized involves moving objects from one cluster to another until no transfer of an object results in a reduction in the error component (Dillon and Goldstein, 1984).

2.4.1.4.2.2 Methods Based on the Trace

Methods based on the Trace, as explained by Dillon and Goldstein (1984), can be described as minimizing (maximizing) the within-group (between-groups) dispersion, where dispersion is the similarity measure, either distance-type or matching-type values. The total dispersion, T , is fixed by the data set and is equal to the sum of the within-group dispersion, W , and the between-group dispersion, B . Three methods based on the trace will be discussed here. The first method is the trace of W . The trace of W attempts to minimize the within-group dispersion W , which is equivalent to maximizing the between

group dispersion, B . The second method is the determinant of W . The determinant of W method attempts to minimize the determinant of the within-group dispersion matrix. The final method is the trace of BW^{-1} , which attempts to maximize the product of the between group dispersion and the inverse of the within-group dispersion matrices. Whichever method is selected, partitions are selected and rearranged so that only those partitions that yield an improvement in the criterion (method goal) are kept (Dillon and Goldstein, 1984).

2.4.1.5 Stopping Rules

As mentioned above, one of the major difficulties when applying cluster analysis is to decide when to stop the clustering process (that is, to decide when you have an appropriate number of clusters). Several methods for determining the optimum number of clusters have been proposed, but none have been generally accepted. The cophenetic correlation coefficient (CPCC) has been suggested as a method for evaluating the results of a cluster analysis, but this method may only be applied to distance-type measures of quantitative data (Dubes and Jain, 1979). Since, this study will convert distance-type measures to matching-type measures, the CPCC is not applicable to our analysis.

Another method that has been suggested for determining the optimal number of clusters is to evaluate the clustering coefficients (Mojena, 1977). Using the average and standard deviation of the clustering coefficients, Mojena (1977) developed a stopping rule. The stopping rule states that clustering should stop at the j th set of clusters when

$$\alpha_{j+1} > \bar{\alpha} + ks_z \quad (2.19)$$

where α_j is the clustering coefficient of the j th set of clusters, α_{j+1} is the clustering coefficient of the $j+1$ th set of clusters, $\bar{\alpha}$ is the average clustering coefficient, s_z is the clustering coefficient standard deviation, and k is the standard deviate. Mojena (1977) offers two heuristics to apply if the inequality is not met for any j th set of clusters. The first is to decide that one cluster is present, which indicates the data are random and no meaningful clusters can be formed. The second alternative is to determine the cluster set that has the maximum standard deviate. The number of clusters in this set represents the appropriate number of clusters. Finally, Mojena (1977) suggests that, if the inequality is never satisfied, another heuristic or stopping rule should be used to determine the appropriate number of clusters.

Another method for determining the optimal number of clusters is the Davies-Bouldin Index. This index measures within-to-between cluster spread to determine the number of clusters. The within-to-between cluster spread for clusters j and k ($R_{j,k}$) is defined as

$$R_{j,k} = \frac{e_j + e_k}{m_{j,k}} \quad (2.20)$$

where, e_j is the average error for the j th cluster, e_k is the average error for the k th cluster, and $m_{j,k}$ is the Euclidean distance between the cluster centers (Backer, 1995). The cluster center is defined as the average parameter value of the projects within the cluster. The standard error is the average difference between the cluster center and the individual project parameter values. We now define R_k as the maximum value of the within-to-between measures for cluster k . The Davies-Bouldin Index value for K clusters can then be fixed as

$$DB(K) = \frac{1}{K} \sum_{k=1}^K R_k \quad (2.21)$$

where $DB(K)$ is the Davies-Bouldin Index value (Backer, 1995). The Davies-Bouldin Index value would be determined for each partition (set of clusters) along the hierarchy. The number of clusters that has the minimum value of $DB(K)$ should represent the appropriate number of clusters based on the within-to-between spreads of the clusters.

A final method of selecting the appropriate number of clusters is comparing the clustering coefficients before and after clusters are formed to quantify the degree of dissimilarity of clusters that have merged. A relatively large difference, or “jump,” implies that two relatively dissimilar clusters have been merged (Aldenderfer and Blashfield, 1984). Thus the number of clusters prior to the jump is the most appropriate number of clusters. The subjective aspect of this method is deciding how large a difference in clustering coefficients constitutes a “jump.” While the previous methods discussed are rooted in statistical theory and computations, the “jump” method is a heuristic that proves useful and readily available when the exact number of clusters is not necessary to meet the objectives of the study.

2.4.2 Discriminant Analysis

Discriminant analysis is a statistical technique for classifying individuals or objects into mutually exclusive, collectively exhaustive groups on the basis of a set of independent variables (Dillon and Goldstein, 1984). “Discriminant analysis can be thought of in terms of a rather simple “scoring system” that assigns to each individual or object in the sample a score that is essentially a weighted average of the individual’s or object’s values

on the set of independent variables" (Dillon and Goldstein, 1984: 361). The score can then be transformed into a probability that gives the likelihood of the individual or object belonging to each of the groups (Dillon and Goldstein, 1984). The goal of discriminant analysis is to minimize the misclassification error rates (Dillon and Goldstein, 1984).

Discriminant analysis involves deriving linear combinations of the independent variables of a data set that will discriminate between previously defined groups (Dillon and Goldstein, 1984). Two assumptions must be made for discriminant analysis to be valid. The first assumption is that the independent variables must have a multivariate normal distribution (Dillon and Goldstein, 1984). The second assumption is that the variance-covariance matrix of the independent variables in each of the two groups must be the same (Dillon and Goldstein, 1984). One of the most common problems for discriminant analysis application is where two predetermined groups are defined and the data must be grouped into one of the sets. Each object will have certain components associated with the observation of the object. For this study, an object would be a remediation project and components would be characteristics of the project (aquifer characteristics, contaminants characteristics, remediation objectives). The assumption is that each object (project) will have a probability of being assigned to each of the two groups (Krzanowski, 1988). An allocation rule can be defined to assign each project observation to a specific group based on this probability (Krzanowski, 1988). The rule can be as simple as assigning the object to the group with the highest probability or as complex as incorporating cost weights into the rule (Krzanowski, 1988).

While the above general discussion of discriminant analysis supposes quantifiable variables, discriminant analysis can also be utilized when the data is non-quantifiable (discrete). As Dillon and Goldstein explain, the major difference in discriminant analysis of non-quantifiable data is the calculation of a discriminant score. A discussion of the discriminant score is provided by Dillon and Goldstein (1984) Chapter 10. The discriminant score is used to evaluate the object against the allocation rule, just as is done for quantitative allocation. However, the performance of discriminant analysis when the data is non-quantifiable is especially poor (Dillon and Goldstein, 1984).

It is appropriate to now discuss the situation when a data set contains both quantifiable and non-quantifiable data. Dillon and Goldstein (1984) provide a discussion of this situation. They use Bayes' Theorem to obtain the multivariate logistic function, which is then used to determine the probability of a project belonging to a certain group. In the context of this study, the parameters, both quantitative and non-quantitative, from individual remediation projects would be evaluated using the multivariate logistic function. Based on the output probability, the allocation rule would be used to classify the project into a group.

The difficulty in applying any type of discriminant analysis to this study is the fact that discriminant analysis is based on predetermined groups. When dealing with environmental remediation projects, predetermining groups would be difficult due to the large number of variables and variety of projects. It was determined that discriminant analysis would not be applicable to this study because we do not have predetermined groups. Therefore, discriminant analysis will not be discussed further in this study.

2.4.3 Classification Using Decision Trees

Classification using decision trees differs from cluster and discriminant analysis in that decision trees do not use mathematical procedures to group data. Rather, expert opinion and a decision scheme are created to classify the data into smaller groups. The classification scheme uses constraints placed on parameters to define smaller groups (Eisenberg and McKone, 1998). A simple example of a decision tree is shown in Figure 2.5 using the parameters chemical class/contaminant type and hydraulic conductivity.

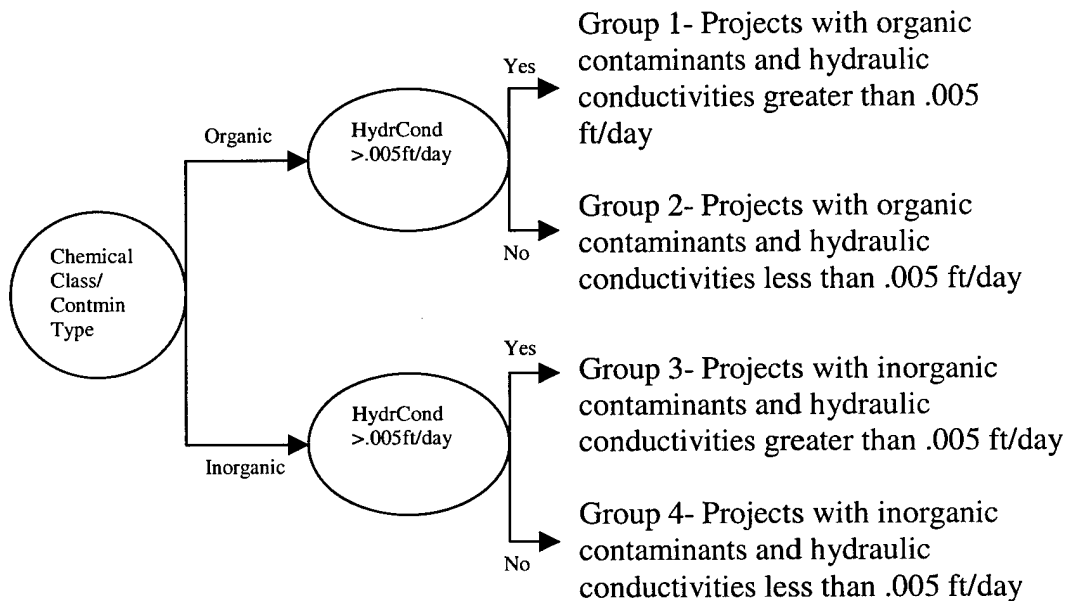


Figure 2.6: Example Decision Tree

The groups that are formed by decision trees are the last step in the decision tree, reached by proceeding through the tree from start to finish. For the above example, the groups are simply defined based on the application of the parameters and constraints, but it is also possible to have an assignment rule as the final step in the process. An example of an assignment rule would be to assign projects that are sorted into Groups 1 and 2 to a predetermined Group 5 while projects sorted into Groups 3 and 4 are assigned to a

predetermined Group 6. It is important to note that the end points of the decision tree have distinct boundaries (defined by the parameters and constraints) where no projects that are within one group can fall into another group and that all projects must belong to a group. As the number of parameters under consideration increases, so does the size of the decision tree. Researcher expertise must be used to define an explicit classification scheme that will classify all possible objects into well-defined and meaningful groups. There should be justification to defend the use of a particular classification scheme over another scheme.

For the purposes of this study, a classification scheme to differentiate between remediation projects would need to be developed based on current literature and expert opinion. The classification scheme would have a set number of categories in which each project would have to fall. A simple example classification would be to have all LNAPL remediation projects sort into one category and all remediation projects not having LNAPL to sort into a different category.

2.4.4 Case Studies of Application of Analytical Techniques to Group Objects

Cluster analysis has been used in a wide range of fields such as:

1. psychology for classifying individuals into personality types,
2. regional analysis for classifying cities into typologies based on demographics and fiscal variables,
3. marketing research to classify customers into segments on the basis of psychographic factors and product use, and

4. chemistry for classification of compounds based on performance properties
(Dillon and Goldstein, 1984).

Decision trees have been used to classify chemical compounds. The following sections present case studies that illustrate the uses of clustering analysis and decision tree grouping techniques. Each case study is concluded with a brief discussion on the relevance of the techniques to our specific problem.

2.4.4.1 Cluster Analysis Example: Cluster Analysis of Environmental Data which are not Interval Scaled but Categorical

The following case study is based on the work of Hannappel and Piepho (1996). Cluster analysis was used to evaluate the similarities between sampling sites using aerial photographs of groynefields (erosion control areas extending from the banks of rivers into the water (Morris, 1992)). "The result of the cluster analysis of sampling sites can be used to decide which sampling sites are chosen to be representative for a larger group" (Hannappel and Piepho, 1996: 335). The analysis begins with 596 groynefields that were manually examined using six parameters. The six parameter that were examined are listed in Table 2.4.

Table 2.4: Cluster Analysis Study Parameters

1. Type of sediments	4 values	Fine Grain, silt, mixed sediments, sand
2. Type of catchment area	4 values	Urban, field, meadow, relatively natural
3. Degree of filling	5 values	Little, 1/3, 1/2, 2/3, completely
4. Course of the river	3 values	Undercut slope, straight, slipoff slope
5. Size in m	4 values	Size ≤ 60 , $60 < \text{size} \leq 90$, $90 < \text{size} \leq 120$, $120 < \text{size}$
6. Form of water body	5 values	No form (completely filled), semicircle, irregular, triangular (increasing to next groynefield), triangular (decreasing to next groynefield)

(Adapted from Hannappel and Piepho, 1996)

The cluster analysis began with the largest number of clusters possible with each groynefield being one cluster (fusion technique). Since the data are not continuous and ordered, the Euclidean distance metric is not a valid similarity measure, so a similarity coefficient is selected. The computational algorithm that is selected is a hierarchical method using group average. The minimum-variance and centroid methods were not chosen because the data were not at least interval scaled, while nearest neighbor was not chosen because of the chaining effect (discussed in Section 2.4.1.4.1). The group average and furthest neighbor methods were considered, but furthest neighbor was not selected because furthest neighbor can lead to ties (the formation of larger groups based on several individual observations having the same similarity measure). Since the data are categorical, ties are more frequent due to the limited number of similarity measures (Hannappel and Piepho, 1996).

The data and the clustering algorithm were analyzed using the statistical computer software package SAS 6.1 for MS-Windows 3.1. The output from the analysis was four clusters containing the entire population of groynefields. The following table shows the results of the cluster analysis.

Table 2.5 Cluster Analysis Results

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Percent of data set	21%	48%	7%	24%
Type of sediment	Mixed sediment Sand	Fine grain <i>Silt</i> <i>Mixed sediment</i>	Mixed sediment Sand	Sand
Type of Catchment area	Relatively natural	Relatively natural <i>Urban</i> <i>Field</i> <i>Meadow</i>	Relatively natural Meadow	Relatively Natural Meadow
Degree of Filling	Little, 1/3	Little, 1/3	2/3, completely	2/3, completely
Course of the river	Straight undercut	Straight undercut	Slipoff	Straight
Size	≤90m	≤90m	>90m	≤90m
Form	Triangular Irregular	Semicircle <i>Triangular</i>	No Form	Irregular <i>Semicircle</i>

Note: Line in nonitalic indicates the property with the highest and second highest percentages in the respective cluster. Italicized line indicates the property is found most frequently in the respective cluster, compared with the other clusters (Hannappel and Piepho, 1996). For example, mixed sediment and sand are the most frequently observed type of sediment in Cluster 1 and silt and mixed sediment are found more frequently in Cluster 2 than in Clusters 1, 3, and 4.

(Hannappel and Piepho, 1996)

This case study demonstrates the ability of cluster analysis to reduce a given data set to meaningful groups of closely related objects. This case study looks at six parameters, each with several levels (possible values or descriptors), to group the data set into four clusters. This case study is very similar to our problem as environmental remediation projects have several parameters, each with more than one possible value, some quantitative and some non-quantitative. This case study demonstrates that cluster analysis is a viable option for analyzing our data set.

2.4.4.2 Decision Tree Classification: Classifying Environmental Pollutants 1: Structure-Activity Relationships for Prediction of Aquatic Toxicity

This case study pertains to the work of Verhaar et al. (1992). This analysis was performed to classify a large number of organic pollutants into one of five classes: [1] inert chemical, [2] less inert chemicals, [3] reactive chemicals, [4] specifically acting chemicals, and [5] chemicals that cannot be classified using this scheme, in order to approximate potential for chemical toxicity. Verhaar et al. (1992) begin by establishing explicit definitions for each of the five categories.

With these classes explicitly defined, a scheme was developed to enable one to assign chemicals, based solely on structural characteristics, to one of the classes. The classification scheme was developed by constructing rules for categorizing the chemicals based on structure. Rules were developed through a literature review and expert opinion. An example rule might read "If a chemical has a log octanol-water partition coefficient between 0 and 6 and a molecular weight less than 600 Daltons, it belongs in class 1, 2 or 3."

With these classification rules in place, 166 chemicals were classified and analyzed. (Verhaar et al., 1992). Fifty chemicals were classified class 1, 40 chemicals as class 2, 42 chemicals as class 3, and 34 chemicals were classified as class 4 (Verhaar, et al., 1992).

The work completed by Verhaar et al. (1992) demonstrates how classification rules may be developed and applied. It is important to note that a great amount of detail, for both group definition and classification rules, is needed for this classification technique to work effectively. It is also important to note that the process is subjective. While

literature is available as an aid, the researcher is ultimately responsible for defining both the group and the classification scheme.

3.0 Methodology

3.1 Overview

This chapter begins with a discussion of the data that will be gathered to aid in the development of a methodology to determine the value of source removal when managing subsurface contamination problems. As discussed in Chapters 1 and 2, the field data for environmental remediation projects will be diverse, with each project having several different parameters that can be used to describe the remediation project. This section will describe the data, data sources, data gap management practices, and outline the parameters that will be used for subsequent analysis. The second section of this chapter describes the selection of a grouping technique to classify the diverse set of remediation projects into subgroups for subsequent analysis. The second section of this chapter also includes an example of the grouping technique used to classify the data set into subgroups. This chapter concludes with a discussion of how the lifecycle cost versus percent source removal curves, similar to those presented in Chapter 2, Figures 2.1 and 2.2, will be constructed.

3.2 Data Management

3.2.1 Data Collection/Parameter Selection

The data set for this analysis will come from real world environmental remediation project information obtained from case studies and technology reports. The data will be both quantitative and non-quantitative in nature. Six of the parameters chosen for this study are quantitative (depth to groundwater table, hydraulic conductivity, contaminant concentration (source), contaminant concentration (dissolved), remedial objective

(source) and remedial objective (plume)) and one variable is non-quantitative (contaminant type). Expert opinion has suggested that these parameters would most influence the cost and performance of an environmental remediation effort (Goltz, 2000). The quantitative variables will have values greater than zero. Contaminant concentration and remedial objectives will be separately specified for the source area and for the plume area. The non-quantitative variable, contaminant type, will have four categories: chlorinated volatile organic compounds (CVOCs); benzene, toluene, ethylbenzene, and xylene compounds (BTEX); semi-volatile organic compounds (SVOCs); and metals. The collected cost data will be separated into the cost for source removal actions and costs for plume containment actions, though ultimately the costs will be combined for the purpose of determining the lifecycle cost of the entire remediation project. Although this study will use the parameters listed above, the methodology that is presented is general, and may be applied using additional or different parameters, as deemed appropriate.

The collected data will be stored in a database constructed using Microsoft Excel. The data base will list each environmental remediation project by its name along with the parameter values associated with it. Each project will be assigned a number from 1 to n, where n is the total number of remediation projects in the database. Project numbers will be used instead of names for easier data management. The data will be converted into consistent units before being entered into the database. The English system will be used to define the units for depth to groundwater (ft) and hydraulic conductivity (ft/day), while metric units will be used for contaminant concentration and remedial objectives (mg/kg for source area concentration and mg/L for dissolved concentration). English units will be used since most information from remediation projects are reported using this system.

It should be noted that either unit system (English or metric) could be used as long as the system remains consistent.

Other parameters will also be included in the database. Cost data for the source removal and contaminant containment technologies will be entered as a dollar value and year so that the net present value can be determined at a later time. The percent source removal will be entered as the percentage of the initial mass of contaminant in the source area that was removed by the remediation project.

3.2.2 Data Gap Management

As data are obtained, some values may not be reported, or reported in an unusable format. These missing values are referred to as data gaps. If a data gap is encountered, the gap will need to be filled or somehow accounted for if the project is to be included in the subsequent analysis. The following paragraphs will outline the data gap management practices that will be used.

For the parameter depth to groundwater table several situations are possible. In some instances, a range of values is reported for this parameter. For this situation, the minimum and maximum value will be entered into the database and the average value will be used for subsequent analysis. If no value is reported for the depth to groundwater table, the United States Geological Survey published value for the area immediately surrounding the project location will be used.

If a range of values is reported for the parameter hydraulic conductivity, the minimum and maximum value will be entered into the database and the average value will be used for subsequent analysis. If no value for hydraulic conductivity is reported, the soil classification will be used to estimate the hydraulic conductivity value. Typical hydraulic conductivity ranges for various soils are presented in Domenico and Schwartz (1998). The average of the minimum and maximum values will be used for subsequent analysis.

For contaminant concentration, the maximum reported contaminant concentration will be used. Reports for remediation projects typically include the maximum contaminant concentration, while only a limited number of reports include average contaminant concentration. Because of their availability, maximum contaminant concentration values will be used for contaminant concentration.

The remedial objective will be obtained from reported values. If the remedial objective is not explicitly stated, the federal standard for maximum contaminant concentration of the particular contaminant (in soil, groundwater, etc.) will be used. The federal standard for maximum contaminant concentrations will come from the Code of Federal Regulations (CFR) Title 40 and EPA published values (CFR, 2000 and EPA, 2000).

For the parameter contaminant type, the reported contaminant chemical will be used to determine which contaminant type, of the four possibilities (CVOC, BTEX, SVOC, metals), will be stored in the data base. If more than one contaminant type is present at a project, then each type that is present will be recorded.

Typically, capital and operations and maintenance costs are available along with the year for which the cost data are reported. The cost data will be converted to 2000 dollars using inflation indices reported by the United States Air Force (USAF, 2000). The index allows for conversion using a multiplier from costs incurred in one year to costs in another year. The net present value cost (Year 2000) will be normalized by the total mass of contaminant treated. This will allow remediation projects with different masses of treated contaminant to be compared. The normalized cost will be calculated by summing the costs of source removal and plume containment (year 2000 dollars) and dividing this sum by the sum of mass treated contaminant in the source and plume, as expressed in the following equation:

$$\frac{\text{NPV of Remediation Project}}{\text{Mass of Contaminant Treated}} = \frac{\text{NPV of source removal} + \text{NPV of plume containment}}{\text{Mass of source removed} + \text{Mass of contaminant in plume treated}} \quad (3.1)$$

The mass of source removed will be determined by multiplying the source contaminant concentration, the source volume, and the percent source removal. The mass of treated contaminant in the plume will be determined by multiplying the contaminant concentration dissolved in the plume by the volume of the plume treated.

The percent source removal will be determined from the reported information. A single percent source removal value will be used if reported. For projects in which the percent source removal is not reported, an attempt will be made to estimate it from the available information. For a contaminated soil source, the fraction removed can be determined by dividing the difference between the initial and final contaminant concentrations by the initial contaminant concentration. For free product (NAPL) sources, the fraction removed will be determined in the same manner, by dividing the difference between the

initial (mass or volume) and final (mass or volume) by the initial (mass or volume) of the contaminant. An example would be a free product layer (volume = 27,000 gallons) floating on the groundwater. Using a treatment technology, 20,000 gallons of free product is recovered. The percent removal would be $[(27,000 \text{ gallons} - 7,000 \text{ gallons}) / 27,000 \text{ gallons}]$ for a percent source removal of .743 or 74.3%. When excavation is used as a source removal technology, we will assume that the percent source removal is 100%. If the excavated material is treated using some treatment technology and then used as backfill at the site, the project percent removal will be the treatment technology removal efficiency. If the excavated material is disposed off-site, and clean-fill used as backfill, the percent source removal will be 100%. The cost of excavation, treatment, and/or disposal will be included in the total cost of the remediation project.

3.3 Grouping the Data

3.3.1 Grouping Technique Selection

Chapter 2 provided an overview of three possible grouping techniques that could be used to classify a diverse data set into subgroups for meaningful analysis. In order to analyze the database described in Section 3.2, cluster analysis will be applied. As mentioned in Chapter 2, discriminant analysis requires predetermined groups for the analysis.

Predetermined groups will not be used in this study; therefore discriminant analysis is not an appropriate classification technique. Decision trees will not be used for this study because the constraints that are used to form groups are subjective and based on expert opinion, which may vary and is not readily reproducible. Cluster analysis is optimal for this study because it examines the data and selects groups based on quantifiable

similarities of the data. As discussed previously, the goal of cluster analysis is to arrive at clusters of objects that display small within-cluster variation relative to between-cluster variation (Dillon and Goldstein, 1984). Cluster analysis is best suited for this study because it does not require predetermined groups and can be applied to data that is quantitative, non-quantitative, or both with relative ease.

3.3.2 Similarity Coefficient Determination

As discussed in Chapter 2, several techniques are available to perform cluster analysis of a given data set. The first requirement is to determine the similarity coefficients that will be used for the analysis. Distance-type measures for quantitative data and matching-type measures for non-quantitative data are available. The data from this study will be both quantitative and non-quantitative. We have two options to manage the combination of quantitative and non-quantitative data. The first option is to develop categories that may be used to transform the quantitative data into non-quantitative data. An example of this can be shown using the quantitative variable of depth to water table. Although the depth to water table can take on any value greater than zero, the data can be categorized non-quantitatively using classes such as 0 to 10 feet, 10 to 20 feet, 20 to 30 feet, and greater than 30 feet. The second option for managing quantitative and non-quantitative data was discussed in Chapter 2, where $1/(1+d_{ij})$ is used to transform a distance-type measure, represented by d_{ij} , into a similarity coefficient. For this study, the second option is chosen. This option is not based on expert opinion, which drives the categories of option 1, so disagreements surrounding the category definitions will not overshadow the analysis. The second option also allows use of the actual value of the quantitative

distance-type measure for the analysis rather than using an arbitrarily defined category for the quantitative variable.

As explained in Chapter 2, scale invariance causes problems in distance-type measures due to the impact of using different units of measure. Scale invariance will be dealt with in this study by standardizing the data, as described in Section 2.4.1.2.

A second determination that is required for cluster analysis is the similarity coefficient calculation for matching-type measures. As described in Chapter 2, an association table may be constructed based on the presence or absence of a given attribute for two remediation projects. Six methods for calculating the similarity coefficients from the association table were presented in Chapter 2. For this study, method (i) will be used to determine the similarity coefficient of matching-type measures. Method (i) is selected because all aspects of the association table are included and weighted evenly.

3.3.3 Clustering Technique Selection

Chapter 2 detailed two techniques, hierarchical and partitioning, for clustering objects based on the similarity measures calculated as described in the previous section.

Hierarchical techniques are a proven clustering method. Partitioning techniques usually assume that the final number of clusters is known or predetermined, which is not the case in this study. Hannappel and Piepho (1996) used hierarchical techniques to analyze a problem exhibiting many of the same characteristics as the one in this study.

Hierarchical techniques will be used to cluster the data in this study.

As discussed in Chapter 2, fusion or division methods are available for hierarchical clustering. Divisive techniques are computationally prohibitive for use in this study, while fusion techniques have been used in previous studies, including Hannappel and Piepho (1996). This study will use fusion techniques as outlined in Chapter 2 in Krzanowski's four-step algorithm.

Krzanowski's algorithm is based on a similarity value of 0 meaning two projects are very similar, while a similarity value of 1 indicates the projects are very dissimilar.

According to Krzanowski's algorithm, after smaller groups have been fused, a new similarity coefficient needs to be calculated to account for multiple members within the new group. To calculate the new similarity coefficient, one of six options must be chosen. The six options, as well as advantages and disadvantages, were explained in Chapter 2. The group average method will be used in this study because it provides for spherical clusters. Also, Hannappel and Piepho (1996) used group average methods to cluster a data set that is analogous to the data set in this study. Group average methods determine the new similarity measure based on the average similarity coefficient between the groups under consideration. The similarity coefficient for group average can be expressed as

$$D_{k,ij} = \frac{n_i}{n_i + n_j} D_{ki} + \frac{n_j}{n_j + n_i} D_{kj} \quad (2.13)$$

where $D_{k,ij}$ is the similarity coefficient of cluster k , and the new cluster, consisting of a combination of clusters i and j , the variable n is the number of objects in the group under consideration, and D_{ki} and D_{kj} are the similarity measures between groups k and i and groups k and j , respectively.

3.3.4 Specifics of Cluster Analysis for this Study

With the grouping technique chosen, it is now appropriate to explicitly describe the methods that will be used to cluster our particular data set. Using the database, a spreadsheet of clustering data will be constructed. This clustering spreadsheet will contain only the data that will be used to cluster the remediation projects into subgroups. Quantitative data will be entered as described in Section 3.2 of this chapter. The quantitative data will be normalized for the clustering spreadsheet. Non-quantitative data will be entered based on attributes, with 1 meaning presence of the attribute at a particular project and 0 meaning absence of the attribute at a particular project. The number of attributes will depend on the number of non-quantitative parameters and the number of possible values for the given parameter. For this study, we have one non-quantitative parameter, contaminant type, that can have four values; CVOC, BTEX, SVOC and Metals. We will arbitrarily assign attribute numbers to each possible parameter value. Thus, the presence of CVOC means Attribute 1 is present in the project; the presence of BTEX means Attribute 2 is present; the presence of SVOC means Attribute 3 is present; and the presence of metals means that Attribute 4 is present.

Each remediation project in the data set would be examined for the presence or absence of each attribute. The number of remediation projects, the number of quantitative variables and the number of attributes under consideration (in this case four) will determine the dimensions of the data entry spreadsheet. The number of remediation projects in the data set will determine the number of rows in the data entry table. The

number of columns will be determined by the total number of quantitative variables plus four attributes.

Once the data have been entered into the clustering spreadsheet, the cluster analysis can begin, using the methods described in Section 3.3.4. The first step in the cluster analysis is to calculate the similarity coefficient between each of the environmental remediation projects. As described above, the similarity coefficient between quantitative variables will be transformed to a matching-type similarity coefficient using $1/(1+d_{ij})$ where d_{ij} is the distance between the variable values. The distance between variable values will be calculated from equation (2.3) which is repeated:

$$d_{ij} = \left\{ \sum_{k=1}^p |X_{ik} - X_{jk}|^2 \right\}^{1/2} \quad (2.3)$$

where X_{ik} is the value of the quantitative variable k of the i th remediation project, X_{jk} is the value of quantitative variable k of the j th remediation project, and k goes from 1 to p , where p is the number of quantitative variables under consideration. In this case p will equal 6. Note that the distance-measure is determined after the data has been normalized.

The similarity coefficient for the non-quantitative variables (S_b) will be determined using the following association table (Table 3.1) and equation (2.4):

Table 3.1 Association Table Construction

		Project A	
		+	-
Project B	+	a	b
	-	c	d

$$S_b = \frac{a + d}{a + b + c + d} \quad (2.4)$$

With two similarity coefficients, one from the quantitative parameters and one from the non-quantitative parameters, the need to combine the measures arises. For the quantitative parameters, the closer the value of the similarity coefficient is to 1, the more similar the projects and the closer to 0, the less similar the projects. The similarity coefficient for the non-quantitative, matching-type parameters behaves in the same manner. As the similarity coefficient approaches 1, the greater the number of matched pairs and the more similar the projects. As the coefficient value approaches 0, the less similar the projects are. Therefore, for both quantitative and non-quantitative parameters, a value of 1 means that the remediation projects are the same. Based on this, we may sum the two similarity coefficients to obtain an overall similarity coefficient (Backer, 1995). As described earlier, according to Backer (1995), an overall coefficient can be determined using equation (2.10):

$$s(i, j) = \alpha s_n(i, j) + (1 - \alpha) s_b(i, j) \quad (2.10)$$

where $s(i, j)$ is the combined similarity coefficient, α is a weighting factor, s_n is the similarity coefficient from quantitative parameters, and s_b is the similarity coefficient from non-quantitative parameters. For our analysis, each of the quantitative and non-quantitative parameters will be weighted equally. Since there are six quantitative parameters (depth to groundwater table, hydraulic conductivity, contaminant concentration (source), contaminant concentration (dissolved), remedial objective (source), and remedial objective (plume)) and the number of non-quantitative parameters is 1 (contaminant type), the weighting factor, α , will be equal to 6/7. This means that the quantitative similarity coefficient will be weighted 6/7 and the non-quantitative similarity

coefficient will be weighted $1/7$. A transformation will be used to subsequently allow us to apply Krzanowski's algorithm. Krzanowski's algorithm is based on 0 indicating identical projects. To accommodate this, the combined similarity measure will be subtracted from 1, so that 0 will identify identical objects and 1 will identify totally dissimilar objects.

Following Krzanowski's algorithm, each of the individual remediation projects will begin as individual clusters. The total similarity coefficient between each of the remediation projects will be determined and subtracted from 1 (to allow the application of the Krzanowski algorithm). The similarity coefficients will be stored in a symmetrical matrix with dimensions of $n \times n$ (where n represents the number of projects). The two remediation projects with the lowest similarity coefficient from the similarity coefficient matrix will be fused to form a new cluster. The similarity coefficient at which the fusion takes place now becomes the clustering coefficient for that cluster. A new similarity matrix will be calculated using the group average methods described above. The new similarity matrix will be symmetrical, but the dimension will decrease by one. Again, the lowest similarity coefficient is determined and the two clusters having the lowest similarity coefficient are fused to form a larger cluster. This process is repeated until one cluster is formed containing all the individual remediation projects.

With the clustering completed, it is appropriate to decide which clustering coefficient will be chosen to identify the clusters that will be used for the construction of the percent mass removal of the source versus total cost plot similar to Figures 2.1 and 2.2. As discussed in Chapter 2, several methods are available to determine the appropriate

number of clusters. This study will apply the “jump” analysis, a subjective heuristic for determining the optimal number of clusters. After completing the full hierarchical fusion of clusters, the clustering coefficients are visually examined to determine where a relatively large jump in the clustering coefficient is seen. The theory behind this analysis is that the relatively large jump in the clustering coefficient indicates relatively dissimilar objects are being clustered.

While other methods could be used to determine the number of clusters in the data set, the jump heuristic will be adequate for this study because the exact number of clusters does not need to be determined to create plots similar to those in Figures 2.1 and 2.2. We only need to group the data into meaningful clusters so we can determine that the projects within the cluster are more similar to one another than projects in other clusters.

There are computer programs available that will perform cluster analysis. However, this study will not use one of these programs in order to fully document the underlying aspects and methods of cluster analysis. Understanding the principles and methods of cluster analysis is vital to the novice user. Choosing a cluster analysis computer program can be challenging. Many of the programs do not allow for mixed data types, quantitative and non-quantitative. Others differ in clustering methods and similarity measure determination. Rather than modifying this study to meet the requirements of a computer program, it was determined that this study would be completed without the use of formal clustering software.

3.3.5 Cluster Analysis Example

In this section, the clustering technique described above is applied to a hypothetical data set so that each step of the grouping technique can be understood both in principle and computationally. Although the numbers used are fictitious, the parameters and methods used for this example are the same as those used to analyze the collected data.

For this example, 7 hypothetical environmental remediation projects will be clustered based on 7 parameters, 6 quantitative and 1 non-quantitative. The 6 quantitative parameters are depth to water table, hydraulic conductivity, contaminant concentration (source), contaminant concentration (dissolved), remedial objective (source), and remedial objective (plume). The non-quantitative parameter is contaminant type, which has four possible values (CVOC, BTEX, SVOC, and metals). The data shown in Table 3.2 are collected from the seven remediation projects and entered into a spreadsheet.

Table 3.2: Cluster Analysis Example: Raw Data

Project	Depth to water table (ft)	Hydraulic conductivity (ft/day)	Contaminant Concentration		Remedial Objective	
			Source (mg/kg)	Dissolved (mg/L)	Source (mg/kg)	Plume (mg/L)
1	5	.000005	200	.08	5	.05
2	25	.00025	1000	.1	5	.1
3	10	.0000001	750	5	2	.05
4	15	.001	500	.8	3	.05
5	75	.000001	100	100	5	.05
6	40	.00075	20	.5	10	.1
7	33	.000008	1500	20	1	.1

Project	Contaminant Type	% source removal	Total cost (\$ millions)	Cost Year
1	BTEX	62	7.5	1995
2	SVOC	60	2.1	1998
3	Metal	25	3.0	1996
4	SVOC	35	4.5	1995
5	CVOC	16	5.0	1993
6	CVOC	85	8.0	1997
7	BTEX	90	10.0	1991

The first step is to normalize the quantitative parameters that will be used in the cluster analysis using the standard deviation of each variable. The standard deviation for each quantitative parameter is listed in Table 3.3.

Table 3.3: Cluster Analysis Example: Quantitative Parameter Standard Deviations

Parameter	Standard Deviation
Depth to Groundwater Table	23.8117 ft
Hydraulic Conductivity	0.00041745 ft/day
Contaminant Concentration (Source)	539.5192 mg/kg
Contaminant Concentration (Dissolved)	36.8343 mg/L
Remedial Objective (Source)	2.9358 mg/kg
Remedial Objective (Dissolved)	0.0268 mg/L

Dividing the value of the parameter by its standard deviation will normalize the data in order to manage the scale invariance problem. Table 3.4 lists the normalized values for the quantitative parameters.

Table 3.4: Cluster Analysis Example: Standardized Data

Project	Depth to water table	Hydraulic conductivity	Contaminant Concentration		Remedial Objective	
			Source	Dissolved	Source	Plume
1	0.2100	0.0120	0.3708	0.0022	1.7031	1.8657
2	1.0499	0.5989	1.8538	0.0027	1.7031	3.7313
3	0.4200	0.0002	1.3904	0.1357	0.6812	1.8657
4	0.6299	2.3955	0.9269	0.0217	1.0219	1.8657
5	3.1497	0.0024	0.1854	2.7149	1.7031	1.8657
6	1.6798	1.7966	0.0371	0.0136	3.4062	3.7313
7	1.3859	0.0192	2.7808	0.5430	0.3406	3.7313

With the quantitative parameters normalized, it is now necessary to incorporate the non-quantitative parameters. For non-quantitative parameters, 1 represents the presence of an attribute and 0 represents the absence of an attribute at a remediation project. Table 3.5 lists the attributes for the 7 projects.

Table 3.5: Cluster Analysis Example: Non-quantitative Attributes

Project	Attribute 1 CVOC	Attribute 2 BTEX	Attribute 3 SVOC	Attribute 4 Metal
1	0	1	0	0
2	0	0	1	0
3	0	0	0	1
4	0	0	1	0
5	1	0	0	0
6	1	0	0	0
7	0	1	0	0

We may now begin the clustering process. The first step in the process is to determine the similarity coefficient between each of the projects. The quantitative parameters will

use the distance-type measure (which can be converted to a matching-type measure), expressed using equation (2.3):

$$d_{ij} = \left\{ \sum_{k=1}^p |X_{ik} - X_{jk}|^2 \right\}^{1/2} \quad (2.3)$$

For this example, $p = 6$, with 1 being depth to water table, 2 being hydraulic conductivity, 3 being contaminant concentration (source), 4 being contaminant concentration (dissolved), 5 being remedial objective (source) and 6 being remedial objective (plume).

X_{ik} and X_{jk} are the k th parameter values from projects i and j . Applying equation (2.3) leads to Table 3.6, which shows the distances between the quantitative parameters.

Table 3.6: Cluster Analysis Example: Quantitative Distance Measures

Project	1	2	3	4	5	6	7
1	0	2.5942	1.4649	2.5751	4.0044	3.4406	3.5850
2	2.5942	0	2.3478	2.8650	4.2878	2.8341	1.8592
3	1.4649	2.3478	0	2.4750	4.0743	4.1912	2.5747
4	2.5751	2.8650	2.4750	0	4.5104	3.3791	3.7245
5	4.0044	4.2878	4.0743	4.5104	0	4.3680	4.4611
6	3.4406	2.8341	4.1912	3.3791	4.3680	0	4.5223
7	3.5850	1.8592	2.5747	3.7245	4.4611	4.5223	0

Calculating $1/(1+d_{ij})$ yields the following similarity coefficients shown in Table 3.7 for the quantitative variables.

Table 3.7: Cluster Analysis Example: Matching-type Similarity Measures for Quantitative Parameters

Project	1	2	3	4	5	6	7
1	1.0000	0.2782	0.4057	0.2797	0.1998	0.2252	0.2181
2	0.2782	1.0000	0.2987	0.2587	0.1891	0.2608	0.3497
3	0.4057	0.2987	1.0000	0.2878	0.1971	0.1926	0.2797
4	0.2797	0.2805	0.2878	1.0000	0.1815	0.2284	0.2117
5	0.1998	0.1891	0.1971	0.1815	1.0000	0.1863	0.1831
6	0.2252	0.2608	0.1926	0.2284	0.1863	1.0000	0.1811
7	0.2181	0.3497	0.2797	0.2117	0.1831	0.1811	1.0000

The non-quantitative variables similarity coefficient will be calculated using association tables based on Table 3.5. An association table will be constructed for each combination of projects. Only the association table for projects 1 and 2 will be shown for space considerations.

		Project 1	
		+	-
Project 2	+	0	1
	-	1	2

This indicates that projects 1 and 2 have no instances where they share the same attributes, two instances where they both lack the same attribute, one instance where Project 1 has an attribute not present at Project 2, and one instance where Project 2 has an attribute not present in Project 1.

Using equation (2.4), the similarity coefficient can be determined. The similarity coefficient for project 1 and 2 is $\frac{0+2}{0+1+1+2}$ which equals .50. The process was repeated for each of the possible combinations of projects. The following similarity table (Table 3.8) for the non-quantitative variables was constructed to show the similarity coefficients for our example.

Table 3.8: Cluster Analysis Example: Matching-type Similarity Measures

Project	1	2	3	4	5	6	7
1	1	.50	.50	.50	.50	.50	1
2	.50	1	.50	1	.50	.50	.50
3	.50	.50	1	.50	.50	.50	.50
4	.50	1	.50	1	.50	.50	.50
5	.50	.50	.50	.50	1	1	.50
6	.50	.50	.50	.50	1	1	.50
7	1	.50	.50	.50	.50	.50	1

Now that similarity coefficients have been determined for both the quantitative and non-quantitative variables, the similarity coefficients can be combined based on equation (2.10), using $\alpha = 6/7$. The new similarity coefficient table is shown in Table 3.9.

Table 3.9: Cluster Analysis Example: Combined Similarity Matrix

Project	1	2	3	4	5	6	7
1	1.0000	0.3099	0.4192	0.3112	0.2427	0.2645	0.3298
2	0.3099	1.0000	0.3275	0.3646	0.2335	0.2950	0.3712
3	0.4192	0.3275	1.0000	0.3181	0.2403	0.2365	0.3112
4	0.3112	0.3833	0.3181	1.0000	0.2270	0.2672	0.2529
5	0.2427	0.2335	0.2403	0.2270	1.0000	0.3025	0.2284
6	0.2645	0.2950	0.2365	0.2672	0.3025	1.0000	0.2266
7	0.3298	0.3712	0.3112	0.2529	0.2284	0.2266	1.0000

The final step is to transform these similarity coefficients by subtracting from unity so that we may apply the algorithm presented by Krzanowski (1988). These results are shown in Table 3.10.

Table 3.10: Clustering Analysis Example: Clustering Similarity Matrix

Project	1	2	3	4	5	6	7
1	0.0000	0.6901	0.5808	0.6888	0.7573	0.7355	0.6702
2	0.6901	0.0000	0.6725	0.6354	0.7665	0.7050	0.6288
3	0.5808	0.6725	0.0000	0.6819	0.7597	0.7635	0.6888
4	0.6888	0.6354	0.6819	0.0000	0.7730	0.7328	0.7471
5	0.7573	0.7665	0.7597	0.7730	0.0000	0.6975	0.7716
6	0.7355	0.7050	0.7635	0.7328	0.6975	0.0000	0.7734
7	0.6702	0.6288	0.6888	0.7471	0.7716	0.7734	0.0000

With the similarity coefficients determined and transformed such that zero is the optimal similarity coefficient, the Krzanowski algorithm will be applied to cluster the data. The first step in the algorithm is to define each individual object as an individual cluster. Step 2 is to select the smallest similarity measure and form a new cluster containing both of the projects associated with the similarity measure. For our example, the lowest similarity coefficient is .5808. This similarity coefficient is between Projects 1 and 3. Projects 1 and 3 are fused at a cluster coefficient equal to the similarity coefficient of .5808. Table 3.11 will be the new similarity coefficient matrix incorporating the new cluster (combining Projects 1 and 3) into the table. The new similarity coefficient matrix will be determined using the group average expression (equation (2.13)) for all similarity coefficients involving the cluster 1,3. The similarity coefficients not involving cluster 1,3 will not change from the previous similarity coefficient matrix. The new similarity matrix will lose one dimension due to the new cluster being formed.

Table 3.11: Cluster Analysis Example: Similarity Coefficient Matrix for 6 Clusters

Project	1,3	2	4	5	6	7
1,3	0	.6813	.6854	.7585	.7495	.6795
2	.6813	0	.6354	.7665	.7050	.6288
3	.6854	.6167	0	.7730	.7328	.7471
5	.7585	.7665	.7730	0	.6975	.7716
6	.7495	.7050	.7328	.6975	0	.7734
7	.6795	.6288	.7471	.7716	.7734	0

An example calculation for the new similarity coefficients involving clusters with more than one member will be presented for cluster 1,3 and cluster 2. The calculation proceeds as follows:

$$D_{2,13} = \frac{n_i}{n_i + n_j} D_{ki} + \frac{n_j}{n_j + n_i} D_{kj} = \frac{1}{1+1} .6901 + \frac{1}{1+1} .6725 = .6813$$

where n_i is the number of objects in cluster 1, n_j is the number of objects in cluster 3, .6901 is the similarity coefficient of clusters 1 and 2 from Table 3.10, and .6725 is the similarity coefficient of clusters 2 and 3 from the Table 3.10 . Basically, the arithmetic mean of the similarities is being determined. Continuing, clusters 2 and 7 are now fused because the similarity coefficient of .6288 is the lowest value in the new similarity coefficient matrix. Clusters 2 and 7 are fused together at a clustering coefficient of .6288 and a new similarity coefficient matrix is formed (Table 3.12).

Table 3.12: Cluster Analysis Example: Similarity Coefficient Matrix for 5 Clusters

Project	1,3	2,7	4	5	6
1,3	0	.6804	.6854	.7585	.7495
2,7	.6804	0	.6913	.7690	.7392
4	.6854	.6913	0	.7730	.7328
5	.7585	.7690	.7730	0	.6975
6	.7495	.7392	.7328	.6975	0

The similarity coefficient for cluster 1,3 and cluster 2,7 is calculated using the following expression: (The remaining similarity coefficients were calculated as described above)

$$D_{13,27} = \frac{n_i}{n_i + n_j} D_{ki} + \frac{n_j}{n_j + n_i} D_{kj} = \frac{1}{1+1} .6813 + \frac{1}{1+1} .6795 = .6804$$

where n_i is the number of objects in cluster 2, n_j is the number of objects in cluster 7, .6813 is the similarity coefficient of clusters 1,3 and 2 from Table 3.11, and .6795 is the similarity coefficient of clusters 1,3 and cluster 7 from Table 3.11. The process repeats until ultimately, one cluster is formed. The clustering coefficient for each set of clusters is listed in Table 3.13.

Table 3.13: Cluster Analysis Example: Clustering Coefficient for 1 Cluster

6 Clusters	Clustering Coefficient = .5808
5 Clusters	Clustering Coefficient = .6288
4 Clusters	Clustering Coefficient = .6804
3 Clusters	Clustering Coefficient = .6883
2 Clusters	Clustering Coefficient = .6975
1 Cluster	Clustering Coefficient = .7538

Visual inspection of the clustering coefficients shows the largest jump between 5 and 4 clusters. This means that two relatively dissimilar clusters were joined to form 4 clusters. Therefore, 5 clusters should be used in subsequent analysis to create plots similar to those in Figures 2.1 and 2.2. It should be noted that a large jump in clustering coefficients is also seen between 2 and 1 clusters. However, this jump is seen higher in the hierarchy than the jump between 4 and 5, so 5 clusters is determined to be appropriate for subsequent analysis.

3.4 Using Clustered Data to Create Lifecycle Cost versus Percent Source Removal Plots

We will now discuss the methods for creating the lifecycle cost versus percent source removal plots similar to those depicted in Figures 2.1 and 2.2. With the optimal number of clusters determined heuristically, the individual remediation projects are grouped so that the within-group variation compared to the between-group variation is minimized. We will therefore assume that the remediation projects within one group are similar enough to directly compare on a single source removal versus lifecycle cost plot.

The remediation projects within the cluster will each have a certain percent source removal associated with the project as well as a certain lifecycle cost of the project

(presented in the raw data set). The lifecycle cost of the project will be normalized by the mass of contaminant treated so that projects of different mass of contaminant treated can be compared on a dollar per unit mass treated basis. A plot will be constructed for each cluster. The horizontal axis of the plot will contain the range of source mass removal (0-100%), while the vertical axis will represent the normalized NPV (in year 2000 dollars). Using these axes, the cost versus source removal data for the individual remediation projects within the selected cluster will be constructed. These plots will be similar to those presented in Figures 2.1 and 2.2. The plots that are created may be used to validate, with real world data, the conceptual view of Kavanaugh and Goldstein (1999).

4.0 Results

4.1 Overview

The results of the cluster analysis and plots of lifecycle cost versus percent source removal described in the previous chapter are reported here.

4.2 Cluster Analysis

Data were collected from 72 environmental remediation projects and are presented in Appendix A. These data were used to cluster the remediation projects as described in Chapter 3. Appendix B lists the clustering coefficients associated with each step in the hierarchy. By performing the “jump” method on the clustering coefficients, the optimal number of clusters was determined to be 24 with a clustering coefficient of .492971. The difference between clustering coefficients for 24 and 23 clusters is .040122. Except for the difference in coefficients between 2 and 1 clusters, this is the largest jump in clustering coefficients. As mentioned in Chapter 3, a large, unexpected jump in clustering coefficients indicates that two relatively dissimilar clusters were fused, resulting in the large increase in clustering coefficients.

The distribution of projects over the 24 clusters is included in Appendix C. Examination of Appendix C shows 1 cluster with 28 projects, 1 cluster with 11 projects, 1 cluster with 5 projects, 1 cluster with 3 projects, 5 clusters with 2 projects, and 15 clusters with a single project. For the purposes of this study, clusters with less than 10 projects will not be considered because these clusters would not prove useful for subsequent analysis.

Cluster 1 and 2 will represent the clusters with 28 and 11 projects, respectively. The following discussion focuses on these individual clusters.

4.2.1 Cluster 1

Table 4.1 lists the projects and parameter values for each of the projects in Cluster 1.

Table 4.1: Cluster 1 Projects and Parameter Values

Project Number	Depth to Water Table (ft)	Hydraulic Conductivity (ft/day)	Contaminant Concentration		Remedial Objective		Contaminant Type
			Source (mg/kg)	Dissolved (mg/L)	Source (mg/kg)	Plume (mg/L)	
2	45	100	1	0.12	1	0.006	CVOC, BTEX
4	11	1.54	400	920	43	0.1	CVOC, BTEX, SVOC
6	10	0.79	1	70	1	0.005	CVOC
7	27.5	42.5	0.0228	4	0.0228	0.007	CVOC
8	14	10	1	0.13	1	0.005	CVOC, Metals
9	22.5	12.82	600	0.79	25	0.005	CVOC
10	20	27.25	1	200	1	0.2	CVOC
16	26.5	150.1	61.2	30	1	0.005	CVOC, BTEX, metals
17	15	40	16	140	16	2.9	CVOC, BTEX, metals
19	7.5	50.5	84.2697	250	1	0.07	CVOC, Metals
24	20	1.7	39.3	10	1	0.16	CVOC
28	5	0.0015	100	4.1	100	0.007	BTEX
34	15	7.085	550	92	0.06	0.003	CVOC, BTEX
36	5.5	5.5	16	0.446	1	0.005	CVOC, BTEX
37	65	68.5	210	10	1	0.005	CVOC
38	8	190	7.8	4.7	0.02	0.005	BTEX
44	25	9	876.39	223	1	0.005	CVOC
45	12.5	141.8377	360	240	1	0.005	CVOC
47	15	0.35	0.02	2.807	0.02	0.029	BTEX
54	4	70.98	1400	4.1	100	0.001	SVOC
59	13.75	2.95	280	0.005	100	0.005	SVOC
62	3.75	28.34	420	0.01	50	0.005	SVOC, CVOC, BTEX
63	35	22.73	3400	142	0.2	0.005	CVOC, BTEX
64	46	28.34	1300	1.9	1	0.001	CVOC, BTEX
65	33	28.34	280	460	1	0.001	SVOC
66	35	2.83	46	37	1	0.005	CVOC
68	4	12.3	8.61434	11.7	1	0.005	CVOC
71	13.5	2.834	1	16	1	0.007	CVOC, BTEX

Insight into the projects that have been clustered may be gained by examining the parameter values of the projects in the cluster. Table 4.2 lists the range, average value, and standard deviation for each of the quantitative parameters, as well as the frequency of the non-quantitative parameter values, for the projects in Cluster 1.

Table 4.2: Cluster 1: Parameter Range, Average, Standard Deviation, and Frequency

	Depth to Water Table (ft)	Hydraulic Conductivity (ft/day)	Contaminant Concentration		Remedial Objective		Contaminant Type
			Source (mg/kg)	Dissolved (mg/L)	Source (mg/kg)	Plume (mg/L)	
Range	3.75-65	.0015-190	.02-3400	.005-920	.02-100	.001-2.9	
Average	19.929	37.825	373.593	102.67	16.083	.127	
Standard Deviation	14.927	50.424	705.129	195.2	32.205	.546	
Frequency							CVOC - 10 BTEX - 3 SVOC - 3 CVOC, BTEX - 6 CVOC, metals - 2 CVOC, BTEX, SVOC - 2 CVOC, BTEX, Metals - 2

As described in Chapter 2, each environmental remediation project can be characterized by a number of parameters. The projects in Cluster 1 can be described as having the parameter values listed in Table 4.2. Inspection of the parameter values indicates that the depth to groundwater table for each of the projects in this cluster is 65 feet or less with the vast majority of the values less than 45 feet. The values for hydraulic conductivity, contaminant concentration (source), and contaminant concentration (dissolved) are widely dispersed and do not allow for any generalizations. The vast majority of parameter values for remedial objective (source) are 25 mg/kg or less which allows for a generalization that projects belonging to Cluster 1 will typically have remedial objective

(source) values below 25 mg/kg. A generalization can also be made for remedial objective (plume) as projects in Cluster 1 typically have remedial objective (plume) values on the order of $1\text{e-}3$ mg/L. The vast majority of the projects in Cluster 1 also have CVOCs present either individually or in conjunction with another contaminant type. By describing the cluster based on typical parameter values, we allow one set of parameter values to represent a much larger number of remediation projects that have been determined to be more similar to each other than any other projects within the data set. This is to say that a project manager who is interested in obtaining information on environmental remediation projects in which the depth to groundwater is 45 ft or less, remedial objective (source) is 25 mg/kg or less, remedial objective (plume) is on the order of $1\text{e-}3$ mg/L and contain CVOC contamination could use the projects listed in Cluster 1 because they have been determined to be more similar to one another than any other projects within the data set (as long as the range of hydraulic conductivity, contaminant concentration (source) and contaminant concentration (dissolved) values from Cluster 1 are acceptable to the project manager).

4.2.2 Cluster 2

Table 4.3 lists the projects and parameter values for each of the projects in Cluster 2.

Table 4.3: Cluster 2 Project and Parameter Values

Site Number	Depth to Water Table (ft)	Hydraulic Conductivity (ft/day)	Contaminant Concentration		Remedial Objective		Contaminant Type
			Source (mg/kg)	Dissolved (mg/L)	Source (mg/kg)	Plume (mg/L)	
23	30	14.17	10	3	10	11.705	BTEX
25	36.5	28.37	9200	14.3	10	11.705	SVOC, BTEX
30	15	70.98	10000	26.576	10	11.705	BTEX
31	18	28.47	565.778	43.28	10	11.705	BTEX
32	10.5	70.84	614.408	47	10	11.705	BTEX
33	5	0.0015	425	1.81	10	11.705	BTEX
40	105	1.41838	5040	10	38.1	11.705	BTEX
41	50	1.4E-07	10200	11.705	38.1	11.705	BTEX
48	80	2.834	11000	10	23	10	BTEX, CVOC
67	6	85.0176	130.725	10	10	11.705	BTEX
70	30	14.17	592	1	10	11.705	BTEX

Table 4.4 lists the range, average value, and standard deviation for each of the quantitative parameters as well as the frequency of the non-quantitative parameter values.

Table 4.4: Cluster 2: Parameter Range, Average, Standard Deviation and Frequency

	Depth to Water Table (ft)	Hydraulic Conductivity (ft/day)	Contaminant Concentration		Remedial Objective		Contaminant Type
			Source (mg/kg)	Dissolved (mg/L)	Source (mg/kg)	Plume (mg/L)	
Range	5-105	1e-7-85.02	10-11000	1-47	10-38.1	10-11.705	
Average	35.09	28.752	4343.45	16.24	16.29	11.55	
Standard Deviation	31.99	31.97	4783.37	15.94	11.46	.514	
Frequency							BTEX – 9 BTEX, SVOC – 1 CVOC, BTEX - 1

The projects in Cluster 2 can be described as having the parameter values listed in Table 4.4. Inspection of the parameter values indicates that the depth to groundwater table for each of the projects in this cluster is 105 feet or less with the vast majority of the values at 50 feet or less. Again the values for hydraulic conductivity, contaminant concentration (source), and contaminant concentration (dissolved) are widely dispersed and do not allow for any generalizations. The vast majority of parameter values for remedial

objective (source) are 10 mg/kg so we can generalize that projects belonging to Cluster 2 will typically have remedial objective (source) values of 10 mg/kg. A generalization can also be made for remedial objective (plume) as projects in Cluster 2 typically have remedial objective (plume) values of 11.705 mg/L. All of the projects in Cluster 2 also have BTEX contamination present either individually or in conjunction with another contaminant type.

4.3 Cost Versus Percent Removal

As described in Chapter 3, once the clusters have been determined, the lifecycle cost of the remediation project will be plotted versus the percent source removal achieved at the project. The lifecycle cost of the project will be normalized using the total mass of contaminant treated at the project, yielding dollars per kg treated at the project. Table 4.5 and 4.6 list the mass of contaminant treated, lifecycle cost, normalized lifecycle cost and percent removal for Cluster 1 and 2, respectively.

Table 4.5: Cluster 1 Mass Contaminant Treated, Lifecycle Cost, Normalized Cost and Percent Source Removal

Project	Mass contaminant Treated (kg)	Lifecycle Cost (year 2000 Dollars)	Normalized Cost (\$/Kg)	Percent Removal
2	1251	26530401	21200	0
4	364904	103766201	284	100
6	8532	23986405	2811	0
7	148	7097421	47914	0
8	6	8891718	1540291	0
9	1419	7583556	5343	95.83
10	40085	8042693	201	0
16	8409	76267703	9070	0
17	91444	45172785	494	100
19	473044	210355915	445	0
24	58514	13735737	235	0
28	3477	3998409	1150	0
34	14985	5305803	354	95
36	146	3540366	24189	93.75
37	72136	51658637	716	87.07
38	2663	358958	135	99
44	49671662	20537861	0	9
45	4038306	38600000	10	99
47	180	2121045	11788	100
54	105745	164071527	1552	100
59	2312	935414	405	100
62	13111	38862089	2964	14.58
63	3110	716645	230	100
64	13276	25636905	1931	83.7
65	193365	75551923	391	100
66	633	597270	944	93.7
68	320	201042	627	99
71	3587	686600	191	0

Table 4.6: Cluster 2 Mass Contaminant Treated, Lifecycle Cost, Normalized Cost and Percent Source Removal

Project	Mass Contaminant Treated (Kg)	Lifecycle Cost (Year 2000 Dollars)	Normalized Cost (\$/Kg)	Percent Source Removal
23	12269	5060612.9	412.47	39.8
25	117265	1413879.9	12.06	52.1
30	167625	559979.84	3.34	20
31	5346	1765830.3	330.31	70
32	4290	346979.17	80.88	0
33	76	164791.67	2170.93	0
40	87319	323246.22	3.70	48.1
41	18342	1079763.7	58.87	98
48	6927	706447.27	101.98	99
67	298469	1588353.6	5.32	99.5
70	367826	498834.5	1.36	90

Looking at Tables 4.5 and 4.6, we observe that the normalized costs (\$/Kg), especially for Cluster 2 projects, appear relatively low. The reason for this is that, as discussed in Section 3.2.2, maximum contaminant concentration was used to estimate the mass of contaminant in the subsurface. This results in the mass of contaminant in the subsurface being overestimated, leading to an overestimate of the mass of contaminant treated and an underestimate of the normalized cost. Typical normalized cost values (\$/Kg treated) for plume containment technologies in the field have been observed ranging from \$330/Kg to \$5650/Kg, with some projects costing as much as \$77,000/Kg (EPA, 2000). Normalized cost values (\$/Kg treated) for source removal technologies in the field have been observed ranging from <\$1/Kg to >\$250000/Kg (EPA, 2000) depending on project characteristics and technology selection. Overall, cost per mass removed is dependent on the mass of contaminant in the subsurface and type of treatment technology used. While it is difficult to compare normalized costs when we have a combination of source removal and containment technologies being applied, we may gain some insight by focusing on projects that only applied plume containment (0% source removal). Looking

at the 12 projects in both Clusters 1 and 2 that have 0% source removal, we find five of the projects have normalized costs for plume containment within the typical range noted above. Four of the seven projects with normalized costs outside of the typical range have costs below the lower bound of the range. The costs for these four projects may be artificially low as a result of using the maximum contaminant concentration to calculate mass treated. Thus, it appears that the majority of projects under consideration (with the exception of 3 very high cost Cluster 1 projects) have normalized costs for plume containment within the typical range.

Looking at Tables 4.5 and 4.6, it appears that projects in Cluster 1 have higher normalized costs than projects in Cluster 2. In fact, even if we do not consider the three most expensive Cluster 1 projects, the average normalized cost of a Cluster 1 project is \$3600/Kg while the average normalized cost of a Cluster 2 project is \$300/Kg. This is because most remediation projects in Cluster 1 incorporated pump and treat technology for plume treatment. Pump and treat technologies are typically very expensive in terms of cost/Kg contaminant treated. In addition, projects in Cluster 1 incorporate a variety of source removal technologies, including excavation, which is very expensive. Projects in Cluster 2 used treatment technologies such as bioventing, free product recovery and natural attenuation, that are relatively cheap in terms of cost/Kg contaminant treated.

Figures 4.1 and 4.2 illustrate normalized cost versus percent source removal for Cluster 1 and 2 respectively.

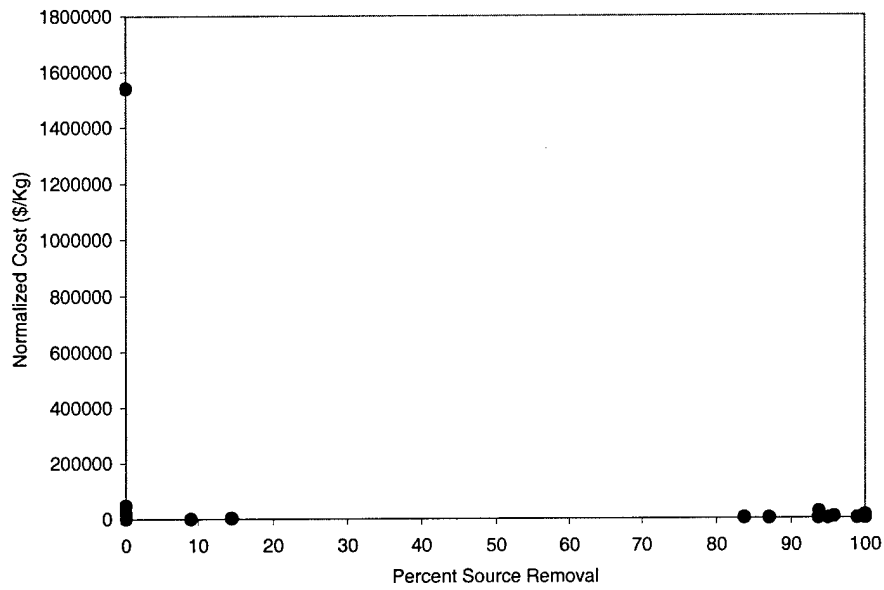


Figure 4.1: Cluster 1 Plot of Normalized Cost Versus Percent Source Removal for All Data Points

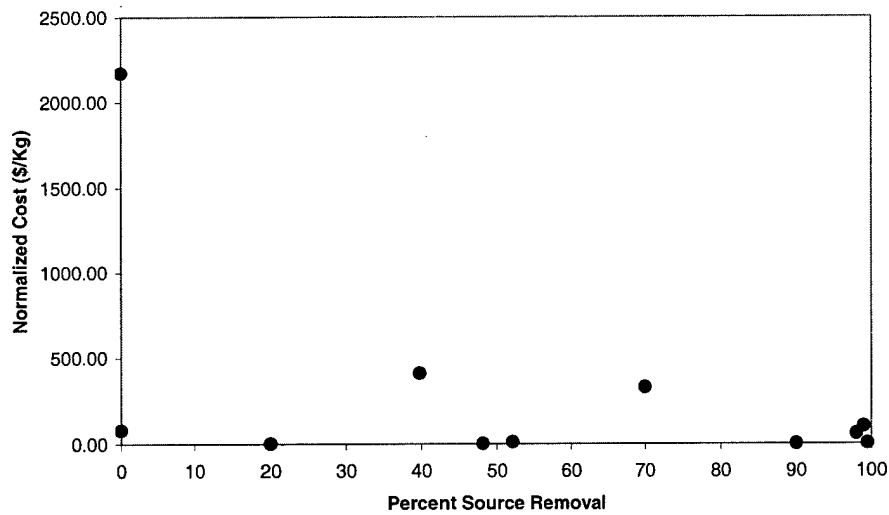


Figure 4.2: Cluster 2 Plot of Normalized Cost Versus Percent Source Removal for All Data Points

In order to gain a better understanding of the above plots, the scale of the normalized cost axis was adjusted to remove some of the outlying points to better show the majority of the data points (Figure 4.3 and 4.4).

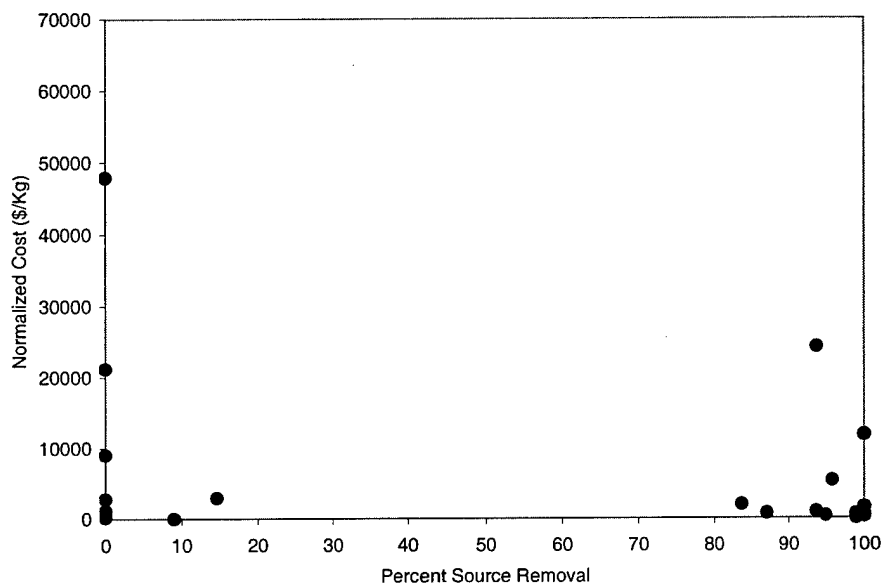


Figure 4.3: Cluster 1 Plot of Normalized Cost Versus Percent Source Removal (Normalized Cost Scale Reduced)

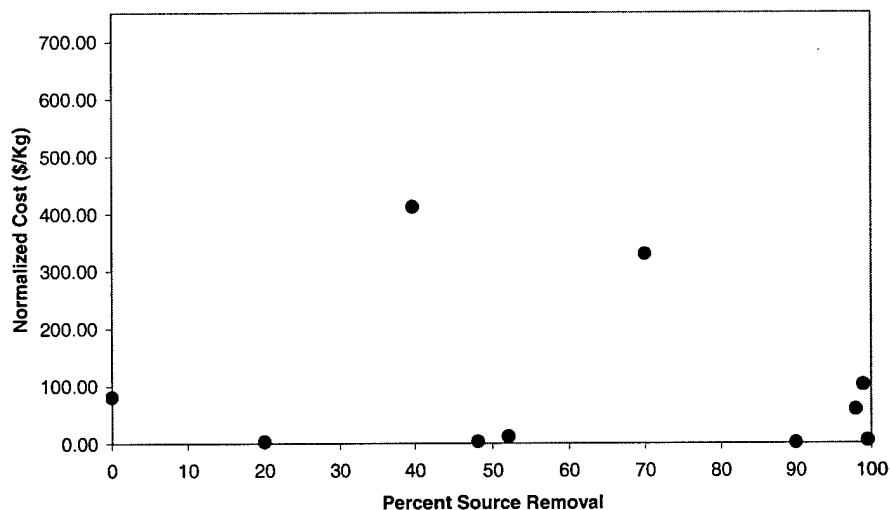


Figure 4.4: Cluster 2 Plot of Normalized Cost Versus Percent Source Removal (Normalized Cost Scale Reduced)

Looking at Figures 4.1 and 4.3 for Cluster 1, one observes the highest normalized lifecycle costs near 0 and 100% source removal with lower values observed at intermediate source removal fractions. Examining Figures 4.2 and 4.4 for Cluster 2, one observes high and low normalized lifecycle costs across the range of source removal fractions with no apparent relationship between lifecycle cost and percent source removal.

5.0 Conclusions

5.1 Summary

In this thesis, cluster analysis was used to group a data set of 72 diverse environmental remediation projects. Seven parameters, six quantitative and one non-quantitative, were used to characterize the projects and perform a hierarchical cluster analysis. Applying the "jump" method to the clustering coefficients, an optimal number of clusters within the data set was identified. The remediation projects in the resulting clusters demonstrate more similarity to remediation projects within the group than to remediation projects in other groups. The optimal number of clusters was determined to be 25, with the largest two clusters containing 28 and 11 environmental remediation projects. The remediation projects within these largest clusters were used to develop plots of lifecycle cost versus percent source removal, similar to Figures 2.1 and 2.2.

5.2 Lifecycle Cost versus Percent Source Removal Plots

The objective of developing lifecycle cost versus percent source removal plots was to validate, with real world data, the concept of Kavanaugh and Goldstein (1999) that for an environmental remediation project a minimum lifecycle cost exists at a certain percent source removal. It was assumed that remediation projects within a given cluster are similar enough to directly compare on a single lifecycle cost versus source removal plot. The lifecycle cost versus percent source removal plots for Cluster 1 (28 projects) and Cluster 2 (11 projects) are presented in Figures 4.1 and 4.2. The lifecycle cost for each project was normalized using the mass of contaminant treated by each project.

Examination of Figures 4.1 and 4.2 shows that no definite conclusion can be drawn concerning the existence of a minimum lifecycle cost at a certain percent source removal. This is to say that the results of this thesis cannot validate nor contradict the conceptualization of Kavanaugh and Goldstein (1999). Examination of Figures 4.1 and 4.3 may provide some support for a minimum lifecycle cost between 0 and 100% source removal. One may speculate that the points on the plot may be fit with a parabola (since the highest normalized lifecycle costs can be seen near 0 and 100% source removal, with lower values at intermediate source removal fractions). However, as the data show no apparent correlation between lifecycle cost and source removal percentage, no conclusion concerning the validity of Kavanaugh and Goldstein (1999)'s conceptualization can be made.

Examination of Figure 4.2 and 4.4 provides little, if any, support for Kavanaugh and Goldstein (1999)'s conceptualization. The data show no apparent correlation between lifecycle cost and percent source removal, so no conclusion concerning the validity of the Kavanaugh and Goldstein (1999) conceptualization can be made.

5.3 Utility of Cluster Analysis

While no attempt was made to validate the clusters used in this study, some insight into the applicability of cluster analysis to environmental remediation can be gained by looking at the clustering coefficients and considering the generalizations made in Sections 4.2.1 and 4.2.2. Recall that a clustering coefficient value of 0 indicates that the fused clusters were exactly identical, while a coefficient value that approaches 1 indicates the fused clusters were very dissimilar. Examining the clustering coefficients in

Appendix B, one observes very low clustering coefficients at lower levels of the hierarchy (which is to be expected since we fused clusters based on the minimum value of the clustering coefficients). Clustering coefficient values observed in this study, especially values near 0, indicate that cluster analysis is very applicable to grouping environmental remediation projects. Cluster analysis is applicable to both quantitative and non-quantitative data over a wide range of parameters (large number of parameters), which is the situation one faces when examining environmental remediation projects.

The cluster analysis in this study resulted in groups of remediation projects that are similar with respect to the other remediation projects in the data base. The clusters could change if different parameters are examined or more data are added to the data base. The key to forming valid clusters of environmental remediation projects is to increase the number of observations in the data base and ensure the parameters under consideration effectively characterize and differentiate the projects. It is possible that increasing the number of projects and/or parameters under consideration may lead to new clusters which would produce results different than those presented in this study.

Since clusters can change based on number of projects included in the analysis or the parameters used to determine the cluster, one may be interested in determining the "correct" set of clusters. No one validation method for cluster analysis is generally applied. Some individuals suggest that no effective method exists for assessing the validity of a cluster analysis solution (Walker, 1998). Others suggest simply replicating the cluster analysis using different similarity measures and algorithms. If true clusters exist in the data set, then they should be apparent independent of the clustering method

(Aldenderfer and Blashfield, 1984). Another option for validating clusters is to use indices, similar to and including the Davies-Bouldin Index described in Section 2.4.1.5. These indices can give a general indication of cluster validity based on the behavior of the index values across the hierarchy. Finally, a multivariate analysis of variance may be applied to determine the degree of differentiation between clusters (Walker, 1998). Basically, validating a cluster analysis remains an open research question. Ultimately, it is up to the investigator to decide whether or not a set of clusters is valid for application to a particular problem.

5.4 Limitations

In order to set the stage for our future research recommendations, it will be useful to discuss the limitations of the current study. One limitation is how parameters were selected to characterize projects. In this study, the parameters were selected based on expert opinion. However, parameters that were not considered may affect project performance and cost. Techniques, such as multivariate data analysis, are available to determine the most influential parameters that affect cost and performance. An analysis of the parameters listed in Chapter 2 (Tables 2.1, 2.2 and 2.3) may be necessary to ensure the parameters used in this study are the appropriate parameters to use to cluster the data. As demonstrated in Chapter 4, hydraulic conductivity and contaminant concentration (dissolved) were widely variable within the clusters. This may indicate that these parameters may not be influential enough to use as clustering parameters.

Another problem in the current study concerns the methodology used to normalize the lifecycle cost. This study used maximum contaminant concentration to determine the mass of contaminant in the subsurface, which inflated the total mass of contaminant treated. The increased mass of contaminant treated resulted in low normalized cost (\$/Kg) values. A more appropriate method for determining the mass of contaminant treated would be to use the actual mass treated, if available, or base the mass of contaminant in the subsurface on the average contaminant concentration, if available.

5.5 Recommendations for Future Study

While the actual results of this study did not permit validation of the Kavanaugh and Goldstein (1999) conceptualization, the methods used to create the lifecycle cost versus percent source removal plots can serve as the foundation for future research. The following discussion provides recommendations for future research to determine the value of source removal.

1. **Reexamine Assumptions and Parameter Selection** – Perform an analysis, such as multivariate data analysis, on parameters listed in Tables 2.1, 2.2 and 2.3 to determine what parameters are most useful in discriminating between projects. Incorporate these parameters into the methodology. Reexamine the assumption that remedial project managers selected the lowest cost alternative at each individual project. Political and regulatory constraints may have inhibited the application of the lowest cost alternative. It may be necessary to use type of technology in the cluster analysis to obtain clusters that incorporate the same or similar technologies, thus eliminating possible discrepancies in normalized cost due to different technologies. Another implicit assumption that should

be reexamined concerns the economies of scale. This study assumed that project scale did not affect lifecycle cost. This may or may not be an appropriate assumption that should be tested in a future study. Another consideration for parameter selection deals with the normalization of the lifecycle cost (which as noted earlier, created problems). More appropriate parameter selection may eliminate the need to normalize cost, especially if the remediation projects can be grouped in such a manner that the volume and mass of contaminant removed are similar within the cluster. Therefore, volume of contamination and mass of contaminant removed should be considered as possible parameters for the cluster analysis. Another assumption that must be reexamined is the assumption that the clusters formed in this analysis were valid. This would require that the clusters be validated, perhaps using methods described in Section 5.3.

2. Data Gathering and Data Gap Management – With the proper parameters selected, gather data from as many remediation projects as possible to increase the overall size of the constructed data base. Project status reports, remedial investigation and feasibility studies, and remedial project managers are the best sources of information for each individual remediation project. Data gap management should be reexamined to ensure the methods in this study are most appropriate. A more appropriate method may be to use minimum, maximum and average parameter value when clustering rather than one fixed parameter value.

3. Computerized Clustering – As noted in Chapter 3, computer programs are available for cluster analysis. Caution must be used, however, when selecting computer programs as some programs do not allow for quantitative and non-quantitative parameters. Other

considerations arise regarding the clustering methods incorporated in the computer programs. Some programs use clustering methods different than those used in this study, such as K-means clustering, or use different methods for determining the similarity coefficients (please see Chapter 2 for discussion of different clustering methods). One must be fully aware of the methods and underlying assumptions when using computer programs to perform cluster analysis.

4. Field Users – If the conceptualization of Kavanaugh and Goldstein (1999) is validated and it is proven that a minimum lifecycle cost is seen at a certain percent source removal, the information must be made accessible to remedial project managers. Remedial project managers must be able to use the characteristics of a particular project to determine which lifecycle cost versus percent source removal plot is appropriate. One possible method is to explicitly define the project characteristics of each plot (characteristics of the cluster) and allow the remediation project manager to select the most appropriate plot based on a comparison of project characteristics. Another possibility is to have the remedial project manager enter the characteristics of a particular project into a computer program, which clusters the new project with projects already included in a data base. This will determine which cluster the new project belongs to, and thus, the appropriate plot of lifecycle cost versus percent source removal.

5. Focus on Other Methods – Finally, it may be necessary to shift focus away from validating the Kavanaugh and Goldstein (1999)'s conceptualization using real world data. Real world data from individual remediation projects may be too complex to form meaningful groups, which is to say that individual remediation projects would not be

similar enough to plot on the same lifecycle cost versus percent source removal plot. One solution to this problem is to model different source removal fractions and ascertain the subsequent effect on lifecycle cost. Modeling would require a great deal of simplifying assumptions, but could provide a useful approach to validating (or disproving) the concept suggested by Kavanaugh and Goldstein (1999).

BIBLIOGRAPHY

- Aldenderfer M. S. and R. K. Blashfield. Cluster Analysis: Sage University Paper series on Quantitative Applications in the Social Sciences, series no 07-044, Newbury Park, CA: Sage Publications, 1984.
- Allard, J., V. Choulakian, R. LeBlanc, S. MacNeill, and S. Mahdi, Analysis 4: Discriminant analysis of seal data, *The Canadian Journal of Statistics*, 28(1): 205-212, 2000.
- American Society for Testing and Materials, *Standard Guide for Site Characterization for Environmental Purposes with Emphasis on Soil, Rock, the Vadose Zone and Groundwater*, Report D5730-98, 1998.
- Anderberg, M. R., Cluster Analysis for Applications. New York: Academic Press, 1973.
- Backer, E., Computer-assisted Reasoning in Cluster Analysis, New York: Prentice Hall, 1995.
- Defense Environmental Restoration Program, *FY 1999 Annual Report to Congress*, Internet access to report. <http://156.80.6.61/derparc/derp/default.htm>, 2 July, 2000.
- Dillon, W. R. and M. Goldstein, Multivariate Analysis Methods and Applications, New York: John Wiley and Sons, 1984.
- Domenico P. A. and F. W. Schwartz, Physical and Chemical Hydrogeology, New York: John Wiley and Sons, 1998.
- DuTeaux, S. B., *A Compendium of Cost Data for Environmental Remediation Technologies 2nd Edition*, DOE Report LA-UR-96-2205 (Internet Access: <http://www.lanl.gov/projects/etcap/>), 1997.
- Federal Remediation Technology Roundtable (FRTR), *FRTR Cost and Performance Remediation Case Studies and Related Information*, Report EPA 542-C-00-001, 2000.
- Federal Remediation Technology Roundtable (FRTR) 1998a, *Guide to Documenting and Managing Cost and Performance Information for Remediation Projects*, Report EPA 542-B-98-007, 1998.
- Federal Remediation Technology Roundtable (FRTR) 1998e, *Remediation Case Studies: Ex Situ Soil Treatment Technologies*, Report EPA 542-R-98-011, 1998.
- Federal Remediation Technology Roundtable (FRTR) 1998b, *Remediation Case Studies: Groundwater Pump and Treat (Chlorinated Solvents)*, Report EPA 542-R-98-013, 1998.

- Federal Remediation Technology Roundtable (FRTR) 1998c, *Remediation Case Studies: Groundwater Pump and Treat (Nonchlorinated Solvents)*, Report EPA 542-R-98-014, 1998.
- Federal Remediation Technology Roundtable (FRTR) 1998d, *Remediation Case Studies: In Situ Treatment Technologies*, Report EPA 542-R-98-012, 1998.
- Federal Remediation Technology Roundtable (FRTR), *Remediation Technologies Screening Matrix and Reference Guide, Second Edition*, Report EPA 542-B-94-013, 1994.
- Federal Remediation Technology Roundtable (FRTR), *Remediation Technologies Screening Matrix and Reference Guide, Third Edition*, Report NTIS PB98-108590, 1997.
- Goltz, M. N., Professor Air Force Institute of Technology, Personal Communication, 2000.
- Hamil, C, Remedial Project Manager Solid State Circuits Missouri Superfund Site, Personal Communication, 2000.
- Hannappel S., and B. Piepho, Cluster Analysis of Environmental Data Which is not Interval Scaled but Categorical, *Chemosphere*, 33(2): 335-342, 1996.
- Kavanaugh, M. and K. Goldstein, Presentation Handouts "Overview of DNAPL Source Zone Characterization/Remediation Technologies," SERDP/ESTCP Partners in Environmental Technology Symposium and Workshop. Arlington, Virginia. 1 December 1999.
- Krzanowski, W. J., Principles of Multivariate Analysis A User's Perspective. Oxford: Clarendon Press, 1988.
- Mojena, R., Hierarchical Grouping Methods and Stopping Rules: An Evaluation, *The Computer Journal*, 20(4): 359-362, 1977.
- Morris, C., Academic Press Dictionary of Science and Technology, San Diego: Academic Press, 1992.
- National Archives and Records Administration, *Code of Federal Regulations*, CFR 40-141. Washington: US Government Printing Office, 2000.
- Parsons Engineering Science, Inc. *Natural Attenuation of Fuel Hydrocarbons Performance and Cost Results from Multiple Air Force Demonstration Sites*, San Antonio: Air Force Center for Environmental Excellence Technology Transfer Division, 1999.
- Sellers, K., Fundamentals of Hazardous Waster Site Remediation. Boca Raton: Lewis Publishers, 1998.

- Stanin, F. T., M. B. Phelps, J. W. Ratz, D. C. Downey, A. Leeson, M. Jenner, P. E. Haas, R. N. Miller, *A General Evaluation of Bioventing for Removal Actions at Air Force/Department of Defense Installations Nationwide*, San Antonio: U.S. Air Force Center For Environmental Excellence Technology Transfer Division, 1996.
- Terhune, J., Introduction to the harp seal data set, *The Canadian Journal of Statistics*, 28(1): 183-185, 2000.
- United States Air Force, *1999 Raw Inflation Indices*, Washington: Assistant Secretary of the Air Force (Financial Management and Comptroller), Internet Access to Report: <http://www.saffm.hq.af.mil/FMC/inf199/inf199.html>, 2000.
- United States Environmental Protection Agency (US EPA), *Evaluation of Groundwater Extraction Remedies Volume 2 - Case Studies 1-19*, Report EPA 540-2-89-054b, Washington: Office of Emergency and Remedial Response, 1989.
- United States Environmental Protection Agency (US EPA), *Groundwater Cleanup: Overview of Operating Experience at 28 Sites*, Report EPA 542-R-99-006, Washington: Office of Solid Waste and Emergency Response, 1999.
- United States Environmental Protection Agency (US EPA), *Progress Toward Implementing Superfund Fiscal Year 1990 Report to Congress*, Report EPA 540-8-91-004, Washington: Office of Emergency and Remedial Response, 1992.
- United States Environmental Protection Agency (US EPA), *Progress Toward Implementing Superfund Fiscal Year 1991 Report to Congress*, Report EPA 540-R-94-016, Washington: Office of Solid Waste and Emergency Response, 1994.
- United States Environmental Protection Agency, *EPA Reach it Database*, Internet Access: <http://www.epareachit.org/index3.html>, 2001.
- United States Environmental Protection Agency (US EPA), *ROD Annual Report FY 1991: Volume 1*, Publication 9355.6-05-1, Washington: Office of Emergency and Remedial Response, 1992a.
- United States Environmental Protection Agency (US EPA), *ROD Annual Report FY 1991: Volume 2*, Publication 9355.6-05-2, Washington: Office of Emergency and Remedial Response, 1992b.
- United States Environmental Protection Agency (US EPA), *ROD Annual Report FY 1992*, Publication 9355.6-06, Washington: Office of Solid Waste and Emergency Response, 1993.
- United States Environmental Protection Agency (US EPA), *Site Characterization for Subsurface Remediation*, Report EPA 625-4-91-026, 1991.

- United States Environmental Protection Agency (US EPA), *State Coalition for Remediation of Dry Cleaners Site Profiles*, Internet Access to Report: <http://www.drycleancoalition.org/profiles>, 2001.
- United States Environmental Protection Agency Region 6 (US EPA 6), *Superfund Site Status Summaries and Site Photos*, Internet Access: <http://www.epa.gov/earth1r6/6sf/6sf.htm>, Jan 2001.
- United States Geological Survey (USGS), Internet Access to National Groundwater Levels: <http://water.usgs.gov/>, 2001
- Verhaar, H. J. M., C. J. van Leeuwen, and J. L. M. Hermans, Classifying Environmental Pollutants. 1: Structure-Activity Relationships for Prediction of Aquatic Toxicity, *Chemosphere*, 25(4): 471-491, 1992.
- Walker, M. D., Class handout, EPOB 5640, Multivariate Analysis for Ecologists. Colorado University, Boulder CO, Jan 1998.
- Wang, M. and C. Zheng, Optimal Remediation Policy Selection under General Conditions, *Groundwater*, 35(5): 757-764, 1997.
- Yu, C., C. Loureiro, J.J. Cheng, L.G. Jones, Y.Y. Wang, Y.P. Chia, and E. Faillace, *Data Collection Handbook to Support Modeling Impacts of Radioactive Material in Soil*, Internet Access: <http://web.ead.anl.gov/resrad/datacoll/dcall.htm>, Environmental Assessment and Information Sciences Division, Argonne National Laboratory, IL, 1993.

Appendix A Data Base

Site Number	Project Name	Reference	Depth to Water Table (ft)	Hydraulic Conductivity (ft/day)	Contaminant Concentration		Remedial Objective		Contaminant Type
					Source (mg/kg)	Dissolved (mg/L)	Source (mg/kg)	Plume (mg/L)	
1	Des Moines TCE, IA	1,2,3	17.5	535	3000	8.467	1	0.005	CVOC
2	Former Firestone Facility CA	1,3	45	100	1	0.12	1	0.006	CVOC, BTEX
3	Former Interil INC, CA	1,3	5	375	10000	0.61	1	0.005	CVOC
4	French Ltd, TX	1,3,8	11	1,54	400	920	43	0.1	CVOC, BTEX, SVOC
5	Gold Coast, FL	1,3	5	1000	7.86	2	1	0.005	CVOC, BTEX, metals
6	JMT Facility, NY	1,3	10	0.79	1	70	1	0.005	CVOC
7	Keefe Environmental Services	1,3	27.5	42.5	0.0228	4	0.0228	0.007	CVOC
8	SCRDI Dixiana	1,3	14	10	1	0.13	1	0.005	CVOC, Metals
9	Sol Lynn/Industrial Transformers	1,3,4	22.5	12.82	600	0.79	25	0.005	CVOC
10	US Avlex	1,3	20	27.25	1	200	1	0.2	CVOC
11	Baird and McGuire	1,3	12.5	24	27800	10	100	3.125	CVOC, SVOC, metals
12	King of Prussia Technical Corp	1,3,8,11	15	78	11300	1.04	483	0.05	CVOC, BTEX, metals
13	LaSalle Electrical	1,3	4	0.22	17000	3.123	10	0.2	CVOC, SVOC
14	Libby Groundwater	1,3,8	15	550	5000	3.2	37	1.05	SVOC
15	Mid-South Wood Products	1,3,4	30	0.85	11000	44	3	0.2	SVOC, metals
16	Solvent Recovery Services	1,3	26.5	150.1	61.2	30	1	0.005	CVOC, BTEX, metals
17	Sylvester/Gilson Road	1,3,8	15	40	16	140	16	2.9	CVOC, BTEX, metals
18	USCG Support Center	1,3,8	6	18.4	14500	3430	483	0.1	CVOC, Metals
19	Western Processing	1,3	7.5	50.5	84,269,629	250	1	0.07	CVOC, Metals
20	Odessa Chromium I	1,3	37.5	3.35	483	72	483	0.1	metals
21	Odessa Chromium II	1,3	37.5	3.35	483	9.9	483	0.1	metals
22	United Chrome	1,3,8	5	30.25	23030	19000	483	10	metals
23	Shaw AFB OU1	1	30	14.17	10	3	10	11.705	BTEX
24	US DOE Kansas City Plant	1,5	20	1.7	39.3	10	1	0.16	CVOC
25	Texas Tower Site, Ft Greely	1	36.5	28.37	9200	14.3	10	11.705	SVOC, BTEX
26	FT-01 Pope Air Force Base	1	3.5	14.17	44000	100	10	0.1	BTEX
27	Keesler AFB, AOC-A (ST-06)	1	7	40	640	22.4	100	18	BTEX
28	Langley AFB IRP site 4	1,8	5	0.0015	100	4.1	100	0.0074	BTEX
29	Eielson AFB, ST-20	1,8	7.055	1480	1500	12	200	0.005	BTEX
30	Hill AFB, UST Site 870	6	15	70.98	10000	26.576	10	11.705	BTEX
31	Elmendorf AFB, Site ST-41	6	18	28.47	565,7778	43.28	10	11.705	BTEX
32	Travis AFB, Site NSGS	6	10.5	70.84	614,4075	47	10	11.705	BTEX
33	Langley AFB, Site SS-04	6	5	0.0015	425	1.81	10	11.705	BTEX
34	Verona Well Field	1,8,12	15	7.085	550	92	0.06	0.003	CVOC, BTEX
35	Davis-Monthan AFB, Site ST-35	1	320	70.98	110	0.51	0.02	0.005	BTEX
36	Tinkham's Garage Superfund Site	1	5.5	5.5	16	0.446	1	0.005	CVOC, BTEX
37	Twin Cities Army Ammunition Plant	1,8	65	68.5	210	10	1	0.005	CVOC
38	Ammor Precast, Agden Utah	1,8	8	190	7.8	4.7	0.02	0.005	BTEX
39	North Fire Training Area, Luke AFB, Arizona	1,8	361	11.7	336	0	44	10	BTEX

Appendix A Data Base

Site Number	Project Name	Units	Depth to Water Table (ft)	Hydraulic Conductivity (ft/day)	Contaminant Source (mg/kg)	Contaminant Concentration Dissolved (mg/L)	Remedial Objective Source (mg/kg)	Plume (mg/L)	Contaminant Type
40	Hill Air Force Base, JP-4 Spill at site 280	1.8	105	1.41837696	5040	10	38.1	11.705	BTEX
41	Hill Air Force Base JP-4 Spill at Site 914	1.7.8	50	1.43113E-07	10200	11.705	38.1	11.705	BTEX
42	POL sites 2 and 5 Area, Holloman AFB	1	15	28.37	17500	11.705	100	11.705	BTEX
43	Bonneville Power Administration Wood Pole Ar	1	37.5	850.3	150	13	23	0.05	SVOC, metals
44	Savannah River A/M site	1	25	9	876.39	223	1	0.005	CVOC
45	Lawrence Livermore National Laboratory Site 3	1	12.5	141.837696	360	240	1	0.005	CVOC
46	Ft Richardson Building 908 South	1	50	42.64	17000	0.005	200	0.005	SVOC, BTEX
47	Kelley AFB, Building 2093 Gas Station	1	15	0.35	0.02	2.807	0.02	0.0294	BTEX
48	Sacramento Army Depot	1	80	2.834	11000	10	23	10	BTEX, CVOC
49	Basket Creek Surface Impoundment	1.8	27.5	28.34	8600	0.005	0.5	0.005	CVOC, BTEX
50	Cimarron Mining Corp. New Mexico	4.11	32.5	2.86	18000	4300	500	0.2	Metals
51	Crystal Chemical Company Texas	4	250	2.834	27000	600	30	0.05	Metals
52	Double Eagle Refinery Company Oklahoma	4	22.5	2.834	13300	0.05	500	0.05	Metals
53	Geneva Industries	4	30	2.834	12200	420	100	0.005	CVOC, SVOC
54	Koopers Co., INC. (Texarkana Plant) Texas	4.11	4	70.98	1400	4.1	100	0.001	SVOC
55	Midland Products	1.3.4	20	0.000283	14000	5100	1	0.2	SVOC
56	North Cavalcade Street Texas	4.8	250	28.34	14394	0.62	1	0.005	SVOC
57	Petro-Chemical systems, INC. Texas	4	21.5	2.86	7000	480	0.02	0.005	BTEX, SVOC
58	Sikes Disposal pits	1.4	140	28.47	58	2.2	10	0.02	CVOC, BTEX, SVOC
59	Brown Wood Preserving, Land treatment only	1	13.75	2.95	280	0.005	100	0.005	SVOC
60	Lowry Air Force Base Underground Storage Ta	1.8	4	22.73	14000	0.0031	500	11.705	BTEX
61	Holloman Air Force Base	1	10	28.34	17500	11.705	100	11.705	BTEX
62	Seymour Recycling Corporation	1.8	3.75	28.34	420	1	50	0.005	SVOC, CVOC, BTEX
63	Jasco Chemical Superfund Site	8	35	22.73	3400	142	0.2	0.005	CVOC, BTEX
64	Garden State Cleaners	8, 12	46	28.34	1300	1.9	1	0.001	CVOC, BTEX
65	Popple Inc Site	8	33	28.34	280	460	1	0.001	SVOC
66	328 Site, Santa Clara California	1	35	2.83	46	37	1	0.005	CVOC
67	Amoco Pipeline	1.8	6	85.0176	130.725	10	10	11.705	BTEX
68	Former Nu Look One Hour Cleaners	9	4	12.3	8.6143358	11.7	1	0.005	CVOC
69	Fourth Street Refinery	4	22.5	2.834	24500	0.05	500	0.05	Metals
70	Shaw AFB SD-29 and ST-30	1	30	14.17	592	1	10	11.705	BTEX
71	Offutt AFB LF-12	1	13.5	2.834	1	16	1	0.007	CVOC, BTEX
72	Solid State Circuits	1.10	90	0.8215	3.1	290	0.027	0.005	CVOC

1. FRTR, 2000

2. US EPA, 1989

3. US EPA, 1999

4. US EPA 6, 2001

5. US EPA, 2001

6. Parsons, 1999

7. Stanin et al, 1996

8. DuTeaux, 1997

9. US EPA, 2001

10. Hamil, 2001

11. US EPA, 1992

12. US EPA, 1994

Other works consulted

FRTR, 1997; US EPA, 1992a; US EPA, 1992b;

US EPA, 1993; USGS, 2001; Yuet al., 1993;

Domenico and Swartz, 1998;

Appendix B

Clustering Coefficients

Number of Clusters	Clustering Coefficient	Difference	Number of Clusters	Clustering Coefficient	Difference
72		0.021699			
71	0.021698541	0.031159	36	0.342117117	0.007581
70	0.052857375	0.004662	35	0.349698031	0.031725
69	0.057519784	0.001316	34	0.381422722	0.017528
68	0.058835913	0.006694	33	0.398950703	0.004175
67	0.06553	0.015979	32	0.403125953	0.017632
66	0.081508776	0.006381	31	0.42075746	0.001722
65	0.087889466	0.008071	30	0.422479345	0.020798
64	0.095959968	0.004201	29	0.443277792	0.021325
63	0.100160874	0.013327	28	0.464603136	0.002046
62	0.113487856	0.014855	27	0.466648979	0.001055
61	0.128342487	0.005581	26	0.4677703524	0.015628
60	0.133923126	0.004362	25	0.483331838	0.009639
59	0.13828536	0.001761	24	0.492970844	0.040122
58	0.140046594	0.016447	23	0.53309241	0.000583
57	0.156493847	0.014312	22	0.533675527	0.010577
56	0.170805971	0.023457	21	0.544252118	0.005608
55	0.194262568	0.000545	20	0.549859775	0.029181
54	0.194807136	0.013646	19	0.579040993	0.001153
53	0.208452744	0.01234	18	0.580194212	0.000408
52	0.220792965	0.000976	17	0.580602002	0.012127
51	0.221768957	0.009568	16	0.592729037	0.019074
50	0.231336587	0.007799	15	0.611802728	0.002663
49	0.239135907	0.005876	14	0.61446612	0.012107
48	0.245011771	0.005211	13	0.626573312	0.008399
47	0.250222728	0.002061	12	0.634972756	0.037284
46	0.252283347	0.004277	11	0.672257182	0.002012
45	0.256560735	0.010914	10	0.674269499	0.001735
44	0.267474431	0.003283	9	0.676004753	0.007118
43	0.270757929	0.009902	8	0.683122525	0.017297
42	0.28066006	0.004904	7	0.700419409	0.007853
41	0.285564319	0.012711	6	0.708272371	0.038378
40	0.298275383	0.012111	5	0.746650227	0.006118
39	0.310385948	0.011617	4	0.75276798	0.003994
38	0.322003253	0.006545	3	0.756762	0.018396
37	0.328548273	0.013569	2	0.775158	0.059017

Appendix C
Breakdown of 24 Clusters

Cluster Number	Projects within the Cluster
1	2,4,6,7,8,9,10,16,17,19,24,28,34,36,37,38,44,45,47,54,59,62,63,64,65,66,68,71
2	22,25,30,31,32,33,40,41,48,67,70
3	12,15,49,53,57
4	1,3,14
5	12,52
6	18,50
7	20,21
8	42,61
9	43,72
10	5
11	11
12	22
13	26
14	27
15	29
16	35
17	39
18	46
19	51
20	55
21	56
22	58
23	60
24	69

Vita

Lt Ben Recker was born in Ottawa, Ohio. He graduated from Ottawa-Glandorf High School in 1995. Later that year, he entered the United States Air Force Academy. He graduated in June 1999 with a Bachelor of Science degree in Environmental Engineering and was commissioned as a Second Lieutenant in the United States Air Force. His first assignment was attending graduate school at the Air Force Institute of Technology. Upon graduation from the Air Force Institute of Technology, he was assigned to the 90th Civil Engineering Squadron, F.E. Warren Air Force Base, Wyoming.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 074-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to an penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) 20-03-2001		2. REPORT TYPE Master's Thesis		3. DATES COVERED (From - To) Sep 1999 - Mar 2001	
4. TITLE AND SUBTITLE DETERMINING THE VALUE OF GROUNDWATER CONTAMINATION SOURCE REMOVAL: A METHODOLOGY				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Recker, Benjamin C., 2Lt, USAF				5d. PROJECT NUMBER If funded, enter ENR #	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 P Street, Building 640 WPAFB OH 45433-7765				8. PERFORMING ORGANIZATION REPORT NUMBER AFIT/GEE/ENV/01M-15	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Center for Environmental Excellence/Technology Transfer Division Attn: Maj Jeff S. Cornell 3207 North Road, Building 532 Brooks AFB TX 78235-5357 (210) 536-4329				10. SPONSOR/MONITOR'S ACRONYM(S) AFCEE/ERT	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT This study examines the tradeoff between extent of source removal and the lifecycle cost of a subsurface remediation project. It has been suggested that the lifecycle cost of a remediation project may be minimized at a certain percent source removal. This study attempts to validate this concept using real world data collected from 72 completed and on-going environmental remediation projects. Project data include total cost, extent of source removal, and site and contamination characteristics. Cluster analysis is used to group the diverse set of individual remediation projects under the assumption that projects within a cluster are similar enough to plot on a single lifecycle cost versus percent source removal plot. From the cluster analysis, two groups of 28 and 11 projects are used to develop lifecycle cost versus percent source removal curves. The resulting curves exhibit no apparent correlation between percent source removal and lifecycle cost. This study concludes with suggestions for future research that may shed light on the value of source removal towards reducing lifecycle cost of a subsurface contamination remediation project.					
15. SUBJECT TERMS Cluster Analysis, Groundwater Remediation, Lifecycle Cost, Source Removal, Remediation Project Data, Cost-Benefit Analysis					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPOR T	b. ABSTR ACT	c. THIS PAGE			19b. TELEPHONE NUMBER (Include area code)
U	U	U	UU	110	Professor Mark N. Goltz, ENV (937) 255-3636, ext 4638

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18

	Form Approved OMB No. 074-0188
--	-----------------------------------