

## **INFORMATION TO USERS**

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

**The quality of this reproduction is dependent upon the quality of the copy submitted.** Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

# **UMI**

A Bell & Howell Information Company  
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA  
313/761-4700 800/521-0600



NONPARAMETRIC STOCHASTIC MODELLING FOR NONSTATIONARY  
SUBSURFACE SOIL VARIABILITY AND INTERCONNECTION

by

Alaa El-Din Ibrahim Ali

A dissertation submitted in partial fulfillment  
of the requirements for the degree

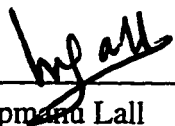
of

DOCTOR OF PHILOSOPHY

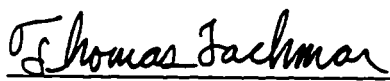
in

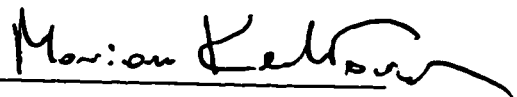
Civil and Environmental Engineering


Approved:


  
\_\_\_\_\_  
Dr. Upmanu Lall  
Major Professor

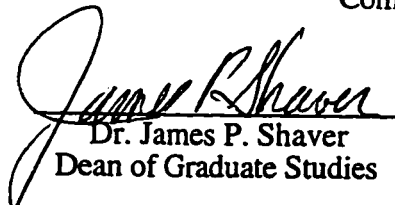
  
\_\_\_\_\_  
Dr. Kevin Hestir  
Committee Member

  
\_\_\_\_\_  
Dr. Thomas Lachmar  
Committee Member

  
\_\_\_\_\_  
Dr. Marian Kemblowski  
Committee Member

  
\_\_\_\_\_  
Dr. Loren Anderson  
Committee Member

  
\_\_\_\_\_  
Dr. Michael Minnotte  
Committee Member

  
\_\_\_\_\_  
Dr. James P. Shaver  
Dean of Graduate Studies

UTAH STATE UNIVERSITY  
Logan, Utah

1996

**UMI Number: 9717025**

---

**UMI Microform 9717025**  
**Copyright 1997, by UMI Company. All rights reserved.**

**This microform edition is protected against unauthorized  
copying under Title 17, United States Code.**

---

**UMI**  
**300 North Zeeb Road**  
**Ann Arbor, MI 48103**

## ABSTRACT

### Nonparametric Stochastic Modelling for Nonstationary Subsurface Soil Variability and Interconnection

by

Alaa El-Din Ibrahim Ali, Doctor of Philosophy

Utah State University, 1996

Major Professor: Dr. Upmanu Lall  
Department: Civil and Environmental Engineering

This study provides some strategies to characterize and simulate nonstationary heterogeneous porous media and to investigate the potential impact of such a heterogeneity on flow and transport. Nonparametric methods are used to develop some techniques for subsurface characterization and simulation using bore hole data. In these methods, the data are allowed to have a larger role in the estimation process than in a parametric model where a particular functional form is assumed a priori for the entire data set, and its parameters are estimated from the data. Such methods lead to two major accomplishments: 1) No data discretization was needed for subsurface characterization, and 2) no stationarity assumption was imposed on the underlying process. Applications to synthetic data from a nonstationary environment demonstrated the efficacy of these models in reproducing important statistics of such an environment. Applications to the Ogden Valley aquifer demonstrated the efficacy of these models in reproducing several attributes of the aquifer system.

Realizations generated by such models were used to identify potential preferential pathways and to estimate measures of the travel time. The simulated annealing technique was used as part of an optimization model to carry out such a task. Application to synthetic data showed the efficacy of such a model in identifying preferential pathways in an environment with dominating clayey layers. Application to the Ogden Valley aquifer indicated several preferential paths between a hypothetical line source and a location in the aquitard.

(195 pages)

To my father  
Ibrahim  
in heaven

## ACKNOWLEDGMENTS

"The more I learn the more I feel ignorant." This is my main conclusion after years of intensive research in the stochastic world. Over these years, I have focused on creating this large body of words, numbers, and images. However, the value of my work lies more in what I have learned rather than what I have produced. I am indebted to my major professor, Dr. Upmanu Lall, for the advice, patience, and encouragement he gave me throughout the course of my stay at Utah State University. I wish to thank Dr. Thomas Lachmar (Geology) for his thorough reading of my dissertation and serving as a committee member. I also wish to thank Dr. Kevin Hestir (Mathematics and Statistics) for serving as a committee member, and for very useful discussions during the course of study. I would like to extend my thanks to the other committee members, Dr. Loren Anderson (Geotechnical), Dr. Marian Kemblowski (Groundwater), and Dr. Michael Minnotte (Mathematics and Statistics). Nathan Rich receives special thanks for providing the main data set used for this study

I am grateful to my mother for her unconditional love, encouragement, and patience. My gratitude goes also to my brothers, Imad and Hisham; sister; brother in law; niece; and nephew for their prayers. I wish to extend my gratitude to my parents-in-law for their continuous support, and encouragement.

As my formal education draws to a close, I am thankful for all those have been my teachers, both formally and informally, especially my father, God's mercy upon him, who had been the best teacher of all.

I will remember fondly those who have become my friends in these four years: Balaji, Moon, Shaleen, Unni, Shymal, and Ashish. Your friendships have made this time not only bearable, but enjoyable and memorable. I would also like to express my gratitude to a



respected friend, Dr. Barakat, for useful discussion at early stages of this work.

Thanks to the staff at the Division of Water Rights: Jerry Olds, Michelle Lemiux, and Allyson Grandy, for their cooperation. The funding for this study through a graduate research assistantship from the Utah Water Research Laboratory, United States Geological Survey, grant no. JER-944, and the Division of Water Rights, State of Utah, is thankfully acknowledged.

The most "thank yous" go to my soul mate, and my wife, Malak, who has with me maintained a bond of love in the midst of both smooth and trying circumstances. I convey my warmest love and sincerest feelings to you. But for you, it would not have been possible to get this work done.

Finally, I dedicate this dissertation to my great father who passed away in November 1995. With all love and respect, I ask you in heaven to accept it for everything you have done for me. With all tears in my heart, I ask you to accept my apology for not being able to complete this work before you departed this world for a better one.

Alaa Ali

## CONTENTS

	Page
ABSTRACT . . . . .	ii
DEDICATION . . . . .	iv
ACKNOWLEDGMENTS . . . . .	v
LIST OF TABLES . . . . .	xi
LIST OF FIGURES . . . . .	xii
CHAPTER	
1. GENERAL INTRODUCTION . . . . .	1
Problem Statement . . . . .	1
Characterization of Aquifer Heterogeneity . . . . .	2
Impact of Aquifer Heterogeneity on Flow and Transport . . . . .	3
Objectives . . . . .	4
Site Used for This Study . . . . .	5
Outline . . . . .	6
References. . . . .	8
2. A KERNEL ESTIMATOR FOR STOCHASTIC SUBSURFACE CHARACTERIZATION. . . . .	10
Abstract . . . . .	10
Introduction . . . . .	10
Background . . . . .	12
Methodology . . . . .	14
A Probabilistic Interpretation of Binary Soil Types . . . . .	14
Model Formulation . . . . .	15
Computation of $\alpha_i$ and $\omega_i$ . . . . .	18
Parameter Specification . . . . .	20
Application . . . . .	23
Control Situation . . . . .	23
Results . . . . .	24
Ogden Valley Aquifer System . . . . .	25
Subsurface Characterization at Ogden Valley . . . . .	27
Testing for Prediction . . . . .	28
Statistical Parameters . . . . .	28

Conclusions . . . . .	30
References. . . . .	33
<b>3. A CONTINUOUS PARAMETER HOMOGENEOUS SEMI-MARKOV MODEL FOR STRATIGRAPHIC ANALYSES FROM BOREHOLE DATA . . . . .</b>	<b>49</b>
Abstract . . . . .	49
Introduction . . . . .	50
Background . . . . .	52
Discrete Parameter Homogeneous Markov Chains (DHMC) . . . . .	52
Continuous-Parameter Homogeneous Markov Chain and Semi-Markov Models . . . . .	55
A Continuous-Parameter, Homogeneous Semi-Markov (CHSM) Simulation Model for Stratigraphy . . . . .	58
Define Initial State Probability Matrix (IPM) and Generate Initial State . . . . .	58
Define Bed Thickness Cumulative Distribution Function for Each Rock Type . . . . .	59
Estimate Off Diagonal Transition Probability Matrix (OTPM) . . . . .	59
Applications . . . . .	60
Application 1 . . . . .	60
Application 2 . . . . .	62
Summary . . . . .	63
References . . . . .	65
<b>4. CONTINUOUS PARAMETER NONHOMOGENEOUS SEMI-MARKOV MODEL FOR STRATIGRAPHIC ANALYSES FROM WELL LOG DATA. . . . .</b>	<b>81</b>
Abstract . . . . .	81
Introduction . . . . .	82
Background . . . . .	83
Continuous Parameter Homogeneous Markov Chain and Semi-Markov Models . . . . .	84
Characterization . . . . .	85
Simulation . . . . .	86
Methodology . . . . .	87
Introduction . . . . .	87

	Characterization of a Continuous Nonstationary Stratigraphic Sequence . . . . .	88
	Kernel Estimator . . . . .	88
	Simulation of Pseudo-Well Logs . . . . .	91
	Application . . . . .	94
	Summary . . . . .	97
	References . . . . .	98
5.	A K-NEAREST NEIGHBOR SIMULATOR OF PSEUDO-BORE HOLE LOGS FOR SUBSURFACE CHARACTERIZATION . . . . .	108
	Abstract . . . . .	108
	Introduction . . . . .	109
	Problem Definition . . . . .	112
	Simulation Algorithm . . . . .	114
	Applications . . . . .	117
	Control Situation . . . . .	117
	Results . . . . .	118
	Ogden Valley Aquifer System . . . . .	118
	Results . . . . .	120
	Discussion and Conclusions . . . . .	121
	References . . . . .	123
6.	IDENTIFYING POTENTIAL PREFERENTIAL PATHS FOR SUBSURFACE TRANSPORT USING SIMULATED ANNEALING . . . . .	133
	Abstract . . . . .	133
	Introduction . . . . .	134
	Background . . . . .	135
	Problem Definition . . . . .	138
	Problem Solution Strategy . . . . .	140
	Introduction to Simulated Annealing Technique . . . . .	141
	Generation of Candidate Solution (Connected Path) . . . . .	143
	Cost Function . . . . .	144
	System Rearrangement . . . . .	145
	Annealing Schedule . . . . .	145
	Termination Criterion . . . . .	147
	Applications . . . . .	148
	Control Situation . . . . .	148
	Results . . . . .	148

Ogden Valley Aquifer System . . . . .	150
Preferential Pathway Identification . . . . .	152
Results. . . . .	152
Discussion and Conclusions . . . . .	153
References . . . . .	155
7. GENERAL SUMMARY . . . . .	173
APPENDICES . . . . .	177
APPENDIX A . . . . .	178
CURRICULUM VITAE . . . . .	179

## LIST OF TABLES

Table	Page
2-1 Indicator Values for Alternative Interpretations of the United Soil Classification System (Johnson and Dreiss, 1989) . . . . .	35
3-1 Markov Chain Properties for the Indian Site . . . . .	67
3-2 Statistics of the Bed Thickness for Three Types of Soil for: (a) the Real Data (Indian site); and Three Sets of 100 Simulated Images Generated Using: (b) CHSM, (c) DHMC with $\Delta z=4$ . Meters, and (d) DHMC with $\Delta z=2$ . Meters . . . . .	67
3-3 The Markov Properties for the Ogden Site . . . . .	68
3-4 Statistics of the Bed Thickness for Three Types of Soil for: (a) the Real Data (Ogden Site) and (b) 100 Simulated Images Based on the CHSM (All Units in Meters) . . . . .	68

## LIST OF FIGURES

Figure	Page
2-1 Illustration of probabilistic interpretation of sand/clay occurrence based on borehole logs . . . . .	36
2-2 Estimation of $\alpha_i$ (.) at elevation $z$ for borehole $i$ (Bandwidth = 15 meters) . . .	37
2-3 The general scheme for the estimation of $\lambda$ at a given depth . . . . .	38
2-4 Vertical cross-sectional view for the distribution of the probability of sand occurrence, drill log locations, and estimation sections, AA' and BB', for the control setting . . . . .	39
2-5 Elevation-northing of the probability of sand occurrence for synthetic data (contour image), and the true function (solid lines) at the site center (Figure 2-4, section A-A') . . . . .	40
2-6 Northing-easting of the probability of sand occurrence for synthetic data (contour image), and the true function (solid lines) at elevation = 30 m (Figure 2-4, section B-B') . . . . .	40
2-7 Variation of average horizontal bandwidth with elevation for the control situation . . . . .	41
2-8 Variation of average vertical bandwidth with distance from center for the control situation . . . . .	42
2-9 A plan view for the borehole sites and the location of the estimation region at Ogden Valley . . . . .	43
2-10 Three-dimensional image of the outside boundaries of the region of estimate . . . . .	44
2-11 Three-dimensional cut out of the aquifer system . . . . .	45
2-12 Elevation-northing contours for $\lambda$ at three different eastings . . . . .	46
2-13 Probability versus elevation at borehole 1, before and after dropping borehole log 1 . . . . .	47
2-14 The variation of the bandwidth in the horizontal with elevation for the Ogden site . . . . .	48

3-1	A stratigraphic section illustrating: (a) the drill log and different sampling intervals $\Delta z$ ; the count matrix, the TPM, and the TIM based on Krumbein (1968) if $\Delta z$ is (b) 5 ft. (i.e., equal to spacing between transition points); (c) 2 ft.; (d) 1 ft.; (e) the TIM based on continuous sampling . . . . .	69
3-2	Simulation procedures for a CHSM . . . . .	70
3-3	Three bore holes representing a subregion in a sedimentary basin in western India [ <i>Sinvhal and Sinvhal</i> , 1992] . . . . .	71
3-4	Bed thickness empirical CDF for the three rock types at the Indian site . . . . .	72
3-5	Simulated images for the Indian site generated by CHSM model . . . . .	73
3-6	Simulated images for the Indian site generated by DHMC using $\Delta z = 4$ meters . . . . .	74
3-7	Boxplots for the estimated unconditional probabilities associated with rock type 1 (sandstone), 2 (shale), 3 (coal) at the Indian site (based on 100 realizations) . . . . .	75
3-8	Boxplots for the estimated transition intensities from rock type $i$ to rock type $j$ , 1 (sandstone), 2 (shale), 3 (coal) at the Indian site (based on 100 realizations) . . . . .	76
3-9	Bore hole data representing the Ogden Valley aquifer . . . . .	77
3-10	Bed thickness empirical CDF for the three rock types at the Ogden site . . . . .	78
3-11	Simulated images for the stratigraphic sequence at Ogden Valley . . . . .	79
3-12	Boxplots for the estimated unconditional probability for sediment types 1 (clay), 2 (silt), 3 (sand) for the Ogden Valley site (based on 100 realizations) . . . . .	80
3-13	Boxplots for the estimated transition intensities from sediment type $i$ to sediment type $j$ , 1 (clay), 2 (silt), 3 (sand) for the Ogden Valley site (based on 100 realizations) . . . . .	80
4-1	Simulation procedures for a homogeneous semi-Markov model . . . . .	101
4-2	Kernel estimation of the transition intensity in the vertical . . . . .	102
4-3	Simulated well logs based on an application of the CNSM model to the data from Figure 3-9 . . . . .	103
4-4	Simulated well logs based on an application of the CNSM model to the data from Figure 3-9 . . . . .	103



4-5	Boxplots for the bed thickness statistics (in meters) of the three soils from 90 CNSM realizations based on: 1) direct resampling from data, 2) an exponential distribution . . . . .	104
4-6	Boxplots for the probabilities (p) and transition intensities (q) averaged over the entire profile for clay (1), silt (2), and sand (3) from 90 CNSM realizations based on: 1) Re-sampling from observed data and 2) an exponential distribution . . . . .	105
4-7	The UPM and TIM elements as a function of elevation for the real and pseudo-well logs. Bed thickness is resampled from the real data . . . . .	106
4-8	The UPM and TIM elements as a function of elevation for the real and pseudo-well logs. Bed thickness is sampled from an exponential distribution . . . . .	107
5-1	A layout of hypothetical drill log locations, the estimation grid, and a hypothetical drill log profile . . . . .	125
5-2	Simulation procedures for a realization using KNN method . . . . .	126
5-3	Elevation-northing of the probability of sand occurrence for synthetic data (contour image), and the true function (solid lines) at the site center (Figure 2-4, section A-A') . . . . .	127
5-4	Northing-easting of the probability of sand occurrence for synthetic data (contour image), and the true function (solid lines) at elevation = 30 m (Figure 2-4, section B-B') . . . . .	127
5-5	Plan view of the drill log data layout and locations of estimate . . . . .	128
5-6	Three-dimensional cut out of the aquifer system . . . . .	129
5-7	Three realizations of the aquifer system which honor the probabilistic image presented in Figure 5-6 . . . . .	130
5-8	The transition intensities, and unconditional probabilities as a function of elevation for the real data and the simulated images . . . . .	131
5-9	Bore holes along axis y-y' in Figure 5-5 . . . . .	132
6-1	Hypothetical flow paths from a contaminant source in an unconfined aquifer to a pumping well in an underlying confined aquifer . . . . .	157
6-2	Assignment of probabilities for the generation of a candidate path . . . . .	158
6-3	Input data for simulated annealing algorithm to identify a preferential pathways . . . . .	159

6-4	Procedures for constructing a connected random path between two points . . . . .	160
6-5	Local perturbation performed on the original path . . . . .	161
6-6	Procedures for simulated annealing in identifying preferential pathways and the associated travel times . . . . .	162
6-7	Procedures for estimating initial values of the control parameter and perturbation length . . . . .	163
6-8	Decrement coefficient distribution with the number of iterations . . . . .	164
6-9	Two images of a hypothetical setting of sand/clay in porous media . . . . .	165
6-10	Control parameter, C, versus travel time at the end of each C step (for the setting in Figure 6-9) . . . . .	166
6-11	Control parameter, C, versus average travel time within each C step (for the setting in Figure 6-9) . . . . .	166
6-12	Plan view of the bore hole data layout and locations of estimate . . . . .	167
6-13	A realization of the KNN method applied to bore hole data from the Ogden Valley aquifer . . . . .	168
6-14	Control parameter, C, versus travel time at the end of each C step (for the Ogden Valley aquifer application) . . . . .	169
6-15	Control parameter, C, versus average travel time within each C step (for the Ogden Valley aquifer application) . . . . .	169
6-16	Probability density functions of the travel time distribution between six points of origin along a line source and destination point D (see Figure 6-13) . . . . .	170
6-17	Preferential pathways between six points of origin along a line source and destination point D . . . . .	171
6-18	Ten realization for optimal pathways between points O and point D . . . . .	172

## CHAPTER 1

### GENERAL INTRODUCTION

#### Problem Statement

Public and governmental concerns regarding subsurface contamination problems have dramatically increased in the last two decades. Waste disposal sites that were thought adequate for groundwater resources protection appeared to be sources of leachate leakage to near-surface groundwater. In such environments, deep aquifers are often used for drinking water supply and, thus, the contamination of such aquifers within the well head protection areas (WHPA) is of major concern. The Safe Drinking Water Act (SDWA) defines a WHPA as "the surface and subsurface area surrounding a well or wellfield that supplies a public water system through which contaminants are likely to pass and eventually reach the water well or wellfield." The assessment of the gross contaminant potential in the deep aquifers is important for the delineation of such areas, and for environmental and governmental legislation.

In alluvial sedimentary environments, shallow and deep aquifer systems are typically separated by discontinuous lenses or layers of markedly different hydraulic conductivity, geometry, and size. Such heterogeneities are a consequence of geologic processes such as deposition, erosion, and sediment diagenesis, which are influenced locally by prior features (e.g., topography, stream location) and globally by historical climatic epochs. The alluvial aquifer system is consequently characterized by lenses and layers of media of rather disparate hydraulic conductivity, and of variable size and geometry. Also, we expect the lens occurrence process to be statistically nonstationary in the vertical and perhaps also in the horizontal. The effective degree of hydraulic interconnectivity of lenses of high

hydraulic conductivity between shallow and deeper aquifers represents a potential for preferential pathways that significantly influence groundwater flow and transport behavior. To identify such pathways, and to assess the gross contaminant potential in the deep aquifers, the reality of the variability and the geologic complexity of natural aquifer systems must be investigated.

### Characterization of Aquifer Heterogeneity

Sources of geologic information are diverse, ranging from hard information, such as drill logs, to soft ones such as seismic data, outcrop exposures, field observations, and trench studies. Useful data extracted from such information are usually sparse and/or not easy to interpret. Methods for aquifer characterization using the available data essentially belong to one of two major classes (deterministic and stochastic).

Geologic models have been the most widely used approach by sedimentologists for the last 40 years. These models provide descriptions for the vertical sequence and lateral extent of soil types and their boundaries based on the available data in a depositional environment. Geologic structures, such as faults and folds, and stratigraphic understanding are basic tools in these models. Reconciliations between such features and data available in a complex environment may be difficult and quite subjective. Also, such models may fail to represent any nonstationarity that may be imposed by the geologic processes and may be present in the data available.

Stochastic methods, in many forms, have been widely used to characterize the spatial structure in natural porous media. Some methods interpret the subsurface soil variability in terms of unconditional probability of observing a particular soil type at a given location in the subsurface [Journel, 1989; John and Dreiss, 1989]. Some other methods investigate the tendency of a certain soil type to follow another, in a stratigraphic sequence profile, in

terms of conditional probability of occurrence of different soil types (e.g., discrete parameter homogeneous Markov chain (HMC) [Harbaugh and Bonham-Carter, 1970; Bayer, 1985; Sinvhal and Sinvhal, 1992]. In these methods, the parameter of interest is viewed as a random function that belongs to a second order stationary stochastic process. Such stationarity ignores any features may be revealed by higher-order moments. Also, these methods impose artificial discretization on the data available to be consistent with soil blocks/layers being modeled. These methods use subjective measures of continuity (e.g., variogram in Kriging, transition probability matrix in HMC), which may not realistically reflect the underlying structure.

In this study, we use nonparametric techniques to interpret subsurface heterogeneity using drill log data. Kernel and nearest neighbor methods are used to characterize and/or simulate nonstationary environments. Models developed here do not impose any discretization on the data available. Some outputs of these models are used for investigation of flow and transport preferential pathways.

#### Impact of Aquifer Heterogeneity on Flow and Transport

The impact of soil heterogeneity on flow transport is traditionally evaluated in terms of equivalent homogeneous soil blocks of one single value for the hydraulic parameter (e.g., hydraulic conductivity). This approach does not account for local interconnectedness between zones of high conductivity, ignores the impact of the potential presence of preferential pathways, and hence, cannot be used to investigate sensitive behavior problems, such as a plume's early travel time. Existing measures for connectivity do not reflect the complex patterns of high conductivity lenses [Fogg, 1986; Journel and Albert, 1988]. Also, methods for investigating interconnectedness do not provide explicitly defined spatial connectivity patterns across a site [Silliman and Wright, 1988].

In this study, given a representation of an aquifer system, we use the simulated annealing technique to develop a model that identifies preferential pathways and an associated measure of the travel time between two locations.

### Objectives

The goal of this study is to model subsurface nonstationary environments using drill log data, and to investigate the impact of the heterogeneity on flow and transport behaviors in these environments.

To characterize such environments from drill log data, two aspects of the depositional process need to be investigated: 1) spatial distribution of lithologic units across drill logs and 2) understanding the stratigraphic sequence. Although a drill log is a real representation of the depositional process in the vertical at a particular location, it does not provide any information about the lateral variation of such a process. To interpret such a process across drill logs, a probabilistic interpretation of lithology observed at drill logs is needed. To simulate lithologic units across drill logs, a scheme that preserves the vertical, and horizontal, continuity of the depositional structure is required. To explore the persistence of a stratigraphic sequence at a site, we need to analyze the transitional property of the depositional process from one lithologic type to another. To carry out these goals, some nonparametric tools are developed.

The investigation of the heterogeneity impact on flow and transport aims at providing the practitioner with an easy tool for groundwater investigations, such as monitoring, and well head protection. To achieve this goal, the simulated annealing method is used to optimize a travel path between two locations in this environment.

Therefore, the focus of this work is twofold: 1) the application of nonparametric methods to subsurface characterization and simulation and 2) the application of the

simulated annealing method, using simulated images from step 1, to the identification of preferential pathways, and travel time estimation. Data used in this study are drill log data. Formally, the specific objectives in this study are:

- 1) Use the kernel method to develop a probabilistic tool that interprets drill log data in a three-dimensional grid.
- 2) Develop a continuous parameter semi-Markov model in a stationary environment.
- 3) Develop a continuous parameter semi-Markov model in a nonstationary environment.
- 4) Use the nearest neighbor method to sequentially simulate the subsurface environment by resampling drill log data.
- 5) Use the simulated annealing method to develop a model that identifies preferential pathways, and provides a measure for the travel time.

#### Site Used for This Study

The Ogden Valley has an aquifer system that is typical of Lake Bonneville sediments that cover large portions of the state of Utah. The site under consideration is located just west of the Wasatch Mountain Range on the relict Weber Delta. The delta consists of broad plains and terraces, and originates along the western base of the Wasatch Range. Topographically, this site is on a plateau formed by the Weber Delta. The plateau is approximately 90 meters above the valley floor. Surface elevations at this site vary from 1400 meters above mean sea level along the western side, to 1540 meters near the eastern side. Depths to bedrock in the basin range from 460 meters on the eastern side to 2300 meters on the west. The available data lie in the upper 20 meters of unconsolidated geologic material that consists of silts, clays, gravels, and sands. The geological units of interest at this site are the Pleistocene Provo Formation and the Pleistocene Alpine Formation.

Groundwater is found in the shallow alluvial deposits, in the sand and gravel deposits of the Provo Formation, and in the sand lenses within the underlying Alpine Formation, which is predominantly clay. The uppermost zones of groundwater are locally discontinuous and exist under unconfined conditions. The groundwater in the sand and silt layers of the upper portions of the underlying Alpine Formation usually exists under confined or semi-confined conditions.

The base of the unconfined aquifer (Provo Formation) rests unconformably on the Alpine Formation. The relatively low permeability clay deposits impede the downward migration of groundwater contaminants and enhance lateral migration along the upper surface of the clay in the downdip direction. Localized pockets of groundwater and dense nonaqueous phase liquid (DNAPL) have been identified in sand and silt lenses within the clay matrix to depths of 30 meters below land surface [Rich, 1995]. Even though the permeability of the Alpine Formation is relatively low, these sand seams may allow for enhanced fluid migration in their primary direction of orientation.

The concern at the study site was deep-migrating, free DNAPL that was detected in the aquifer system. Site characterization of the subsurface is important in identifying possible travel paths and zones of DNAPL.

## Outline

In this study, the objectives are addressed in a multiple-paper format. This work is divided into seven chapters, including the introduction and the conclusion chapters. These chapters are outlined as follows.

In Chapter 2, a kernel estimator is developed to interpret the subsurface variability from drill log data in a three-dimensional framework. In this model, soil variability in the vertical is viewed as a nonstationary process. Such a process is assumed to be



nonhomogeneous in the horizontal. The model parameters are allowed to vary in both the vertical and horizontal directions. The product of this model is a description for the probability of a particular type of soil occurring at any point within a three-dimensional grid. The model is applied to data from a nonstationary synthetic environment, and to data from the Ogden Valley aquifer system.

In Chapter 3, a treatment of sample discretization in stratigraphic modeling is presented. A continuous semi-Markov model is developed to generate stratigraphic sequences that honor some transition properties as seen in the data available. In this model, a transition intensity matrix is presented and used instead of the traditional transition probability matrix to describe the continuous nature of the depositional process. Monte Carlo simulation that honors the bed thickness and the transition intensity matrix is then performed. Applications to data from a sedimentary basin located in western India and data from Ogden Valley aquifer are presented.

In Chapter 4, a continuous nonhomogeneous semi-Markov model is developed to simulate a nonstationary stratigraphic sequence. A nonparametric method is used to describe the continuous variation of the transition intensity. Monte Carlo is used to generate stratigraphic sequences that honor the variation of the transition intensity and the bed thickness. The bed thickness is resampled using 1) exponential distribution and 2) a bootstrap (resample the bed thickness data with replacement). The model is applied to data from the Ogden Valley aquifer.

Chapter 5 presents the  $K^{\text{th}}$  nearest neighbor (KNN) method for subsurface simulation using drill log data. This simulator presumes each drill log to be the outcome of a spatial stochastic process modeled as a Markov random field. The random field is resampled using a KNN probability density estimator. Entire drill logs are resampled as pseudo-logs onto horizontal lattice coordinates. The vertical continuity of soil type is thus explicitly

preserved. The horizontal continuity/structure of the soils is preserved implicitly by conditioning the resampling of a drill log at each lattice location using the real as well as the pseudo drill logs that are the KNN of the current lattice location.

Chapter 6 investigates the impact of soil heterogeneity on the groundwater flow and transport. Modeling results from Chapter 5 are used for the aquifer representation needed for such investigations. Here, the simulated annealing technique is adopted to identify the preferential pathways, and to estimate a measure of the travel time, between two locations in the aquifer system. Applications to a synthetic setting and to the Ogden Valley aquifer are presented.

A brief discussion including recommendations of the new models developed is presented in Chapter VII.

## References

- Bayer, U., *Pattern Recognition Problems in Geology and Paleontology*, 229 pp., Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1985.
- Fogg, G. E., Groundwater flow and sand body interconnectedness in a thick, multiple-aquifer system. *Water Resour. Res.*, 22 (5), 679-694, 1986.
- Harbaugh, J. W., and G. Bonham-Carter, *Computer Simulation in Geology*, 575 pp., Wiley-Interscience, New York, 1970.
- John, N. M., and S.J. Dreiss, Hydrostratigraphic interpretation using indicator geostatistics, *Water Resour. Res.*, 25 (12), 2501-2510, 1989.
- Journel, A.G., *Fundamental of Geostatistics in Five Lessons*, 40 pp., Stanford Center for Reservoir Forecasting, Stanford University, Stanford, Calif., 1989.
- Journel, A. G., and F. G. Albert, Focusing on spatial connectivity of extreme-valued attributes: Stochastic indicator models of reservoir heterogeneities, *Society of Petroleum Engineer*, 18324, 621-632, 1988.
- Rich, N., DLOG3D application at operable unit 2 Hill Air Force Base, M.S. thesis, 93 pp., Utah State Univ., Logan, 1995.
- Silliman S. E., and A. L. Wright, Stochastic analysis of paths of high hydraulic conductivity in porous media, *Water Resour. Res.*, 24 (11), 1901-1910, 1988.

Sinvhal A., and H. Sinvhal, *Seismic Modelling and Pattern Recognition in Oil Exploration*, 178 pp., Kluwer Academic Publishers, Dordrecht, Boston, London, 1992.

## CHAPTER 2

### A KERNEL ESTIMATOR FOR STOCHASTIC SUBSURFACE CHARACTERIZATION<sup>1</sup>

#### Abstract

A nonparametric statistical methodology based on kernel function estimation is developed for assessing the probability that a particular location in the aquifer has high or low hydraulic conductivity using bore hole information. The approach presented is an alternative to indicator Kriging. Soils are classified through a binary indicator function defined as 0 for a low and as 1 for a high conductivity soil. Estimates of the probability of occurrence of a high or low conductivity soil are made on a three-dimensional grid. Each such estimate is formed as a local weighted average of the indicator function values that lie within an averaging interval or bandwidth of the point of estimate. A different vertical bandwidth is chosen at each borehole log. Horizontal bandwidths are selected independently at each horizontal level. These bandwidths are chosen by cross validation. Observations closer to the point of estimate are weighted higher using a kernel or weight function. Unlike Kriging, the underlying stochastic process is not assumed to be stationary. An application using data from Lake Bonneville deposits in Ogden Valley, Utah is presented.

#### Introduction

In alluvial sedimentary environments, shallow and deep aquifer systems are typically separated by discontinuous lenses or layers of markedly different hydraulic conductivity, geometry, and size. The preferential pathways for groundwater flow generated by such

---

<sup>1</sup>Coauthored by Alaa Ali and Upmanu Lall.

heterogeneities can control contaminant transport. Consequently, a quantitative characterization of the subsurface that highlights such macroscopic features can be useful for decisions on groundwater monitoring and for improved groundwater flow and transport model calibration. A particular concern for a water management agency is the identification of possible hydraulic connections between an upper, possibly contaminated, aquifer and a deep aquifer that is separated by an aquitard of variable thickness. Such sites may range in size from a few square miles to basin scale. Our particular interest here was in describing paleo-lake sediments associated with the geologic stands of Lake Bonneville.

The primary sources of information on subsurface hydraulic properties are pumping tests and bore hole logs. Existing pumping test methodologies are inappropriate for the identification of subsurface heterogeneities since they focus on the estimation of average hydraulic parameters, and the aquifer response to pumping is damped and smoothed over the discontinuities in hydraulic conductivity. Hence pumping tests can provide only limited information about the spatial variability of hydraulic conductivity. On the other hand, bore hole logs provide only local qualitative stratigraphic information. The traditional geological interpretation of bore hole logs is usually qualitative and does not formally account for the uncertainty associated with this localized information. *Johnson and Dreiss* [1989] pioneered an approach based on Indicator Kriging [*Journal*, 1989] for quantitatively interpreting bore hole logs. This method is not applicable to a statistically nonstationary environment. Sedimentary environments need not be stationary.

The objective of the work presented here is to develop a probabilistic interpretation of an aquifer's heterogeneity from bore hole information. We estimate the probability of occurrence of high hydraulic conductivity zones using kernel estimation methods. These estimates are formed as localized weighted averages of a binary indicator function (0 for low conductivity soils, and 1 for high conductivity soils) assigned to each bore hole log.

Unlike Kriging, stationarity of the indicator function is not required. The estimates from the kernel method may be used for simulation and decision making in the same way as the estimates from Indicator Kriging. Our goal in this manuscript is to provide a concise, expository presentation of the methodology for the practitioner. An application using data from Lake Bonneville deposits in Ogden Valley, Utah is presented.

### Background

This presentation is restricted to statistical analyses aimed at characterizing large-scale subsurface features using bore hole data. The primary approach currently used for this purpose is indicator Kriging. In this section, a brief review of the attributes of this method is offered together with some comments on its applicability for sedimentary environments.

A minimal description of the aquifer system from the perspective of groundwater flow and contaminant transport and flow is facilitated by considering two zones whose hydraulic conductivities may differ by one or more orders of magnitude. Bore hole data may then be classified as a binary sequence that corresponds to such a distinction. *Johnson and Dreiss* [1989] discuss two possible schemes based on interpretations of the Unified Soil Classification System (USCS) for the binary classification of bore hole data into high and low conductivity regions (see Table 1). Interpretation (a) highlights layers of low conductivity and emphasizes the separation between the high conductivity layers, while interpretation (b) highlights the spatial structure of zones of high conductivity. To emphasize different attributes of the data, one could systematically vary the threshold between the high and low conductivity material. Henceforth, high conductivity soil is called "sand" and low conductivity soil is called "clay."

Indicator Kriging [*Journel*, 1983, 1989; *Journel and Huijbergts*, 1978; *Isaaks and Srivastava*, 1989; *Cressie*, 1991] is a stochastic estimator that interpolates the binary

indicator data to estimate the probability of exceedance of the threshold originally used to classify the data into the binary scheme. Kriging entails two primary activities. The first is the inference of an appropriate variogram from the data, and the second is the use of this variogram together with the data to develop a best linear unbiased estimate (and its mean square error) of the parameter of interest (the probability of exceeding the threshold in this case) at a given location. The variogram recognizes the degree of spatial correlation between observations, and uses that information to determine how individual data points should be weighted to form a weighted average of the data at a point of estimate. Second order stationarity of the increments of the spatial random field represented by the indicator data is typically assumed. Reviews of the strengths and weaknesses of Kriging are offered by *Yakowitz and Szidarovsky* [1985], *Journel* [1989], *Isaaks and Srivastava* [1989], and *Owosina et al.* [1992]. In a situation where Kriging is used, it may be difficult to identify the correct variogram from the data, and to address the nonstationarity and anisotropy of the stochastic process. The variogram is critical for effective Kriging, and uncertainty as to its functional form and parameters can adversely impact the estimation scheme. Indicator Kriging applications with bore hole data require a discretization of the bore hole into sections of a specified length. The resulting estimates depend on the resolution of such a discretization. The associated information loss has not been formally studied to our knowledge.

A number of conceptual representations of sedimentary processes [*Merriam*, 1976; *Kendall et al.*, 1991; *Wanless*, 1991; and *Doveton*, 1991] have shown that the deposition process in alluvial sedimentary environments may often be nonstationary. Modifications of Kriging to consider trends in the target function, and also in the variogram through moving window estimates, are described by *Cressie* [1991]. However, data requirements for these procedures can be much higher than for ordinary Kriging, and the

optimal selection of such windows can be difficult. An approach that explicitly considers such nonstationarities may be a viable alternative to indicator Kriging. Such a method is presented here. No discretization of the bore hole log is needed.

### Methodology

Indicator Kriging and the kernel methods introduced here estimate the probability of the occurrence of "sand" or "clay" subsequent to the binary classification of soil types through an interpolation or averaging of the indicator data. The implications of such an estimation process are first described through a conceptual example. The kernel estimation framework employed here follows.

#### **A Probabilistic Interpretation of Binary Soil Types**

Two hypothetical bore holes, d1 and d2, several hundred meters apart, are shown in Figure 2-1a. Note that each bore hole log indicates two sand layers separated by a clay layer. The layer thicknesses vary between d1 and d2. For an averaging interval or neighborhood completely contained within a sand layer, the probability of sand occurrence,  $\lambda$ , is clearly 1. Similarly  $\lambda$  is 0 if the interval is strictly in a clay layer. Note that the probability of clay occurrence is  $(1-\lambda)$ . A symmetric averaging interval, centered at a clay/sand interface, leads to a probability of sand occurrence of 0.5 (half sand, half clay). This is consistent with the definition of probability as a relative frequency of occurrence, and recognizes that the exact location of the interface may be uncertain given that each bore hole samples only a small locale. The probability of sand occurrence  $\lambda$  varies with depth for each bore hole (from 1 to 0 and back to 1 again). Given the different layer thickness at the two bore holes,  $\lambda$  also varies between bore holes at a given depth. The construction of isolines of  $\lambda$  between bore holes implicitly assumes that  $\lambda$  varies smoothly between the



bore holes. A geologist's interpretation where interfaces between sand and clay may be directly connected in the description of a stratigraphic section may correspond to the isoline for  $\lambda=0.5$ .

The bore hole d2 in Figure 2-1a is now replaced by one with rapidly alternating layers of sand and clay with random thickness as shown in Figure 2-1b. Assuming that the thickness of such layers is much smaller than the averaging interval, the average probability of sand occurrence near d2 is 0.5 over the whole depth. Contours of  $\lambda$  between the two bore holes can still be constructed under the assumption that there is a smooth variation in the probability of observing sand between d1 and d2.

Variations in  $\lambda$  reflect changes in the proportion of sand and clay. The value of  $\lambda$  indicates the average proportion at the point of estimate. A transition across  $\lambda = 0.5$  can indicate the location of a large scale change in soil type or interface, as in Figure 2-1a. On the other hand, a  $\lambda=0.5$  can suggest an equal mixing of soil types (or equivalently, the possibility of an interface at every location) as in Figure 2-1b.

### **Model Formulation**

An estimate of  $\lambda$  can be obtained by considering some neighborhood of the point at which the estimate is desired and by determining the fraction of the volume of that neighborhood that is sand. This can be achieved by averaging over the values of the binary indicator function  $I(x,z)$  at observed locations in this neighborhood.

A special attribute of the bore hole data is continuous vertical sampling and limited horizontal sampling. The rate of variation or continuity of the sedimentation process is also likely to be quite different horizontally than in the vertical. Consequently, it is important to design an averaging strategy that properly treats such orientation and sampling differences.

A nonparametric approach based on kernel estimation is presented here. Kernel estimators are popular for probability density estimation and regression where prior

assumptions of the functional form of the underlying behavior are not desirable. The estimates are local approximations of the target function that use only information close to the point of estimate. The data are thus allowed to have a larger role in the estimation process than in a parametric model where a particular functional form is assumed a priori for the entire data set, and its parameters are estimated from the data.

Kernel estimators are weighted moving averages of a target function, where the weight is prescribed through a kernel function and the moving average is taken over an appropriately determined span or bandwidth. The kernel function is usually chosen to be a symmetric probability density function with finite variance, that has the role of a weight function. The reader is referred to *Härdle* [1989], *Silverman* [1986], and *Scott* [1992] for accessible monographs on kernel estimation and to *Lall* [1995] for a review of hydrologic applications.

The estimates of  $\lambda$  are formed on a three-dimensional spatial grid to aid subsequent contouring. At each estimation point  $(x, y, z)$ , a weighted average of the indicator function values within a prescribed neighborhood of the estimation node is formed as:

$$\bar{\lambda}(x,y,z) = \frac{\sum_{i=1}^n K_r(U_i(x,y,z))}{\sum_{j=1}^n K_r(U_j(x,y,z))} \left\{ \int_{z-h_{v,i}}^{z+h_{v,i}} \frac{1}{h_{v,i}} * I_i(\zeta) * K_v(q_i(z,\zeta)) d\zeta \right\} \quad (2-1)$$

where:

$n$  = Number of bore holes.

$I_i(\zeta)$  = Indicator function at bore hole  $i$  and elevation  $\zeta$ ; it is 1 if soil at elevation  $\zeta$  in bore hole  $i$  is of high conductivity and 0 otherwise.

$K_r(.) = \frac{15}{16}(1-U_i^2)^2$  if  $|U_i| \leq 1$ , and 0 if  $|U_i| \geq 1$ . This is a bisquare kernel function used in the horizontal.

$x, y, z$  = East, North, and Elevation (above datum) coordinates of point of estimate, respectively.

$x_i, y_i$  = East and North coordinates of drill site  $i$ , respectively.

$r_i(x, y) = \sqrt{(x-x_i)^2 + (y-y_i)^2}$  = the horizontal distance between point  $(x, y)$  and drill site  $i$ .

$h_{r,z}$  = A bandwidth or averaging interval in the horizontal at elevation  $z$ .

$$U_i(x, y, z) = \frac{r_i(x, y)}{h_{r,z}}$$

$K_v(.) = \frac{15}{16}(1-q_i^2)^2$  if  $|q_i| \leq 1$ , and 0 if  $|q_i| \geq 1$ . This is a bisquare kernel function used in the vertical.

$$q_i(z, \zeta) = \frac{z - \zeta}{h_{v,i}}$$

$h_{v,i}$  = A bandwidth or averaging interval in the vertical at drill site  $i$ .

The expression inside the summation in equation 2-1 can be written through two terms:  $\omega_i(.) = K_r(.) / \sum K_r(.)$  and  $\alpha_i(.)$  = term between brackets  $\{ \}$ . The term  $\alpha_i(.)$  represents the contribution of a vertical interval of width  $2h_{v,i}$  centered at elevation  $z$  at bore hole log  $i$  to the estimate of  $\lambda(x, y, z)$ . The term  $\omega_i(.)$  determines how such contributions from bore hole logs that lie within some radius  $h_{r,z}$  of the point of estimate  $(x, y, z)$  are averaged to estimate  $\lambda$ . The bandwidths  $h_{v,i}$  and  $h_{r,z}$  vary over the aquifer and are chosen using cross validation. Cross validation involves dropping an observation and estimating the function of interest (e.g.,  $\lambda$ ) at that point using the remaining data. The estimator's performance may be judged either by considering the sum of squares of differences between the observed (but deleted) value and its estimate using the remaining data, or by considering the likelihood of such an estimate using the remaining data.

The estimate  $\lambda$  is formed through a convolution of weight function  $K_v(.)$  and  $K_r(.)$  (with varying spans  $h_{v,i}$  and  $h_{r,z}$ ) with the indicator function  $I_i(\zeta)$ . The continuous sampling in the vertical is accounted for through the integral that appears in  $\alpha_i(.)$ , and the

discrete sampling in the horizontal is considered through the summation of the normalized weights  $\omega_i(\cdot)$ . Definition and estimation of  $\alpha(\cdot)$  and  $\omega(\cdot)$  are discussed below.

### Computation of $\alpha_i$ and $\omega_i$

At each bore hole  $i$ , the quantity  $\alpha_i(z)$  can be defined using the observations that lie within a vertical distance  $h_{v,i}$  from the horizontal level  $z$  as follows:

$$\alpha_i(z) = \frac{1}{h_{v,i}} * \int_{z-h_{v,i}}^{z+h_{v,i}} I_i(\zeta) * K_v\left(\frac{z-\zeta}{h_{v,i}}\right) d\zeta \quad (2-2)$$

Equation 2-2 computes the weighted fraction of sand at point  $z$  within an interval extending a distance  $h_{v,i}$  from point  $z$ . The estimate  $\alpha_i(z)$  is interpreted as the probability of sand occurrence at depth  $z$  at bore hole log  $i$  based only on the information at bore hole log  $i$ . The final estimate of  $\lambda$  at the same location will account for the information from the other bore hole logs.

We illustrate the estimation of  $\alpha_i(z)$  using a single bore hole log in Figure 2-2. The shaded areas under the kernel are the weights corresponding to the layers for which the indicator function equals 1. The unshaded areas under the kernel are the weights corresponding to layers where the indicator function is 0. The weight function serves to localize the estimate by giving higher weight to observations near the estimation point than for observations that are laterally further. Since continuous sampling is available vertically at each bore hole log site  $i$ , the estimate  $\alpha_i(\cdot)$  can be formed by integrating over the weighted indicator function over the desired interval as shown in Figure 2-2. The weight function,  $K_v(\cdot)$ , is defined such that upon integration, the resulting weights sum to 1.

The estimates  $\alpha_i(\cdot)$  for each bore hole log,  $i$ , at each elevation,  $z$ , are used to form the estimate  $\tilde{\lambda}(x,y,z)$ . A weighted average is formed through the normalized discrete weights

$\omega_i(\cdot)$ . Note that the functional form of the kernel used horizontally,  $K_r(\cdot)$ , is the same as that for the kernel used vertically. However, given the discrete horizontal sampling, the weights are normalized by dividing each  $K_r(\cdot)$  by the sum of the  $K_r(\cdot)$  values for all bore hole logs that lie within a radius  $h_{r,z}$  of the point of estimate.

An illustration of how the computation of  $\lambda$  proceeds using the weights  $\omega_i(\cdot)$  using the estimates  $\alpha_i(\cdot)$  is shown in Figure 2-3. In this figure, three bore hole logs fall within a radial bandwidth ( $h_{r,z}$ ) of the point of estimate. The bore hole used for the computation of  $\alpha_i(\cdot)$  in Figure 2-2 is shown as bore hole 1 in Figure 2-3.

Given that a symmetric kernel function is used, it is expedient to work in radial coordinates with the origin at the point of estimate in order to develop the weight sequence. Note that where there is a high degree of asymmetry in the geometry of the site (in plan view), or of the sampling locations or of the variation in  $\lambda(z)$ , it may be better to choose unequal bandwidths  $h_x$  and  $h_y$ . In the interest of parsimony, we shall stay with the single  $h_r$  at this stage. Alternately, one could work in rescaled coordinates  $x'$ ,  $y'$  that have been symmeterized.

A problem with forming a kernel estimate is encountered near the boundaries of the sample (i.e., bore hole log). Near the endpoints, the averaging intervals on one side extend across the boundary where there are no observations. Thus, we do not have a symmetric weighted moving average, leading to the effective center of the average not being at the point of estimate. Several solutions to this problem are offered in the literature [Müller, 1991; Silverman, 1986]. We adopted the use of reflection at the boundary. Reflection results in an unbiased estimate if the derivative of the target function (e.g.,  $\lambda(z)$ ) is small near the endpoints. Synthetic observations are generated across the boundary that are a mirror image of the data in the domain (mirror placed at the boundary). These synthetic observations are used only within a bandwidth of the boundary for an estimation point

within in the interior.

Given irregular geometry of the sampling locations, the boundary problem is more difficult to solve horizontally. We address this problem by restricting our estimation and cross validation to points that are strictly in the interior of the data set.

### **Parameter Specification**

The parameters to be chosen are the kernel function and the vertical and horizontal bandwidths. The selection of a kernel function is not as critical as the selection of its bandwidths. We shall first discuss the kernel choice, and then present the methods for the bandwidths selection.

#### Choice of Kernel

*Scott* [1992] shows that the shape of kernel does not significantly affect the mean square error (MSE) of estimate. Different kernels can be made equivalent in terms of the MSE by appropriately varying the bandwidth. It is desirable to base the choice of the kernel function on other considerations, e.g., the degree of differentiability required and the computational effort involved. The kernel function is usually symmetric to yield unbiased estimates by using a symmetric distribution of the weights on both sides of the point of estimate. It is also positive everywhere. *Scott* [1992] shows that the optimal positive kernel, which minimizes the MSE, is the Epanechnikov kernel,  $\{K(t) = \frac{3}{4}(1-t^2); -1 < t < 1\}$ , whose MSE efficiency is said to be 1. However, other kernels have nearly the same efficiency by adjusting their bandwidth such that a similar MSE is obtained. A smoother kernel (differentiable at  $t=1$ ) is preferred as it gives better differentiability and continuity of the resulting estimate. In this study, the bisquare kernel  $K(t) = \frac{15}{16}*(1-t^2)^2$ , which is better in this regard, was used. Its MSE efficiency is .9939.

### Bandwidth Selection

For a situation where  $\lambda''$  (the second derivative) is zero, i.e., no variation in  $\lambda$ , the best estimate will be obtained by taking the whole domain as the averaging neighborhood. On the other hand if  $\lambda''$  is high, i.e., a high variation in  $\lambda$ , smaller neighborhoods are desirable. However, if a very small neighborhood is chosen, variability of the estimate increases. Bandwidth selection thus represents a trade-off between bias and variance of estimate. Vertical bandwidths are estimated at each bore hole log, and horizontal bandwidths are estimated at each elevation  $z$  at which an estimate is needed.

*Vertical bandwidth selection at bore hole  $i$ .* The estimation of  $\alpha_i(\cdot)$  can be thought of as a local estimate of the relative frequency of observing sand from a one-dimensional vector of observations. This is analogous to probability density estimation or intensity estimation for a nonhomogeneous Poisson Process. From the example presented in Figure 2-2, it is clear that the textural characteristics may have significant vertical variation at different bore holes. This suggests that it may be advantageous to vary the bandwidth  $h_v$  for each bore hole by directly considering the sample attributes at the bore hole site. Here, we use a data-driven or automatic procedure at each bore hole to estimate an appropriate bandwidth  $h_{v,i}$ . The method used is a modified version of maximum likelihood cross validation (MLCV) [Silverman 1986]. In this method, a value  $\alpha_{i,-\delta}(\zeta)$  is first estimated using equation 2-2 for bore hole  $i$ , at the midpoint of each sand interval of thickness  $\delta$  (e.g., 1 foot) and midpoint elevation  $z$ , using all the layers except that interval. The log likelihood function ( $L_i$ ) at bore hole  $i$  is considered as the sum of piecewise log integrals of  $\alpha_{i,-\delta}(\zeta)$  over the layers of high hydraulic conductivity at bore hole log site  $i$ . This is a measure of the joint probability of occurrence of the observed values under the current choice of the bandwidth. Note that the joint probability or likelihood is given by the product of  $\alpha_{i,-\delta}(z)$ . Taking logarithms allows us to write the likelihood function as a sum, which is

computationally easier to deal with. The optimal bandwidth is one that maximizes the value of  $L_i$ . The cross validated log likelihood function is given as:

$$L_i = \sum_{j=1}^m \int_{z_{i,j}}^{z_{i,j+1}} \log(\alpha_{i,-\delta}(\zeta)) d\zeta \quad (2-3)$$

where  $m$  is the number of layers with  $I_i(\zeta)=1$ .

*Horizontal bandwidth selection at depth  $z$ .* In the horizontal, we are interpolating or regressing information across bore hole logs to points that have not been sampled. An appropriate method for selecting the bandwidth in this context [Härdle, 1989] is least square cross validation (LSCV). Here, we treat the estimates  $\alpha_i(z)$  as pseudo observations and cross validation proceeds by estimating  $\tilde{\lambda}_{-i}(x,y,z)$  at bore hole  $i$  without using the estimated  $\alpha_i(z)$  at that bore hole log. The probability  $\lambda_{-i}(z)$  is the probability of sand occurrence at bore hole  $i$  and elevation  $z$  using all data points except bore hole  $i$ . A cross validated sum of squares of errors is then formed between  $\alpha_i(z)$  and  $\tilde{\lambda}_{-i}(x,y,z)$  as shown in equation 2-4. The optimal  $h_r(z)$  is determined as the bandwidth that minimizes the LSCV function  $S_z$  given below:

$$S_z = \sum_{i=1}^{n_z} \left( \alpha_i(z) - \tilde{\lambda}_{-i}(x,y,z) \right)^2 \quad (2-4)$$

where:

$n_z$  = the number of bore hole logs with observation at level  $z$ .

At some locations, for level  $z$ , the bandwidth  $h_r(z)$  may capture too few, i.e., less than 5, bore holes to get a reliable estimate. For such a situation, a minimum number of bore hole logs to be used in the estimate must be defined.



### Application

The methodology presented is applied to two case studies. The first one is a control situation, where bore holes are sampled from an environment where a prior probability distribution of sand occurrence is assumed. The second uses bore hole data from a site in Ogden, Utah. All dimensions mentioned in this section are in meters (m.).

#### Control Situation

The control setting is a sedimentary environment with sand occurrence prescribed by the probability distribution:

$$\lambda(x,y,z) = \sqrt{\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2} \text{ if } \left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2 \leq 1 \quad (2-5)$$

$$= 1 \text{ else}$$

where:  $-250 \leq x \leq 250$ ,  $-250 \leq y \leq 250$ , and  $0 \leq z \leq 75$ .

A cross-sectional view of this environment through isolines of  $\lambda$  is shown in Figure 2-4. The values of a, b, and c are 250, 250, and 75 m, respectively. This environment was sampled by 80-m deep bore holes at 144 equally spaced locations. At each pseudo-bore hole, sand or clay units were randomly generated at a 1-m vertical discretization. The 1-m section is assigned sand if  $\lambda(x,y,z) > u$ , and clay if  $\lambda(x,y,z) \leq u$ . Here, u is a uniform random number between 0 and 1, and  $\lambda(x,y,z)$  is computed from equation 2-5, and (x,y,z) are the coordinates of the center of the 1 m. section.

Twenty-five such data sets with 144 drill logs each are simulated. In each simulation, the kernel method is applied to estimate  $\lambda(x,y,z)$  on a 10 m by 10 m by 1 m grid in the x, y, and z directions, respectively.

## Results

The estimated contours of  $\tilde{\lambda}$  from one realization using the kernel method are shown together with the true isolines of  $\lambda$  in Figures 2-5 and 2-6. The estimated  $\tilde{\lambda}$  values appear to be reasonable.

This hypothetical example also provides insight into the variation of the automatically chosen bandwidths relative to the underlying heterogeneity. For the 25 data sets, the average vertical bandwidth is estimated at each of 144 locations, and the average horizontal bandwidth is vertically estimated at 1-m intervals.

The variation of the average horizontal bandwidth by elevation is shown in Figure 2-7. The bandwidth is maximum at the bottom, where  $\lambda$  is nearly constant (it is equal to 1) and decreases upward as the heterogeneity in  $\lambda$  increases with  $(x,y)$ . This variation is expected. A slight increase in the bandwidth and then a decrease is observed between elevations 20 to 50 m. From Härdle [1989], it is known that the optimal bandwidth for kernel regression is proportional to  $(\sigma/\lambda'')^{2/5}$ , where  $\sigma^2$  is the variance of the residuals from regression, and  $\lambda''$  is the second derivative in the horizontal direction of the target function  $\lambda$ . In the context of the estimator presented here, the variance of the residuals corresponds to the variability in realizations generated from a particular value of  $\lambda$ . Since the samples were generated essentially using a binomial distribution with parameter  $\lambda(x,y,z)$  at each section, this variance is  $\lambda(1-\lambda)$ . This variance is maximum when  $\lambda=0.5$ , which occurs near the middle of the site. However,  $\lambda''$  is zero at the bottom of the site, and is a maximum at the top. Consequently, the general decrease in the bandwidth upwards, punctuated with a slight increase near the middle, is consistent with theoretical expectation.

The average vertical bandwidth as a function of distance from the center of the site is shown in Figure 2-8. A LOWESS smooth [Cleveland and Devlin, 1988] (solid line)

shows the vertical bandwidth increases and then decreases with distance from the center of the site. This behavior is consistent with the expectation of the optimal bandwidth. For density estimation, *Silverman* [1986] shows that the optimal bandwidth may be proportional to  $\sigma \cdot n^{-0.2}$ , where  $\sigma$  is the standard deviation of the data and  $n$  is the sample size. In our context,  $\sigma$  is the standard deviation of the locations in  $z$  at each bore hole where sand occurs, and  $n$  is the number of sand units at the bore hole. For the control situation used here, the standard deviation,  $s$ , increases with distance from the center and so does  $n$ . The resulting pattern of  $s \cdot n^{-0.2}$  was found to be very similar to the optimal bandwidth computed by MLCV.

### **Ogden Valley Aquifer System**

The Ogden Valley has an aquifer system that is typical of Lake Bonneville sediments that cover large portions of the state of Utah. The site under consideration is located just west of the Wasatch Mountain Range on the relict Weber Delta. The delta consists of broad plains and terraces, and originates on the western base of the Wasatch Range. Topographically, this site is on a plateau formed by the Weber Delta. The plateau is approximately 90 m above the valley floor. Surface elevations at this site vary from 1400 m above mean sea level along the western side, to 1540 m near the eastern side. Depths to bedrock in the basin ranges from 460 m on the western side to 2300 m on the east. The available data lie in the upper 20 meters of the unconsolidated geologic material which consist of silts, clays, gravels, and sands. The geologic units of interest at this site are the Pleistocene Provo Formation, and the Pleistocene Alpine Formation.

Groundwater is found in the shallow alluvial deposits, in the sand and gravel deposits of the Provo Formation, and in the sand lenses within the underlying Alpine Formation, which is predominantly clay. The uppermost zones of groundwater are locally discontinuous and exist under unconfined conditions. The groundwater in the sand and silt

layers of the upper portions of the underlying Alpine Formation usually exists under confined or semiconfined conditions.

The base of the unconfined aquifer (Provo Formation) rests unconformably on the Alpine Formation. The relatively low permeability clay deposits impede the downward migration of groundwater contaminants and enhance lateral migration along the upper surface of the clay in the downdip direction. Localized pockets of groundwater and DNAPL have been identified in sand and silt lenses within the clay matrix to depths of 30 m below land surface. Even though the permeability of the Alpine Formation is relatively low, these sand seams may allow for enhanced fluid migration in their primary direction of orientation.

The concern at the study site was deep-migrating DNAPL that was detected in the aquifer system. Site characterization of the subsurface is important in identifying possible travel paths and zones of DNAPL. We chose to assign 1 to soils of permeability higher than silt, and 0 otherwise. However, any other threshold may be selected.

The site is bounded by 570000 m, 570280 m east, and 89125 m, 89430 m north. Ninety-three irregularly spaced bore holes were available for this site. Bore hole depths ranged from 5 to 36 m. A layout for the bore hole locations is shown in Figure 2-9. The subdomain used for estimation is also shown in this figure. The estimates of the probability of sand occurrence are obtained at nodes of a three-dimensional grid bounded by 570125 m, to 570200 m east, 89150 m, to 89275 m north, and 1410 m, to 1430 m, in elevation.

### **Subsurface Characterization at Ogden Valley**

The characterized environment is presented in Figures 2-10, 2-11, and 2-12. These figures show, probabilistically, the structure of the environment. Figures 2-10 and 2-11 provide three-dimensional images of the estimated probability of sand occurrence. The

purpose of these figures is to help visualize the strata and its variations. Figure 2-10 provides a view for the outer boundaries of the estimation region. Figure 2-11 provides selected cutouts in the aquifer system. Figure 2-12 represents elevation-northing contour images for the probability of sand occurrence at three eastings: 570135, 570160, 570185 m. Figure 2-12 describes the likely behavior of soil strata at these locations. Distinct probability zones are observed and they indicate likely features in the geologic formation as discussed below.

We observe the Provo Formation (Figures 2-10 to 2-12) generally above elevation 1426 m ( $\tilde{\lambda} > .6$ ), and the Alpine Formation below 1418 m ( $\tilde{\lambda} < .4$ ), with a transition zone in between ( $.6 < \tilde{\lambda} < .4$ ). Recall that an interface between sand and clay is marked by  $\lambda = 0.5$ . In the Provo Formation, we notice an increase of probability of sand occurrence towards the North-East with sand dipping to the north at a 1:25 slope. This is identified as probability zone A in Figures 2-10, 2-11, and 2-12. Probability zone B in Figures 2-10, 2-11, and 2-12 indicates a likely transition zone located directly below zone A, and extending downward to elevation 1418 m. This zone represents a nearly equal chance of finding sand or clay.

In the Alpine Formation, the soil is expected to be mainly clay interrupted by sand lenses. Probability zones C, D and E in the maps of  $\tilde{\lambda}$  (Figures 2-10 to 2-12) show the possibility of such features. Zone C (corresponding to  $\tilde{\lambda} < .1$ ) represents the main clay matrix. Zone D (corresponding to  $\tilde{\lambda} > .5$ ) indicates a transition zone to a sand lens. It is located between elevations 1417 and 1415 m with width ranging between 29 and 69 m near east 570135 m. From Figure 2-11, this zone extends between 570125, 570144 m east, and between 89209, 89264 north. Zone E ( $\tilde{\lambda} > .4$ ) exhibits an increase in the  $\tilde{\lambda}$  value above the background value suggesting a likelihood of observing a sand lens. This zone is observed

at the same level near south and is shown by Figure 2-11 to extend east. Such locations represent potential zones of DNAPL saturation and may represent potential for preferential pathway between the uppermost formation and the deep confined aquifers.

### **Testing for Prediction**

The model of this study was validated in three steps: 1) estimate the probability of sand occurrence at an existing drill site using all data, 2) reestimate the probability at the same site after excluding the bore hole log at that site, and 3) compare the results obtained in 1, and 2, and the row data of that bore hole log.

We applied these procedures to a site as shown in Figure 2-9. Figure 2-13 shows the plots for the  $\tilde{\lambda}$  estimated at bore hole 1, before and after dropping bore hole log 1. Also, Figure 2-13 shows the bore hole log at sites 1 through 6. The bore holes are placed in the figure according to their radial distances from bore hole 1. A reasonable match is seen between the estimated probabilities before and after dropping the bore hole log at site 1. The sand/clay interface location at site 1 is seen to correspond to a probability of .5 as expected. Results near the bottom boundary may not be realistic due to 1) the boundary effect and/or 2) the contribution of some drill logs other than the one seen in Figure 2-13.

### **Statistical Parameters**

The estimated vertical and horizontal bandwidths are useful for understanding the observed fluctuations in  $\lambda$  and hence the soil type. As mentioned earlier, the bandwidth tends to shrink when the local variation in  $\lambda$  is high and tends to increase when such a variation is low.

In the vertical plane, the bandwidth is found to be between 2 to 6 m. Interestingly, this is similar to the average layer thickness in the area. The bandwidth was set equal to the square root of the bore hole log length (L) at some locations, where bore hole logs exhibit

an unusual lithologic sequence, (e.g., less than three layers, very short bore holes with only one layer, etc). The optimal bandwidth in the vertical is expected to be proportional to  $\sigma \cdot n^{-2}$ , as discussed earlier. The relationship found for this data set was  $h_v = .25 \cdot s \cdot n^{-2}$ , where  $h_v$  is the bandwidth at a bore hole log,  $n$  is the number of 1 m units of sand, and  $s$  is the standard deviation of the locations of the sand units.

In the horizontal, the bandwidth ranged from 29 to 139 m reflecting the lateral extent of the soil layers. Figure 2-14 shows a plot for the horizontal bandwidth in the horizontal versus elevation. The dotted lines correspond roughly to the different probability features presented above. Two interesting features are revealed from Figure 2-14:

1) Sudden drops in the value of the bandwidth in the horizontal reflect layer discontinuity in the vertical as shown in Figures 2-10, 2-11, and 2-12. The average spacing between such drops (3 m) falls in the range of the vertical bandwidth values (2-6 m), which is seen to be consistent with the average thickness of sand lenses observed in Figures 2-10, 2-11, and 2-12.

2) Bandwidth values are consistent with the corresponding soil structure. Bandwidths within zone A are relatively small. Note from Figures 2-10 to 2-12 zone A is laterally heterogeneous much like the top section of the synthetic example, where the horizontal bandwidth was also small.

This heterogeneity appears to extend down to elevation 1422 m (Figure 2-12). From Figures 2-10 to 2-12, we see that  $\tilde{\lambda}$  is approximately constant at .5 between elevation 1419 and 1422 m. The horizontal bandwidth is correspondingly at the upper bound (139 m) in this area. Below elevation 1417 m, the bandwidth structure is consistent with the observation of zone C being interrupted by zones D and E. Within the uninterrupted locations in zone C, (1417-1416, and 1413-1410), the bandwidth is at the upper bound. The bandwidth shrinks as the heterogeneity imparted by zones D and E is encountered. We

notice that the bandwidth value at the elevation of zones D and E is nearly equal to the width of the sand lens, zone D, at the same elevation (29 m at elevation 1417). Also, we observe an increase of the average width of sand lens and the bandwidth value with depth (50 m at elevation 1414 m). These observations suggest that the estimated bandwidth may be a useful measure of soil continuity at a certain depth.

### Conclusions

A new method for characterizing subsurface geology using kernel methods with bore hole data was presented here. A product of this work is a probabilistic image that can provide insight into subsurface heterogeneities. The methodology is particularly useful where the sedimentary environment is nonstationary. No discretization of the spatial domain beyond the resolution of the available data is needed. The nearly continuous information in the vertical at each bore hole is treated differently from and integrated with the sparse information in the horizontal. In these respects the methodology presented improves on Indicator Kriging, which is useful in the same context. Applications to a control situation and to a real data set demonstrated the efficacy of the kernel estimator in nonstationary situations, and for identifying possible heterogeneities in the subsurface from scattered bore hole data.

Kriging-based methodologies are popular and they are considered superior to other interpolation methods since they use information on the spatial correlation structure exhibited by the data. Such information is not used explicitly by the method developed here. Where the degree of spatial correlation is high, Kriging may yield results superior to the kernel method espoused here. However, in our experience [*Johnson and Dreiss, 1989*], (a) sparsity of sampling in the horizontal leads to a rather weak correlation structure (the nugget can be as high as 70%) particularly as the threshold used for generating the



indicator function moves towards the extremes of hydraulic conductivity at the site; (b) the nearly continuous sampling in the vertical at each bore hole contrasts with the sparse horizontal sampling and can dominate and bias variogram selection; (c) estimation of three-dimensional anisotropic variogram is difficult and its misspecification can negate the benefit of using the spatial correlation information; and (d) the cross validatory location adaptive bandwidth choice in the kernel method implicitly accounts for some of the local correlation structure in the data. Kriging has the advantage that if the nugget is zero, it is an exact interpolator, whereas the kernel method will always estimate a probability even at locations where the soil types are known. This need not be a negative attribute since we have already adopted a probabilistic perspective and recognize the bore hole to represent rather localized information. As the number of bore holes approaches infinity, theoretically, the estimated bandwidth will approach zero, asymptotically honoring the observed information.

Both Indicator Kriging and the kernel method require a relatively large number of bore holes to work effectively. *Hughes and Lettenmaier* [1982] indicate that at least 50 observations may be needed to reliably estimate a variogram in a two-dimensional setting. A similar number of observations are needed for the kernel method to be effective. Moving window Kriging [*Cressie*, 1991] is sometimes advocated to deal with nonstationarities. The sample size requirements are even higher in this case.

Incorporation of qualitative geologic information is not considered in our formulation of the kernel estimator. Co-Kriging [*Xu et al.*, 1992] can accommodate such additional information. A similar extension of kernel estimators is theoretically possible, but has not yet been pursued.

A limitation of the kernel methods presented is the increased bias of estimate near the boundaries of the data set. This bias restricts the usability of the method to the interior of the data set. Some corrections for this bias are possible at the expense of increased variance

of estimate. Kriging suffers from similar problems. However, this aspect of Kriging has not usually been pointed out.

A fixed bandwidth was selected in the vertical at each bore hole log and a fixed bandwidth was selected in the horizontal at each elevation of estimate. If adequate data were available, it would be desirable to choose local bandwidths at each point of estimate in the horizontal and in the vertical. This would recognize the local characteristic scale of variation of the underlying process. Strategies for such local choices can be readily developed following proposals in the statistics literature [*Mielniczuk et al.*, 1989]. It is unclear whether the extra computational effort and possibly increased estimation variance will be justified.

A limitation of the model presented is the restriction to two types of soil. As in Indicator Kriging, one can implicitly estimate the cumulative distribution function associated with different hydraulic conductivity soils by varying the threshold at which the soil types are partitioned. If a direct consideration of multiple soil types is desired, an alternate formulation of the kernel method will be necessary. In this case an a priori assignment of a representative hydraulic conductivity (or its logarithm) of each soil type may be used to classify the bore hole data instead of a binary classification. The integer part of the logarithm of the hydraulic conductivity may be a useful classifier since it separates the data into classes separated by an order of magnitude.

A simulation technique is under development to condition the current estimator on the hard data to obtain the random field in terms of simulated pixels. Methodology to identify preferential pathways for contaminant transport by mapping path with a high joint probability of sand is also under development.

Practically, we feel that the approach presented here belongs in the geostatistician's toolbox as much as Kriging does. The conceptual decomposition used here, of the spatial

domain into the vertical, where sampling is continuous and the horizontal where sampling is irregular, is an interesting and powerful building block. Extensions to consider Markovian dependence in soil type and to develop estimates of statistics (e.g., covariance) of the random field, as well as of measures of connectivity, are at an advanced stage of testing and development. FORTRAN source for the algorithms presented is available by request from the authors.

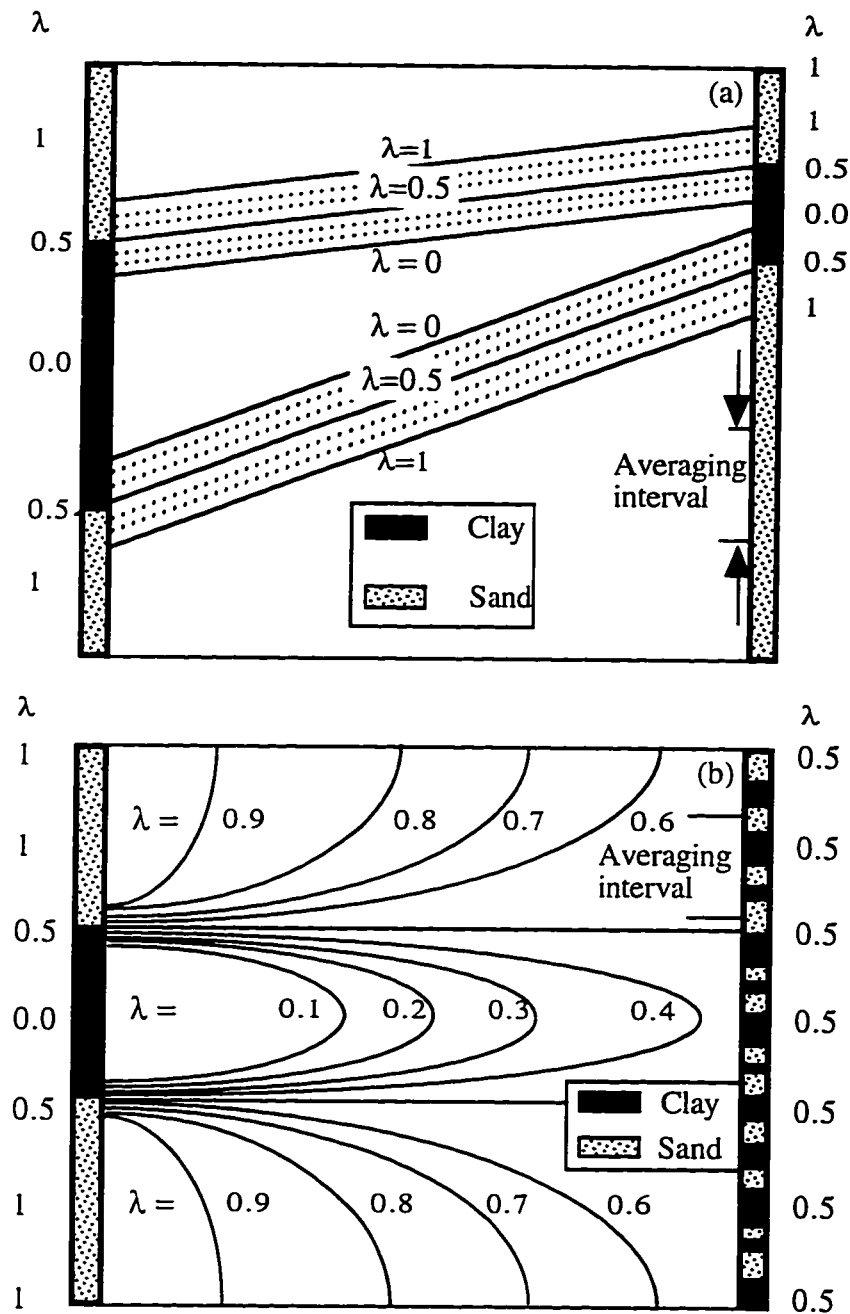
### References

- Cleveland, W. S., and E. J. Devlin, Locally weighted regression: An approach to regression analysis by local fitting, *J. Amer. Stat. Assoc.*, 83 (403), 596-610, 1988.
- Cressie, N, *Statistics For Spatial Data*, 900 pp., John Wiley & Sons Inc., New York, 1991.
- Doveton J. H., Lithofacies and geochemical facies profiles from nuclear wire-line logs, New subsurface templates for sedimentary modelling, in *Sedimentary Modelling: Computer Simulations and Methods for Improved Parameter Definition*, edited by E. Franseen, W. Watney, C. Kendall, and W. Ross, Kansas Geological Survey Bulletin 233, Lawrence, Kan., pp. 101-121, 1991.
- Härdle, W., *Applied Nonparametric Regression*, 333 pp., Cambridge University Press., Cambridge, Mass., 1989.
- Hughes J. P., and D. P. Lettenmaier., Data requirements for Kriging: Estimation and network design, *Water Resour. Res.*, 17 (6), 1641-1650, 1982.
- Isaaks, E., and R. Srivastava, *An Introduction to Applied Geostatistics*, 561 pp., Oxford University Press, New York, 1989.
- Johnson, N. M., and S. J. Dreiss, Hydrostratigraphic interpretation using indicator geostatistics, *Water Resources Research*, 25 (12), 2501-2510, 1989.
- Journal, A. G., Non-parametric estimation of spatial distributions, *Math Geol.*, 15 (3), 445-468, 1983.
- Journal, A.G., *Fundamental of Geostatistics in Five Lessons*, 40 pp., Stanford Center for Reservoir Forecasting, Stanford University, Stanford, Calif., 1989.
- Journal, A. G., and Ch. J. Huijbergts, *Mining Geostatistics*, 600 pp., Academic Press, London, New York, San Francisco, 1978.
- Kendall, C. G. St. C., M. Philip, S. John, R. Cannon, J. Perlmutter, J. Bezdek, and G. Biswas. Simulation of the sedimentary fill of basins, in *Sedimentary Modelling:*

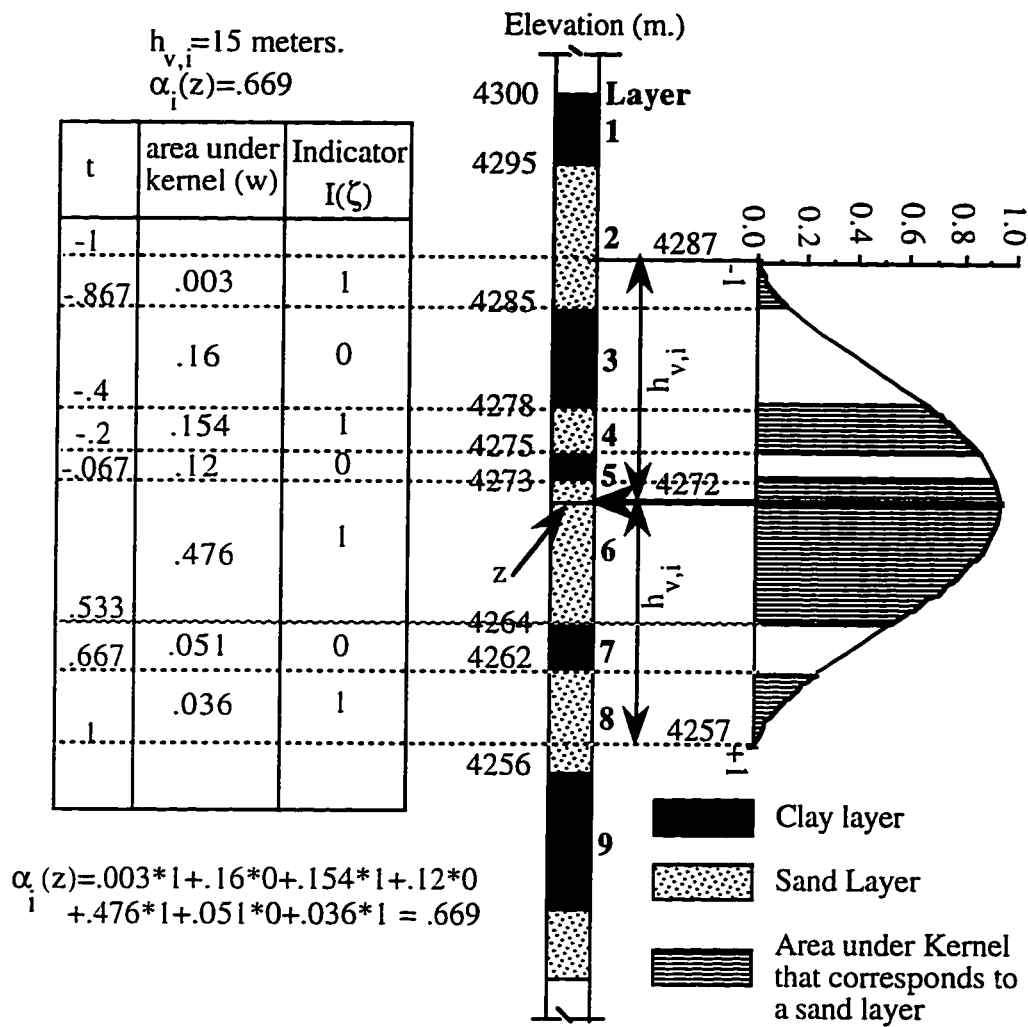
- Computer Simulations and Methods for Improved Parameter Definition*, edited by E. Franseen, W. Watney, C. Kendall, and W. Ross, Kansas Geological Survey Bulletin 233, Lawrence, Kan., pp. 8-30, 1991.
- Lall, U., Nonparametric function estimation: Recent hydrologic applications, *Reviews of Geophysics, US National Report 1991-1994*, 1093-1102, 1995.
- Merriam, D. F., *Random Process in Geology*, 168 pp., Springer-Verlag, New York, 1976.
- Mielniczuk, J. P., P. Sarda, and P. Vieu, Local data-driven bandwidth choice for density estimation, *J. Stat. Plan. Infer.*, 23, 53-69, 1989.
- Müller, H. G., Smooth optimum kernel estimators near endpoints, *Biometrika*, 78 (3), 521-530, 1991.
- Owosina, A., U. Lall, T. Sangoyomi, and K. Bosworth, Methods for assessing the space and time variability of groundwater data, *Report No. 14-08-0001-G1738*, 246 pp., Utah Water Research Laboratory, Utah State University, Logan, Utah, 1992.
- Scott, D. W., *Multivariate Density Estimation, Theory, Practice, and Visualization*, 317 pp., John Wiley and Sons, New York, 1992.
- Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, 175 pp., Chapman and Hall, New York, 1986.
- Wanless H. R., Observational foundation for sequence modelling, in *Sedimentary Modelling: Computer Simulations and Methods for Improved Parameter Definition*, edited by E. Franseen, W. Watney, C. Kendall, and W. Ross Kansas Geological Survey, Lawrence, KS 66047, Bulletin 233, pp. 43-62, 1991.
- Yakowitz, S. J., and F. Szidarovsky, A comparison of Kriging with nonparametric regression method, *J. Multivar. Anal.*, 16(1), 21-53, 1985.
- Xu, W., T. T. Tran, R. M. Sirvastava, and A. G. Journel, Integrating seismic data in reservoir modelling: The collocated cokriging alternative, *Soc. Pet. Eng.*, 24742, 833-842, 1992.

**Table 2-1.** Indicator Values for Alternative Interpretations of the United Soil Classification System [*Johnson and Dreiss, 1989*]

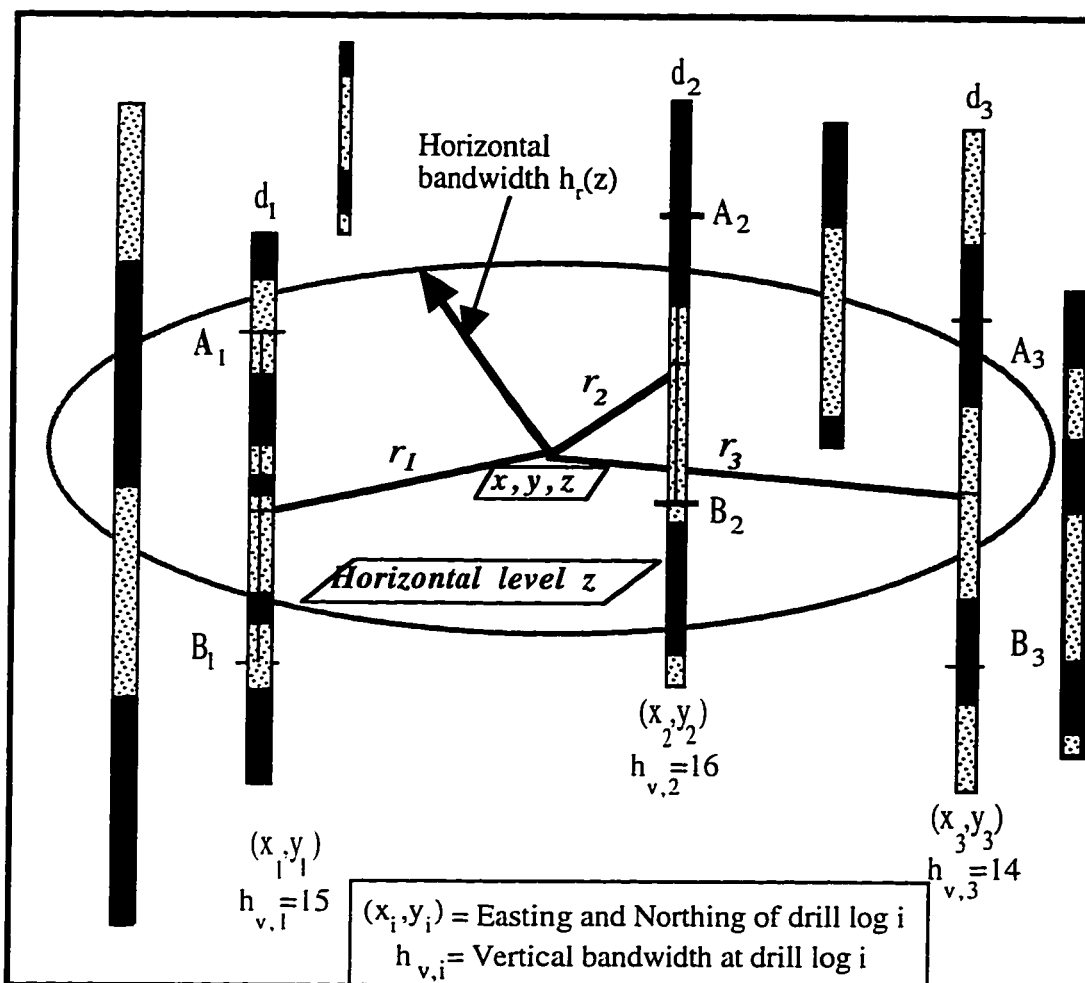
High permeability Indicator=1		Low permeability Indicator=0
<hr/>		
	Interpretation (a)	
GW, SW, GM, SM		ML, CL, OL
GP, SP, GC, SC		MH, CH, OH
	Interpretation (b)	
GW, SW		GM, SM, ML, CL, OL
GP, SP		GC, SC, MH, CH, OH
<hr/>		
G, gravel; S, sand; M, silt; C, clay; O, organics; W, well graded, (i.e., poorly sorted); P, poorly graded; L/H, Low/High plasticity.		



**Figure 2-1.** Illustration of probabilistic interpretation of sand/clay occurrence based on borehole logs. The contour lines,  $\lambda$ , are the probabilities of sand occurrence for a probabilistic solution. (a) A homogeneous depositional environment that can be solved deterministically. (b) Highly spatially variable depositional environment which may not be solved deterministically.



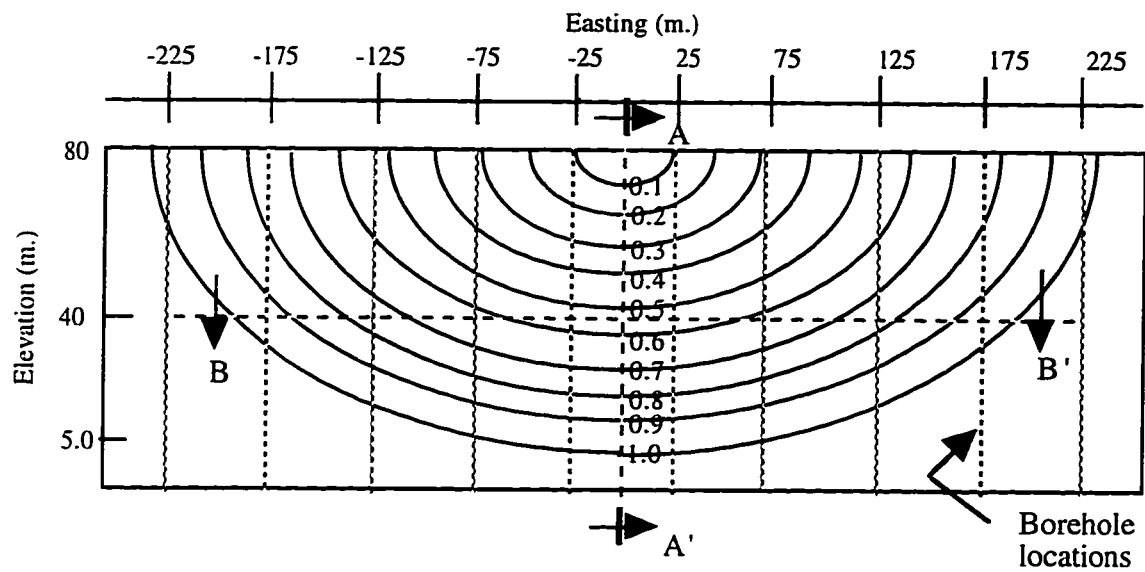
**Figure 2-2.** Estimation of  $\alpha_i(\cdot)$  at elevation  $z$  for borehole  $i$  (Bandwidth = 15 meters).



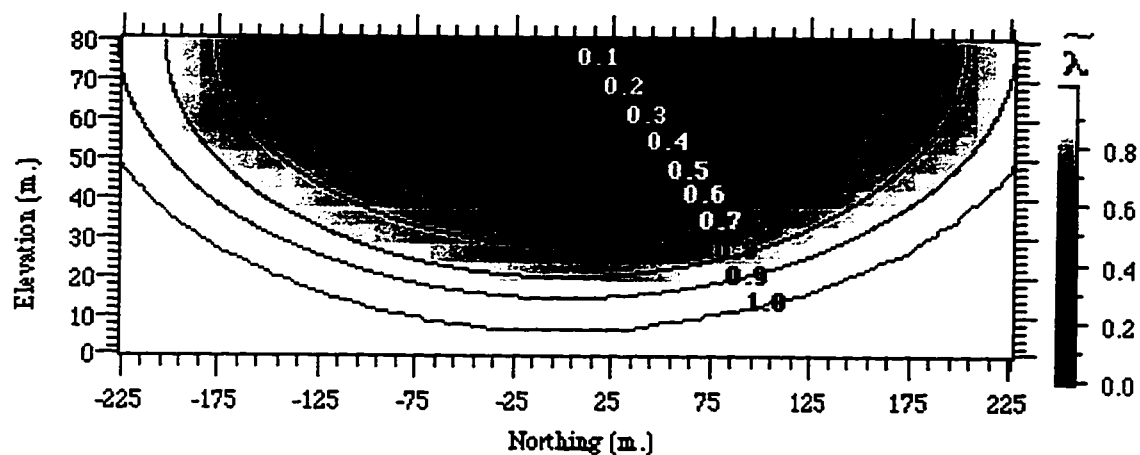
drill log #	$\alpha$	Distance $r$ to point $(x, y, z)$	$K_r(u(z))$	$\omega(x, y)$	$\omega^* \alpha$	$\tilde{\lambda}(x, y, z_j) = \sum \omega^* \alpha$ $= .09 + 0.477 + .099$ $= .666$
1	0.69	1200	0.122	0.131	0.09	
2	0.85	750	0.527	0.561	0.477	
3	0.32	1000	0.289	0.308	0.099	

**Figure 2-3.** The general scheme for the estimation of  $\lambda$  at a given depth. Note that only the points that lie between  $[A_i, B_i]$  at each borehole  $i$  that lies within  $h_r$  from the point of estimate  $(x, y, z)$  contribute to the estimate. In this illustration, three boreholes contribute to the estimate.

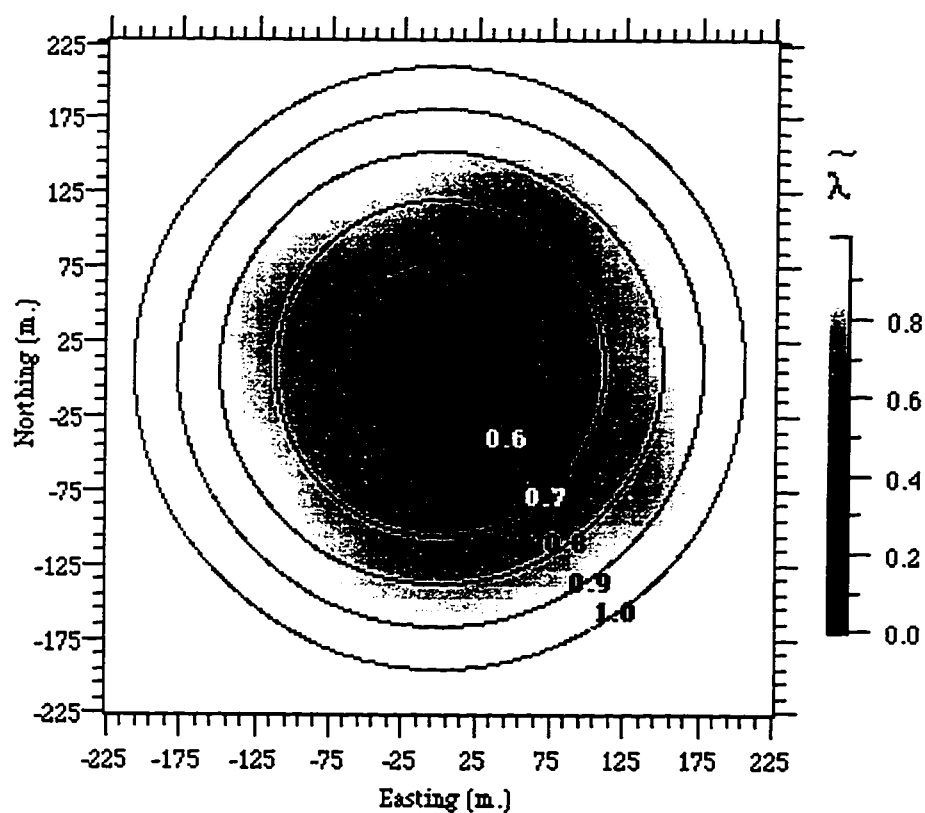




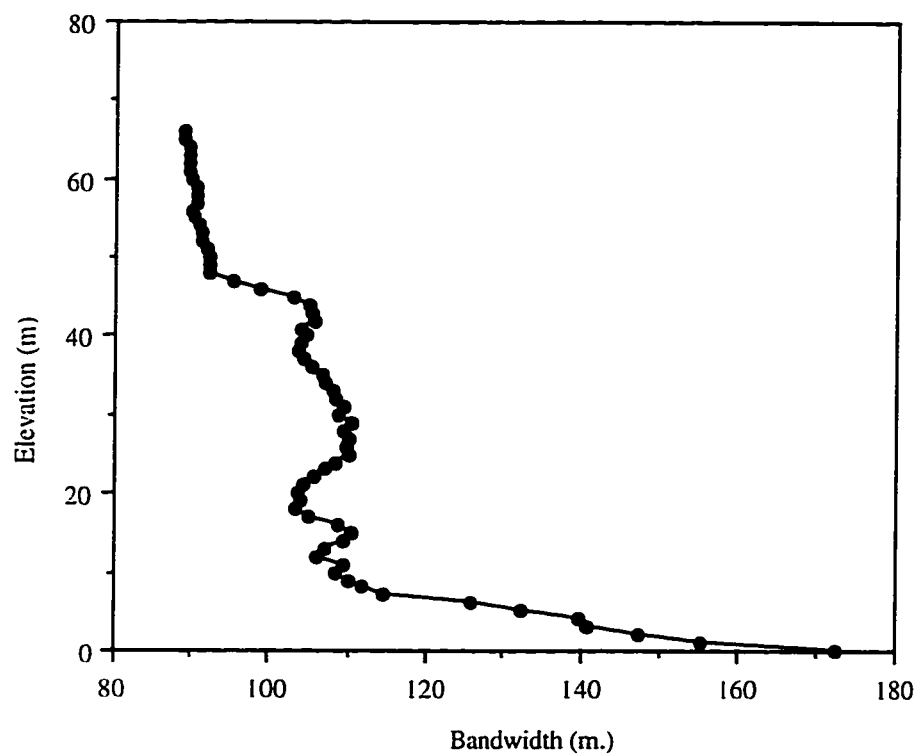
**Figure 2-4.** Vertical cross-sectional view for the distribution of the probability of sand occurrence, drill log locations, and estimation sections, AA' and BB', for the control setting.



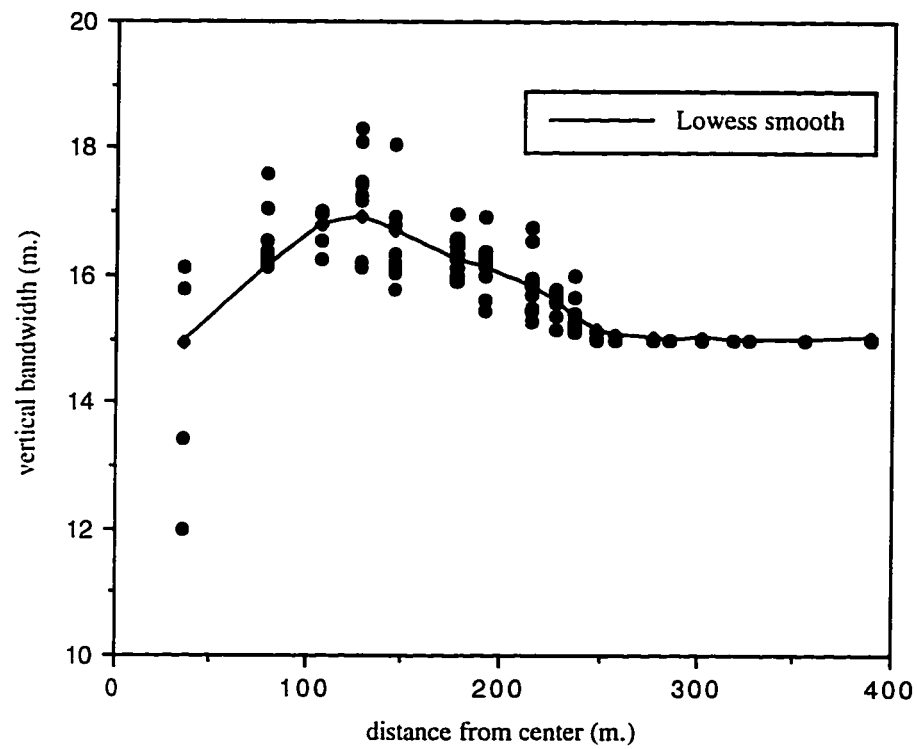
**Figure 2-5.** Elevation-northing of the probability of sand occurrence for synthetic data (contour image), and the true function (solid lines) at the site center (Figure 2-4, section A-A').



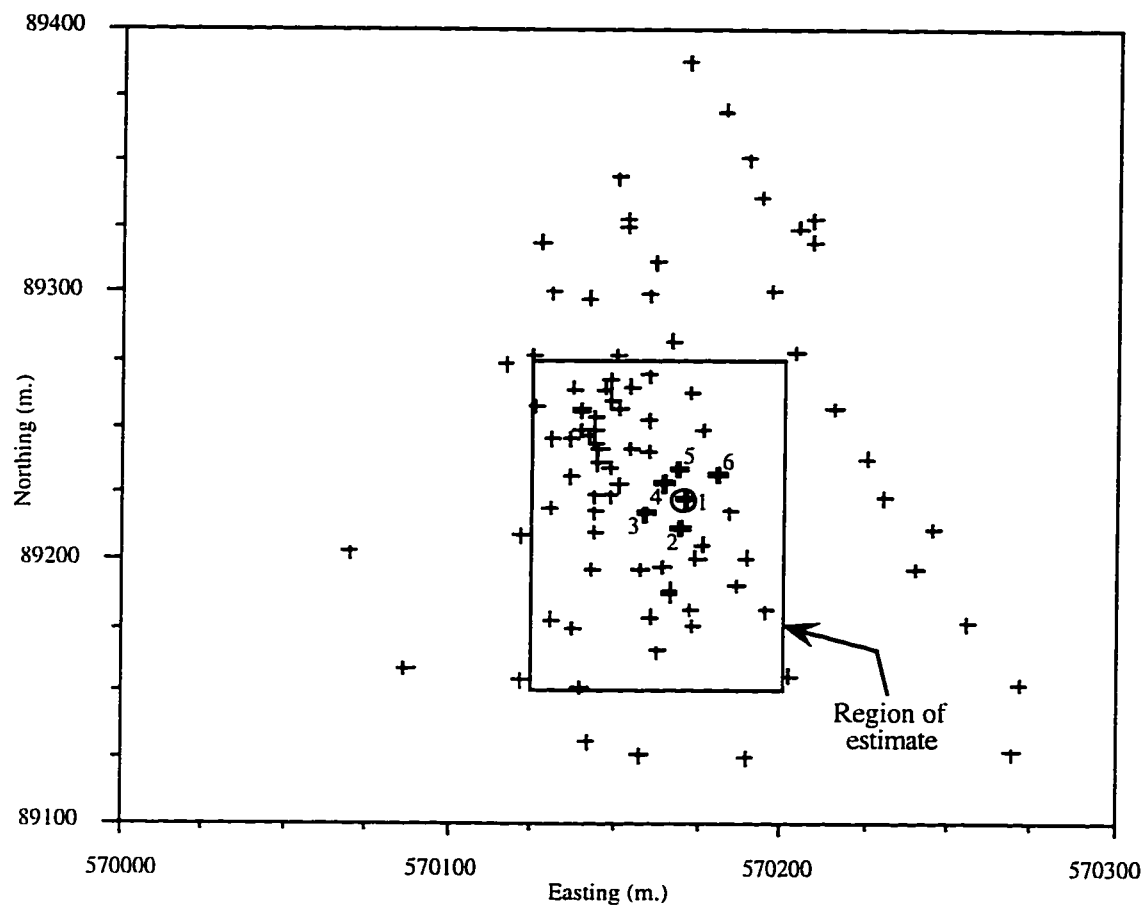
**Figure 2-6.** Northing-easting of the probability of sand occurrence for synthetic data (contour image), and the true function (solid lines) at elevation = 30 m (Figure 2-4, section B-B').



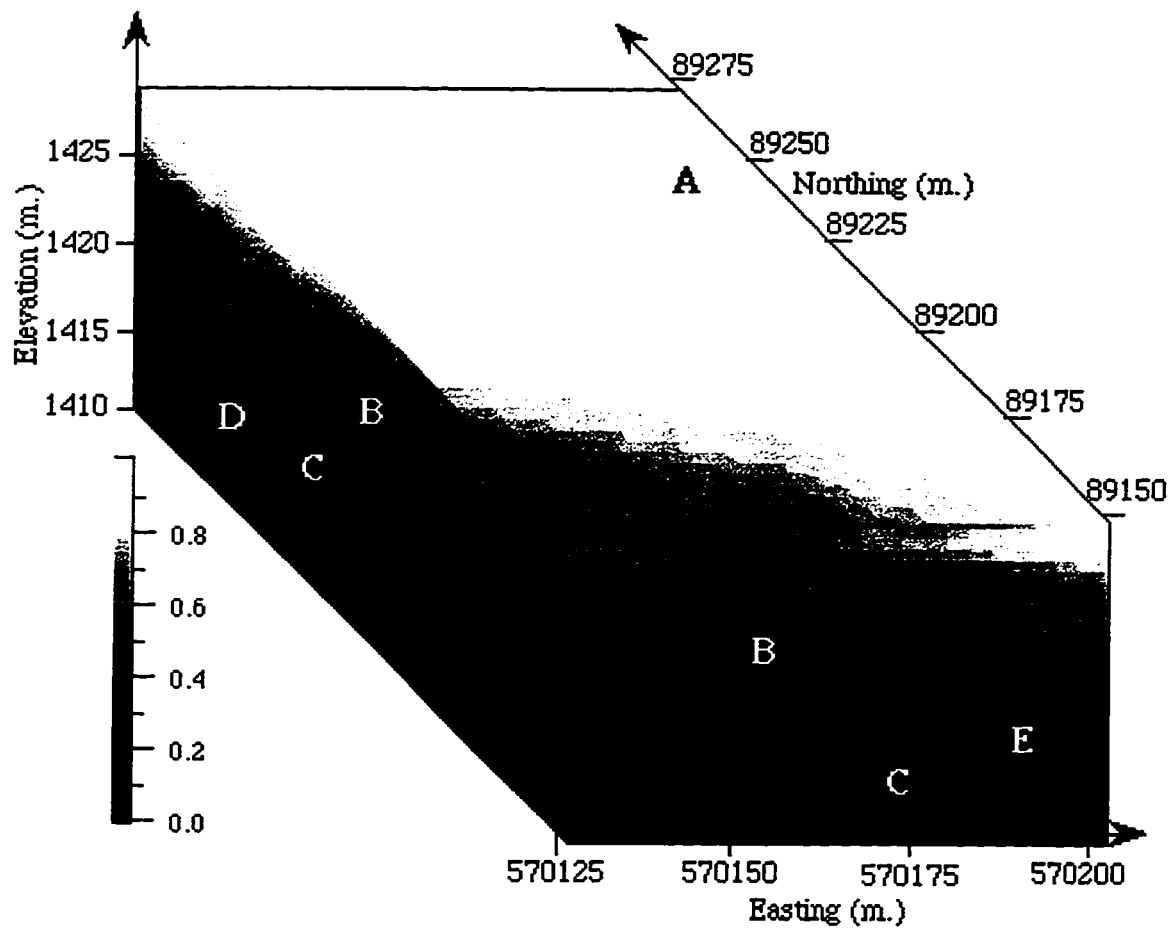
**Figure 2-7.** Variation of average horizontal bandwidth with elevation for the control situation.



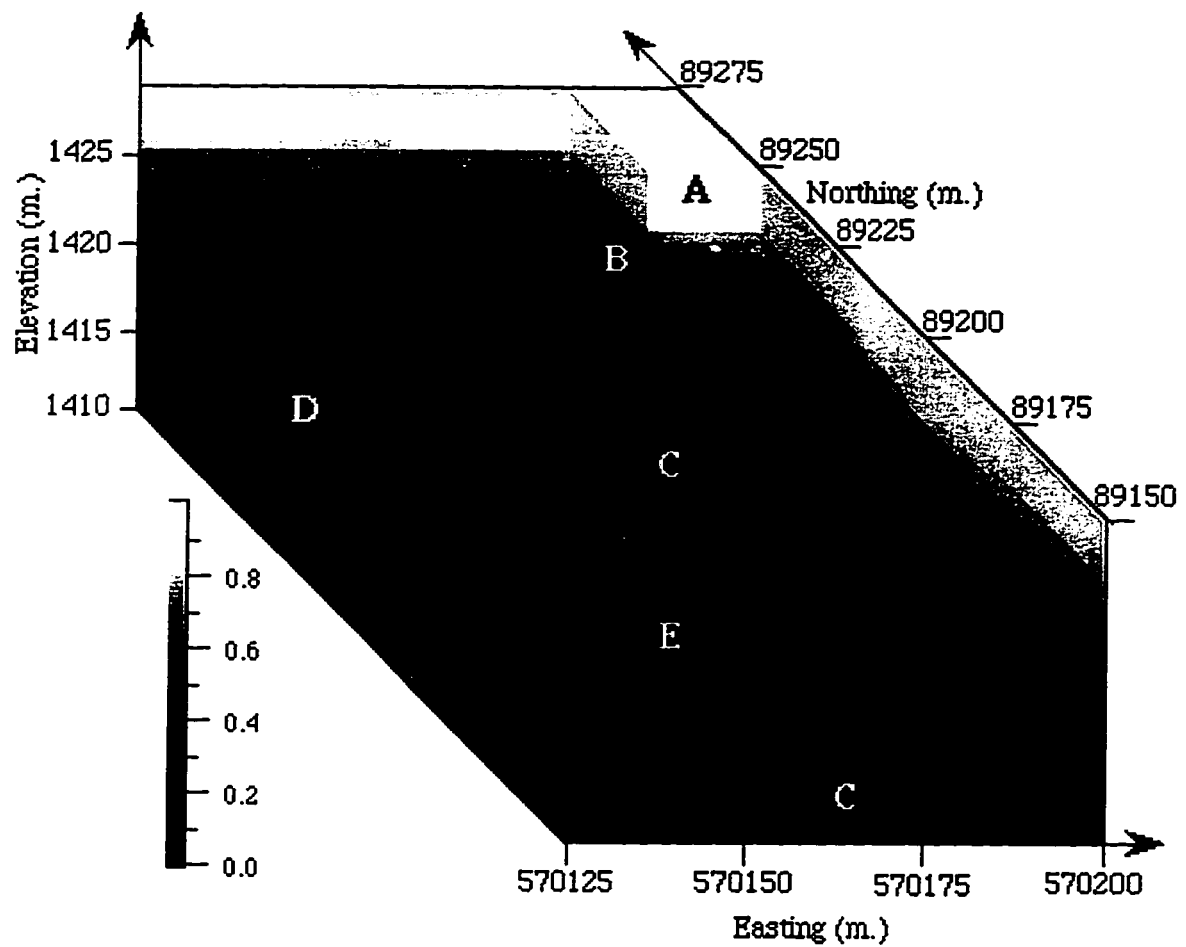
**Figure 2-8.** Variation of average vertical bandwidth with distance from center for the control situation.



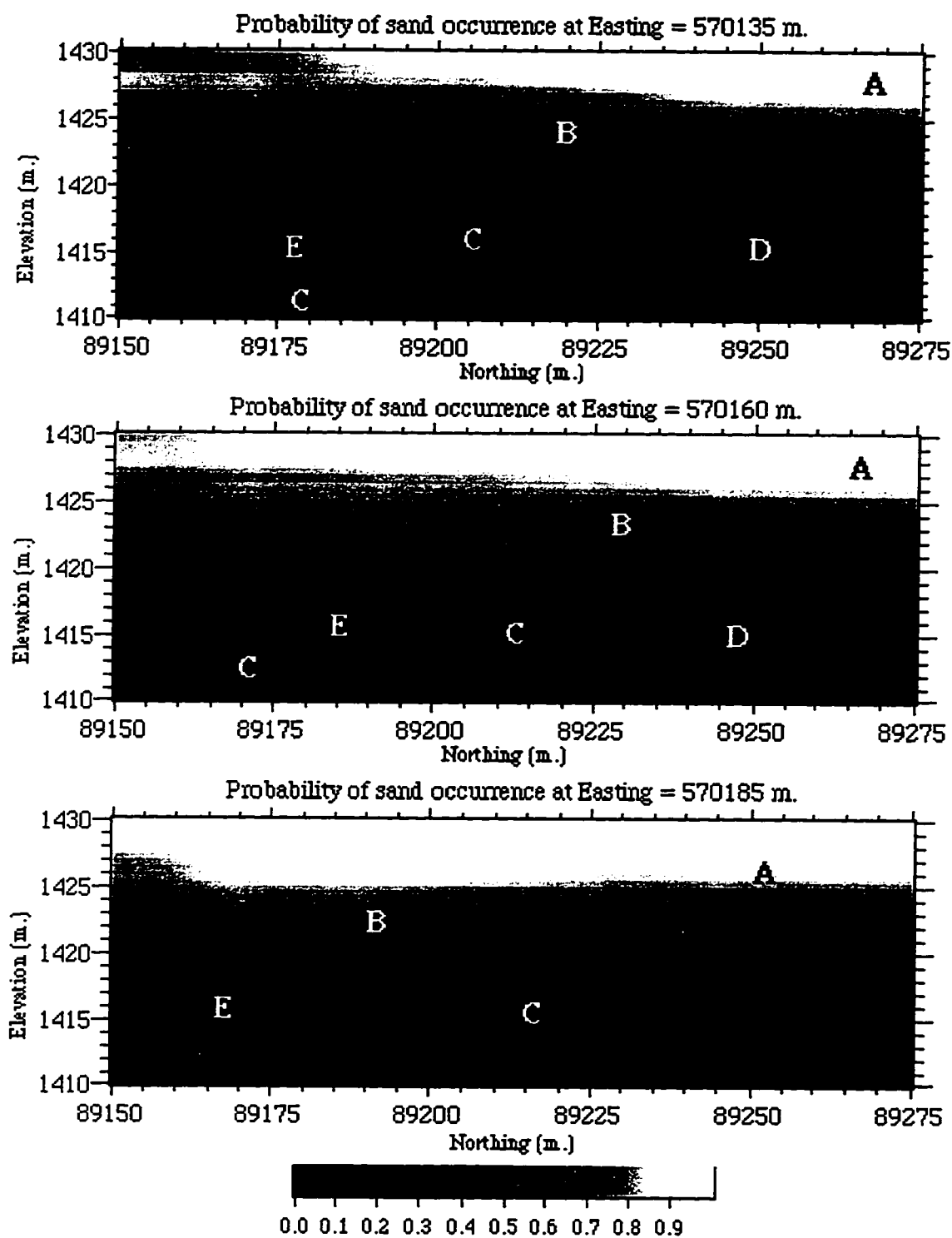
**Figure 2-9.** A plan view for the borehole sites and the location of the estimation region at Ogden Valley. Borehole sites 1 through 6 are used to test predictions of soil types.



**Figure 2-10.** Three-dimensional image of the outside boundaries of the region of estimate. A, B, C, D, and E are probabilistic zones that indicate the likelihood of geologic features occurrence.

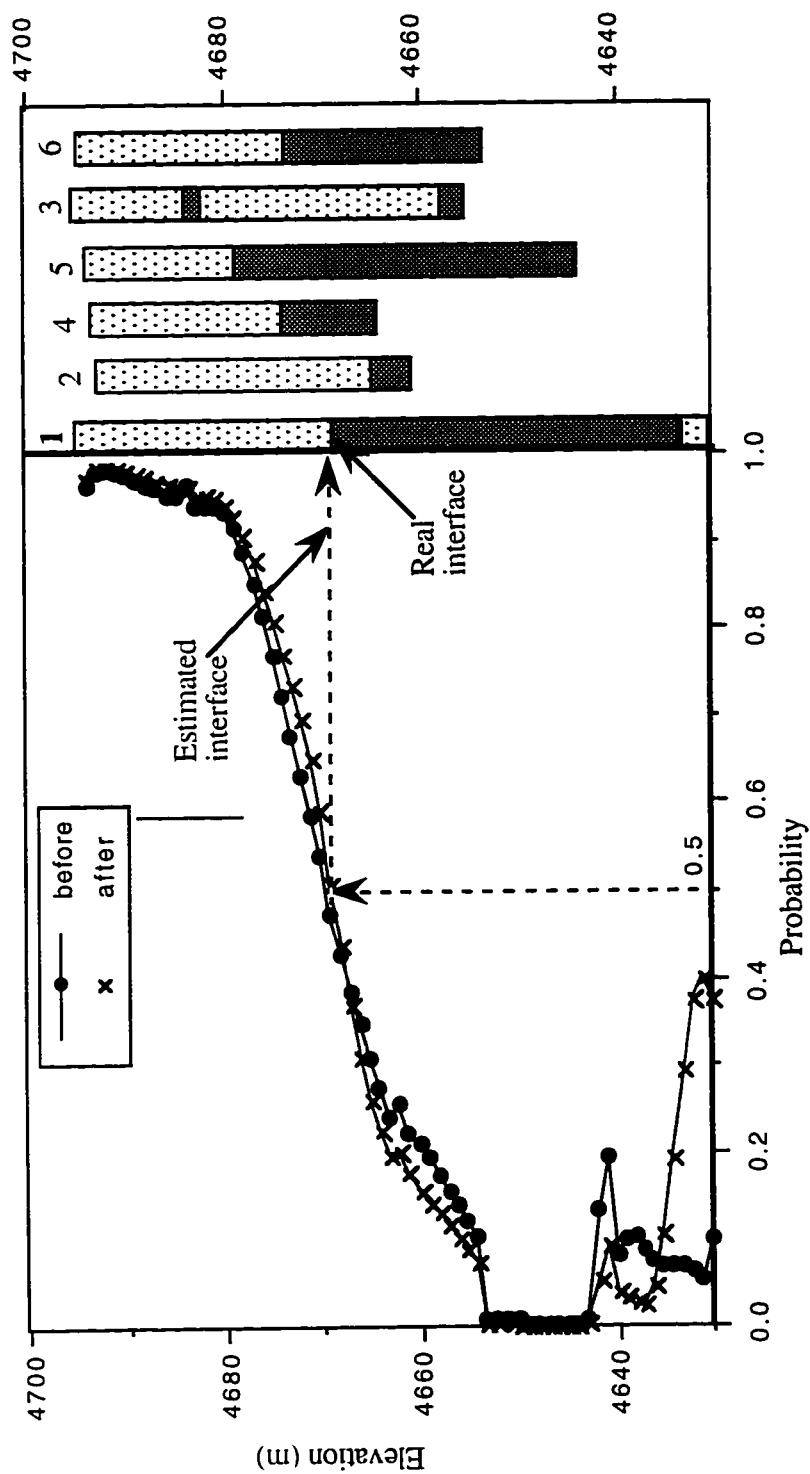


**Figure 2-11.** Three-dimensional cut out of the aquifer system. A, B, C, D, and E are probabilistic zones that indicate the likelihood of geologic features occurrence.

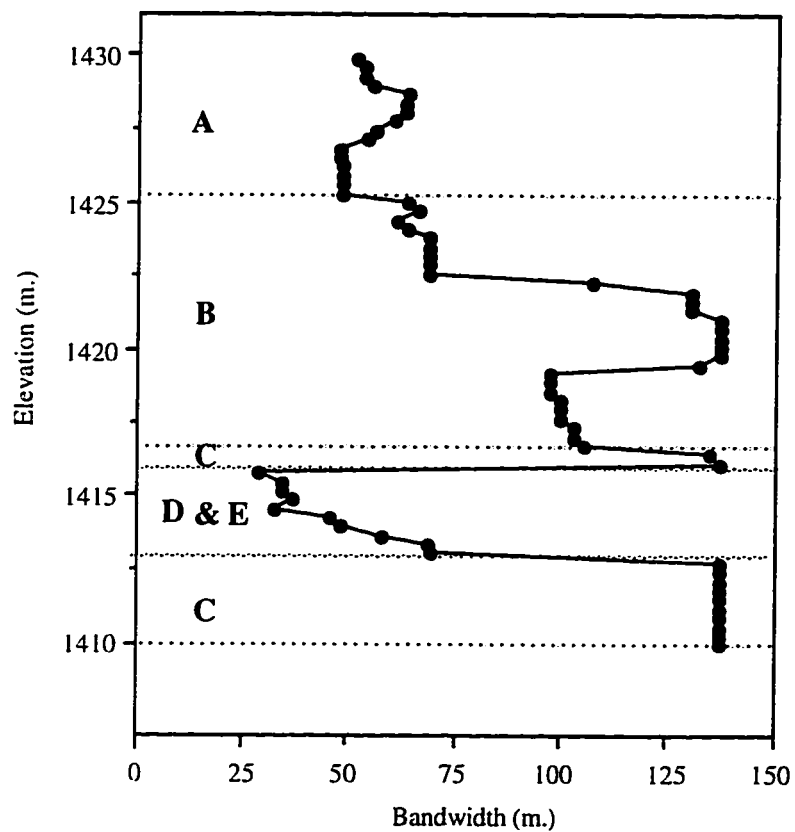


**Figure 2-12.** Elevation-northing contours for  $\lambda$  at three different eastings. A, B, C, D, and E are probabilistic zones that indicate the likelihood of geologic features occurrence.





**Figure 2-13.** Probability versus elevation at borehole 1, before and after dropping borehole log 1. The borehole logs for site 1, and the surrounding sites, 2, 3, 4, 5, and 6, placed according to their distance from site 1 in ascending order, are shown to the right.



CHAPTER 3  
A CONTINUOUS PARAMETER HOMOGENEOUS  
SEMI-MARKOV MODEL FOR STRATIGRAPHIC  
ANALYSES FROM BOREHOLE DATA<sup>1</sup>

Abstract

Markov chain models are often used to represent subsurface stratigraphy and to simulate likely representations of the subsurface. The data used are cores or boreholes. Rock types are classified into a finite set of states, and a state transition probability matrix is estimated for some discretization,  $\Delta z$ , of the vertical sampling domain. The estimated transition probability matrix is sensitive to the  $\Delta z$  value used. No objective methods for selecting a  $\Delta z$  are available. The underlying deposition process is continuous rather than discrete. Hence, information is lost by discretizing the sampling domain, irrespective of the  $\Delta z$  value used.

A simulation and stratigraphic analysis strategy based on a continuous parameter semi-Markov Chain is presented in this chapter. A finite set of rock types is considered, and a state transition intensity matrix is estimated from the borehole data. The sample space is not discretized during estimation. A simulation strategy is developed using the transition intensity matrix to determine the state-by-state transitions, and by bootstrapping (sampling with replacement from the empirical distribution function) a layer thickness corresponding to the new state from the borehole data. Example applications are provided to demonstrate the utility of the method.

---

<sup>1</sup>Coauthored by Alaa Ali and Upmanu Lall.

## Introduction

Stratigraphy in a sedimentary basin is usually characterized from well logs. Well log information is often classified into distinct rock or rock types. The vertical and horizontal stratigraphic changes are generally subtle in nature and can be probabilistically formulated and simulated in terms of such data. The simulated stratigraphy helps locate stratigraphic traps of hydrocarbons, and in the search for oil, gas, or water in delineating reservoirs. Two types of probabilistic models for analyzing subsurface stratigraphy from such data are possible. First, one may consider an estimate of the unconditional probability of observing a particular rock type at a given location in the subsurface. Indicator Kriging [Journel, 1989] and kernel intensity and regression estimators [Ali and Lall, 1993] have been used for this purpose. These methods work with a binary or categorical data set corresponding to rock types or hydraulic conductivity thresholds. Such estimates of the probability of occurrence of a rock type are useful for assessing the spatial variability of rock parameters over the site. However, they do not explicitly consider persistence in the occurrence of a particular type of rock, or the tendency of a certain rock type to follow another, e.g., coal seams commonly follow a seat-earth, and channel-conglomerates may be succeeded by point-bar sandstones [Miall, 1973]. A Markov chain (MC) approach is often used [Anderson and Goodman, 1957; Krumbein, 1968; Harbaugh and Bonham-Carter, 1970; Bayer, 1985; Sinvhall and Sinvhall, 1992] to assess the conditional probability of occurrence of different rock types through a one-step state transition probability matrix. Such an approach allows for a better understanding of the degree of connectivity between different types of rock and is hence useful for analyzing issues related to fluid transport and mixing, and oil traps exploration. Typically, MC models are applied for transitions along the depositional sequence, i.e., in the vertical.

Transitions across states are usually considered across fixed stages or steps using a vertical discretization,  $\Delta z$ , that is prescribed by the investigator. An appropriate  $\Delta z$  value depends on the scale of stratigraphic fluctuation, i.e., how rapidly rock types change in the vertical, and on the resolution of the borehole log. A large  $\Delta z$  relative to the scale of fluctuation leads to loss of information about the depositional sequence. On the other hand, if  $\Delta z$  is taken to be too small, the number of transitions out of a given rock type or states may be very few, leading to a loss of the dependence structure and, hence, to transition probability estimates that are not directly representative of the underlying process scale. This is a classical problem associated with discretization of an underlying continuous process. Despite much research [e.g., *Carr et al.*, 1966; *Krumbein*, 1967; *Miall*, 1973; *Sinvhal and Sinvhal*, 1992], a consensus for the optimal choice of  $\Delta z$  in a given setting has not emerged.

An alternative to such an approach is to use semi-Markov models [*Krumbein and Dacey*, 1969; *Dacey and Krumbein*, 1970; *Schwarzacher*, 1975] where transitions between rock types occur at discrete time intervals, but the deposition rate of a particular rock type during such an interval is allowed to vary randomly. A continuous space model for lithology results. Past efforts at applying such models have suffered from a need to discretize the sample while estimating model parameters and/or assumptions of the parametric form of the probability distribution of rock thickness for each episode of deposition.

In this chapter we present a methodology for estimating parameters of a homogeneous Markov chain model in the vertical in a manner that obviates the need for a priori discretization of the domain. Transition intensities (relative frequency of state transitions per unit length) rather than transition probabilities are estimated from the data. A simulation strategy using these estimated transition intensities and a resampling of layer thicknesses is

also developed. This approach is best classified as a semi-Markov chain model. Applications to a data set from India and to one from Utah are presented to demonstrate the utility of the continuous parameter approach.

### Background

A brief review of discrete and continuous parameter Markov chain models as used for stratigraphic modelling is offered in this section.

#### **Discrete Parameter Homogeneous Markov Chains (DHMC)**

Markov chain (MC) models have been used for subsurface modelling since the 1950s. The occurrence of lithologies is viewed as a stochastic process. The lithology is modelled as a random state variable that takes discrete values,  $X(z)$  (rock types: 1, 2, 3 ), as deposition,  $z$ , progresses. The values of the random variable,  $X$ , are called the states of the MC, while the deposition axis,  $z$ , is called the parameter of the MC. If the random variable,  $X$ , is considered to occur at discrete points (sampled at a resolution of  $\Delta z$ ) along the  $z$  axis, the chain is called a discrete parameter MC. On the other hand, if the random variable  $X$  is allowed to occur at any point, the chain is called a continuous-parameter MC. The Markovian property is stated as:

$$P\{X(z_i) = k \mid X(z_j) = l, X(z_1) = m \dots\} = P\{X(z_i) = k \mid X(z_j) = l\} \quad (z_1 < z_j < z_i) \quad (3-1)$$

where  $z_i, z_j$  are the depths at points  $i$  and  $j$ , respectively.

This property indicates that our knowledge of the state at point  $j$  is enough to infer the state at point  $i$ . First-order dependence is expressed in equation (3-1). Models with higher order dependence can be considered.

In reality, the deposition process varies continuously with time and, hence, with depth. The deposition type, rate, and its persistence vary over the site. However, a first-order, homogeneous MC with a discrete-parameter space (DHMC) has usually been used for the

representation of such a process. The model is prescribed through a state transition probability matrix (TPM).

The TPM is a two-dimensional matrix that summarizes the relative transition frequencies from one state to another (Figure 3- 1). The TPM is computed from a tally matrix. Given  $n$  states, or types of rocks, an element,  $f_{km}(\Delta z)$ , in the tally matrix is the total number of transitions from state  $k$  to state  $m$  at a prescribed sampling interval,  $\Delta z$ . An element in the corresponding TPM,  $p_{km}(z_i, z_{i+1})$ , is the transition probability from state  $k$  at  $z_i$  to state  $m$  at  $z_{i+1}$ , and is given as:

$$p_{km}(\Delta z) = f_{km}(\Delta z) / \sum_{m=1}^n f_{km}(\Delta z) \quad (3-2)$$

where  $\Delta z$  is the sample discretization ( $z_{i+1} - z_i$ ).

A problem associated with the TPM calculation is to define an appropriate sampling discretization,  $\Delta z$ . The sampling discretization,  $\Delta z$ , is assumed to represent a lithologic unit within which the deposition is of the same type. There are two ways to discretize the sampling domain: 1) using break points for rock types and 2) using a fixed  $\Delta z$ .

In the first approach, a sedimentological unit is presumed to have a uniform lithological composition, and transitions are considered across lithologic boundaries. Each unit, regardless of its thickness, represents a step for the MC. The lithologic thickness is not used. This approach may only be valid if the sedimentological units are of nearly the same thickness.

The second approach provides an arbitrary increase in the number of observations. An equal interval sampling provides information about the lithologic thickness but it may result in information loss for lithologic transition. *Bayer* [1985] showed that the tally matrix at a prescribed  $\Delta z$  is equal to the sum of two matrices. The first matrix tallies the transitions to other states and has diagonal elements of value zero, while the second one tallies the

lithologic frequency with off-diagonal elements of value zero. The corresponding probability matrices represent transition (conditional) and total (unconditional) probabilities, respectively.

An illustration of TPM calculations for a synthetic situation with three types of rocks, with each layer of the same thickness (5 ft), and a systematic alternation of layers is shown in Figure 3- 1. For values of  $\Delta z$  larger than or equal to the average bed thickness (e.g., 5 ft. in Figure 3- 1), there is a significant loss of information as many actual transitions are missed. For smaller values of  $\Delta z$  ( 2, or 1 ft in Figure 3- 1), the unconditional probability matrix becomes dominant in the TPM and the resulting transition probabilities become small. If the sampling interval is much smaller than the average bed thickness, the transition probability is extremely small [Davis, 1973]. If the beds have roughly the same thicknesses, a value of  $\Delta z$  near the average thickness may provide a TPM that approximates the process to a certain extent. On the other hand, if the bed thickness is highly variable, a useful choice of  $\Delta z$  may be difficult. This is usually the case in practice.

This sampling problem was the focus of several studies. *Krumbein* [1967] suggested sampling intervals between 2 to 10 ft. *Sinvhal and Sinvhal* [1992] suggested an interval of 2m (6 ft) or its multiple, and *Miall* [1973] found that a sampling interval slightly less than the average bed thickness gives satisfactory results. In general, no theoretical rules have emerged.

Using the DHMC, simulation proceeds in a series of space steps  $\Delta z$  based on the estimated TPM. Successive rock types are randomly sampled in steps of  $\Delta z$  based on the TPM entries. There is usually no explicit consideration of layer thickness statistics. The thrust of this chapter is to develop a Markov chain model of subsurface stratigraphy that explicitly considers layer thickness statistics, recognizes the continuous nature of the



underlying depositional process, and does not suffer from the problems associated with sample discretization.

### **Continuous-Parameter Homogeneous Markov Chain and Semi-Markov Models**

Given the continuous nature of the depositional processes, a continuous-parameter homogeneous Markov chain (CHMC) representation may be more appropriate than the DHMC. *Keiding and Anderson* [1989] defined the transition intensity,  $q_{km}$ , as the ratio of the number of transitions from state  $k$  to state  $m$  to the total number of events of state  $k$ . Using this definition, the transition intensity,  $q_{km}$ , may be defined directly as the ratio of the total number of  $k$ -to- $m$  transitions to the total thickness of state  $k$  observed along the entire profile of length  $L$ , and is given as:

$$q_{km} = n_{km} / L_k^* \quad (3-3)$$

where  $n_{km}$  is the number of transitions from state  $k$  to state  $m$ , and  $L_k^*$  is the total length of rock type  $k$  from which transitions to other rock types occur.  $L_k^*$  does not include the last layer in each borehole from which no transition takes place. A transition intensity matrix (TIM) comprised of elements  $q_{km}$  is used to describe a continuous parameter MC [Trivedi, 1982, p. 362].

A computation of the TIM using equation (3-3) for the synthetic data discussed earlier is shown in Figure 3- 1(e). The quantity  $q_{km}$  is informative as it is a measure of both the lithologic transition and thickness. The direct definition of,  $q_{km}$ , in equation (3-3) is independent of  $\Delta z$ .

*Krumbein* [1968] used a continuous parameter Markov chain to simulate stratigraphic sequences. The TPM is first estimated using a discretization,  $\Delta z$ . A TIM element,  $q_{km}$ , is then derived from the TPM as follows:

$$q_{km} = \frac{p_{km}}{\sum_{j=1 \neq k}^n p_{kj}} M_k \quad (3-4)$$

where  $M_k = -\ln(p_{kk})/\Delta z$ .

Krumbein's definition provides a formal relation between the transition intensities and probabilities, but it suffers from the problems of sampling discretization because it is derived from a TPM estimated at a resolution of  $\Delta z$ . The computation of  $q_{km}$  and its sensitivity to  $\Delta z$  are shown in Figure 3- 1. Clearly the direct estimates (equation 3-3) are to be preferred. The direct estimates are based on the lithologic transitions and are independent of any sampling discretization.

*Schwarzacher* [1975] discusses the use of semi-Markov chain models for sedimentological analyses. The structure of such a model differs from that of a Markov chain model in the following manner. A Markov transition matrix, **A**, governs the occurrence of successive rock types, as in the MC model. However, instead of considering transitions at discrete distance markers, one considers transitions out of a state to other states, after a certain thickness,  $z_j$ , has been randomly deposited in a state  $j$  that follows a probability distribution  $f(z_j)$ . The physical interpretation is that the transitions take place after fixed deposition time intervals (rather than at fixed vertical horizons) during which a certain (random) sedimentation rate for that state is observed. Elements of the matrix **A** are the probabilities with which state transitions take place at these time markers. Transitions to the same state are allowed in **A**. Thus, the semi-Markov model leads to a discrete-time but continuous-space formulation. *Schwarzacher* [1975] discusses theoretical aspects of the semi-Markov model, provides expressions for the probability of being in a certain state given the initial state, and discusses parameter estimation with data from stratigraphic sections. He notes that problems similar to the analysis of Markov chains arise depending on how one samples the records (i.e., focusing only on lithologic transitions, equally

spaced sampling, or bed-by-bed sampling). A parametric probability density model (negative exponential or gamma) is assumed in his derivations. The layer thicknesses for a classical Markov chain model are geometrically distributed. An advantage of the semi-Markov model is that other possibly more representative models for layer thickness probabilities can be utilized.

A particular semi-Markov formulation results if the matrix  $A$  is considered to have a zero diagonal, i.e., only transitions out of the current state are considered at the end of each period. Here, one no longer uses the discrete time deposition sequence analogy, and considers simply an episodic lithologic sequence with a layer of a certain rock type with some thickness deposited during each episode. The next deposition episode corresponds to a rock of a different type. In each case, a probability distribution,  $f(z_k)$ , for the layer thickness  $z_k$  may be considered for deposition of the rock of type  $k$ . This is the approach followed in this study, and also by *Krumbein and Dacey* [1969] and *Dacey and Krumbein* [1970], who developed a simulation strategy using the semi-Markov model that is based on an equally spaced sampling (i.e., discretization  $\Delta z$ ) of the boreholes. The thickness of a rock layer is described by them using an exponential probability distribution fitted to the data on bed thickness for that rock type. A layer thickness is first generated from this fitted distribution. The next rock type is determined by sampling from an off-diagonal transition probability matrix (OTPM) formed to recognize transitions to other rock types only. A nonzero element,  $\dot{p}_{km}$ , in the OTPM is then estimated as:

$$\dot{p}_{km} = \frac{q_{km}}{M_k} \quad m=1..n; \neq k; \quad \dot{p}_{kk}=0 \quad (3-5)$$

Recall that the transition intensities,  $q_{km}$ , are estimated using equation (3-4), where the TPM was evaluated using a discretization of  $\Delta z$ . *Harbaugh and Bonham-Carter* [1970] noted that bed thicknesses are more likely to be log-normally distributed. However, *Potter*

and Blakely [1967] noted that sampling the bed thicknesses directly from the empirical distribution function of the data (i.e., bootstrapping) is more attractive. All the methods discussed in these references are sensitive to the sample discretization, since the TIM is derived using equation 3-4, rather than the direct estimate of equation (3-3).

#### A Continuous-Parameter, Homogeneous Semi-Markov (CHSM) Simulation Model for Stratigraphy

A strategy for generating pseudo-bore holes using an empirical semi-Markov model is presented. The proposed simulation strategy is similar to the models used by *Krumbein and Dacey* [1969] and by *Potter and Blakely* [1967]. Layer thicknesses for a particular rock type are resampled from observed thicknesses for that type of rock, and transitions to other states are modelled by the OTPM. The procedures used for simulating a stratigraphic sequence upwards from the lowest horizon considered are outlined below and in the flow chart (Figure 3- 2).

#### **Define Initial State Probability Matrix (IPM) and Generate Initial State**

The first layer of the stratigraphic sequence is sampled from the initial probability matrix (IPM) of rock types,  $\pi_k$ , defined as :

$$\pi_k = \frac{n_k}{N} \quad (3-6)$$

where  $n_k$  is the number of bore holes starting with rock  $k$ , and  $N$  is the total number of bore holes. Since a homogeneous Markov chain is assumed to describe the stratigraphy, one may wish to sample the initial state from the unconditional probability matrix (UPM) of rock types given by:

$$p_k = \frac{L_k}{L} \quad (3-7)$$

where  $L_k$  is the total length of rock type  $k$ , and  $L$  is the total bore hole length.

However, in a Markov process (as opposed to a purely random process), one does need a specification of the initial condition that corresponds to the observed one at  $z=0$ . Using the IPM is an appropriate way to provide such an initial state.

### **Define Bed Thickness Cumulative Distribution Function for Each Rock Type**

The cumulative distribution function,  $F(d_k)$ , of the thickness,  $d_k$ , of the  $k^{\text{th}}$  rock type is defined as:

$$F(d_k \leq D) = \frac{n_{d_k, D}}{N_{d_k}} \quad (3-8)$$

where  $n_{d_k, D}$  is the number of bed thicknesses of rock type  $k$  that are smaller than  $D$ , and  $N_{d_k}$  is the total number of beds of rock type  $k$ .

A bed thickness for the  $i^{\text{th}}$  layer that is assigned the rock type  $k$ , can then be sampled using  $F(d_k)$ . This strategy will honor the observed data, but will not allow the generation of bed thicknesses that are smaller than or larger than those recorded by the boreholes. Parametric distributions for  $d_k$  may be more attractive where there is only limited data. In our experiments, the strategy proposed here generally worked better than the assumption of an exponential distribution for the layer thicknesses. Kernel estimators for estimating the cumulative distribution function [Lall *et al.*, 1993] or the quantile function [Moon and Lall, 1994] may also be useful.

### **Estimate Off Diagonal Transition Probability Matrix (OTPM)**

An off-diagonal TPM is used to determine transitions out of a rock type into other rock types. An element  $p_{km}$  is computed as:

$$p_{km} = \frac{q_{km}}{\sum_{j=1 \neq k}^n q_{kj}} \quad m=1..n; \neq k; \quad p_{kk}=0 \quad (3-9)$$

The denominator in equation (3-9) is shown to be equal to  $M_k$  by *Krumbein* [1968]. Equation 3-9 is interpreted by observing that the probability of transition out of a state at the end of the deposition episode is seen to be proportional to the transition intensity out of the state. Given a current state  $k$ , the next state  $m$  is determined using the transition probabilities,  $p_{km}$ , to the  $(n-1)$  states excluding  $k$ , as estimated by equation (3-9).

The simulation proceeds until a desired profile length has been simulated. Each simulated sequence can be considered to be a synthetic bore hole.

### Applications

Applications to two field data sets are presented. The purpose of the first application is to compare the CHSM model and the DHMC model with data from a published example. The purpose of the second application is to show the capability of the CHSM model in preserving attributes of fine resolution data. The adequacy of fit is judged by the ability of the simulations to reproduce three sets of statistics: 1) the UPM of the original data (to measure how well the relative frequency of each rock type is preserved); 2) the TIM of the original data (to measure preservation of attributes related to rock type continuity); and 3) the bed thickness statistics, such as CDF, mean, and standard deviation, for each rock type. These simulation statistics are compared with the corresponding statistics estimated from the original bore hole data.

#### Application 1

The study area in this section is a sedimentary basin located in western India. These data were published as a DHMC example by *Sinvhal and Sinvhal* [1992]. The formation under consideration is of Middle Eocene age. It is 200-300 m thick in the middle and thins out to 30-40 m. It is comprised of rhythmic layers of sandstone, coal, and shale (i.e., three types of rocks) belonging to a deltaic environment. A data set of three boreholes (see

Figure 3-3) was used in modelling the deeper part of the basin margin. The data were sampled every 2 m, i.e., at a rather coarse resolution. *Sinhval and Sinhval* [1992] used a DHMC with  $\Delta z=4$  m for this data set. The data appear to be nonstationary, e.g., coal is found only near the lower part of the section. No transitions are observed between sandstone and coal, which is consistent with the geologic principles. Sandstone is a consequence of a high energy depositional process, which is usually followed by a low energy depositional process leading to shale formations. The low energy process is then followed by coal depositions.

Estimates of the TPM, IPM, TIM, and OTPM are presented in Table 3-1. The cumulative distribution functions  $F(d_k)$  for each of the three types of rock are shown in Figure 3-4 (solid line).

Pseudo-bore holes are simulated using the previous quantities and procedures in Figure 3-2 for CHSM simulation, and the conventional procedures for DHMC simulation [Harbaugh and Bonham-Carter, 1970]. A set of 100 realizations was created for each model. Five realizations from each model are shown in Figures 3- 5, and 3-6.

Realizations obtained by the CHSM show more realistic pictures of the environment than those obtained by DHMC because the bed thickness is naturally sampled from data rather as a multiple of  $\Delta z$  (see Figure 3-3, 3-5, and 3-6). This problem is significant when the bed thickness is relatively small. DHMC realizations show much thicker coal layers than those observed, due to the choice of  $\Delta z$ . Statistical comparisons of the two approaches bear out this observation.

Figure 3-7 shows boxplots for the UPM elements of the 100 realizations obtained from CHSM and DHMC, respectively. Both models appear to reproduce the UPM.

Figure 3-8 shows boxplots of the TIM elements of the 100 realizations obtained from the CHSM and the DHMC. Note that the DHMC poorly reproduces the TIM element from

coal to shale. The coal layers in the data are thinner than the  $\Delta z$  used (4 m). Consequently, the simulated coal layers are too thick, resulting in a lower transition intensity out of these layers than observed. This problem is exacerbated as bed thicknesses for different types of layers vary.

Table 3-2 shows the statistics (mean and standard deviation) of the bed thickness for 1) the real data, 2) simulated images based on the CHSM, 3) simulated images based on DHMC with  $\Delta z=4$  m, and 4) simulated images based on DHMC with  $\Delta z=2$  m. We observe for the DHMC that the thickness statistics for each rock type generally improve with  $\Delta z=2$  m. This improvement is dramatic for shale and coal. However, the standard deviation of sandstone and shale thicknesses is now underestimated. The overall performance of the CHSM statistics appears to be better than that of the DHMC with the  $\Delta z=4$  m used by Sinhvai and Sinhvai, and comparable to or better than that of the DHMC fitted at the data resolution of 2 m. Figure 3-4 shows three plots, one for each rock type, in which the empirical CDF of the bed thickness is plotted for the real data, and images based on CHSM, DHMC with  $\Delta z=4$  m, and DHMC with  $\Delta z=2$  m. We notice that the CDFs of layer thickness are reproduced better by CHSM than the DHMC at either resolution.

## **Application 2**

The Ogden Valley has an aquifer system that is typical of Lake Bonneville sediments that cover large portions of the state of Utah. The site under consideration is located just west of the Wasatch Mountain Range on the relict Weber Delta. The delta consists of broad plains and terraces, and originates along the western base of the Wasatch Range. Topographically, this site is on a plateau formed by the Weber Delta. The plateau is approximately 90 m above the valley floor. Surface elevations at this site vary from 1400 m above mean sea level along the western side, to 1540 m near the eastern side. Depth to



bedrock in the basin ranges from 460 m on the western side to 2300 m on the east. The available data lie in the upper 20 m of unconsolidated geologic material that ranges from a very low permeability rock, (clay) to a very high permeability rock (sand and gravel). rocks observed along borehole profiles had been classified in to 14 types. We considered three major classes of rocks: 1) low permeability layers (clay), 2) high permeability layers (sand and gravel), and (3) mixed permeability layers (silty sand, silty gravel). The data used for the application were from 10 boreholes (see Figure 3-9). The vertical resolution of these data is 1 cm. The statistics computed from the original data are reported in Table 3-3, and CDFs of layer thickness for each rock type are provided in Figure 3-10 (solid line). One hundred pseudo-bore holes were generated using the CHSM model. Five of these realizations are shown in Figure 3-11. They are seen to be visually consistent with the real bore holes shown in Figure 3-9.

Boxplots of the UPM and TPM (Figures 3-12 and 3-13), a table of the mean and standard deviation of the bed thickness for each rock type (Table 3-4), and plots of the CDF of layer thickness for each rock type (Figure 3-10) all show that the model preserves selected statistics of this high resolution data set. The CDFs from the simulations are seen to smooth the empirical CDFs from the observed data. This is an expected result. Some small biases in the thickness statistics are also evident.

### Summary

A methodology for the simulation of stratigraphic sequences using a continuous parameter homogeneous semi-Markov model is presented in this chapter. The simplicity and nonparametric flavor of the discrete-parameter Markov chain model is retained, while problems and loss of information stemming from the sampling discretization in the discrete approach and in the work of *Krumbein* [1968] and *Potter and Blakely* [1967] are

circumvented. The applications presented demonstrated the utility of the method in preserving the statistical attributes of the stratigraphy as seen from the boreholes. The simulator presented should be useful in a number of contexts. One may be interested in testing hypotheses related to cyclicity in deposition at a site. The CHSM Monte Carlo simulations could be used to test the significance of any such claims or the adequacy of a physically or conceptually based model of deposition relative to the sampling of the environment using drill logs. Near surface groundwater contamination and the possibility of this contaminated water coming in contact with water in deeper semiconfined aquifers is a major concern in the western United States. Investigations of the geologic structure of the basin through simulations of likely stratigraphic sections at the site may be useful in analyzing such issues. Formal tools for adapting the CHSM model for such analyses are being developed.

A limitation of the model presented is that the layer thicknesses for each rock type are restricted to those observed in the boreholes. Where this is of concern, a parametric (e.g., log normal) or nonparametric (e.g. Kernel density estimate) estimate of the cumulative distribution of bed thickness may be used. The advantage of the approach used here is that arbitrary underlying probability densities (e.g., a bimodal density corresponding to layers either being thick or thin) can be automatically reproduced.

In practice, the assumption of a homogeneous Markov process generating the sedimentary environment is not likely to be valid. Work is in progress to extend the method presented here to model nonstationarity in the vertical by allowing the CHSM parameters to vary smoothly over the site, in the vertical as well as in the horizontal.

#### NOMENCLATURE

MC        Markov Chain

DHMC    Discrete-Parameter Homogeneous Markov Chain

TPM	Transition Probability Matrix
CHMC	Continuous-Parameter Homogeneous Markov Chain
TIM	Transition Intensity Matrix
OTPM	Off Diagonal Transition Probability Matrix
CHSM	Continuous-Parameter, Homogeneous Semi-Markov
IPM	Initial-State Probability Matrix
CDF	Cumulative Distribution Function
UPM	Unconditional Probability Matrix

#### References

- Ali, A. I., and U. Lall, Interpretation of drill log data: DLOG3D-A probabilistic tool for analyzing subsurface soil variability, *Report No. RR-93-HWR-UI/001*, 45 pp., Utah Water Research Laboratory, Utah State University, Logan, Utah, 1993.
- Anderson, T. W., and L. A. Goodman, Statistical inference about Markov Chains, *Ann. Math. Stat.*, 28, 89-110, 1957.
- Bayer, U., *Pattern Recognition Problems in Geology and Paleontology*, 229 pp., Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1985.
- Carr, D. D., A. Horowitz, S. V. Harbar, K. F. Ridge, R. Rooney, W. T. Straw, W. Webb, and P. E. Potter, Stratigraphic sections, bedding sequences, and random processes, *Science*, 154, 1162-1164, 1966.
- Dacey, M.F., and W. C., Krumbein, Markovian models in stratigraphy, *J. Int. Assoc. Math. Geol.*, 2, 175-191, 1970.
- Davis, J. C., *Statistics and Data Analysis in Geology*, 550 pp., John Wiley and Sons, Inc., New York, 1973.
- Harbaugh, J. W., and G. Bonham-Carter, *Computer Simulation in Geology*, 575 pp., Wiley-Interscience, New York, 1970.
- Journel, A. G., *Fundamental of Geostatistics in Five Lessons*, Stanford Center for Reservoir Forecasting, Stanford University, Stanford, Calif., 40 pp., 1989.
- Keiding, N., and P. K. Anderson, Nonparametric estimation of transition intensities and transition probabilities, *Appl. Stat.*, 38(2), 319-329, 1989.
- Krumbein, W. C., FORTRAN IV computer programs for Markov chain experiments in geology, *Computer Contribution 13*, 38 pp., Geological Survey of Kansas, 1967.

- Krumbein, W. C., FORTRAN IV computer program for simulation of transgression and regression with continuous time markov models, *Computer Contribution* 26, 38 pp., Geological Survey of Kansas, 1968.
- Krumbein, W. C., and M. F., Dacey, Markov chains and embedded Markov chains in geology, *J. Int. Assoc. Math. Geol.*, 1, 79-96, 1969.
- Lall, U., Y. Moon, and K. Bosworth, Kernel flood frequency estimators: Bandwidth selection and kernel choice, *Water Resour. Res.*, 29(4), 1003-1015, 1993.
- Miall, A. D., Markov chain analysis applied to an ancient alluvial plain succession, *Sedimentology*, 20, 347-364, 1973.
- Moon, Y., and U. Lall, A kernel quantile function estimator for flood frequency analysis, *Water Resour. Res.*, 30(11), 3095-3104, 1994.
- Potter, P. E., and R. F., Blakely, Generation of a synthetic vertical profile of a fluvial sandstone body, *J. Soc. Pet. Eng.*, 243-251, 1967.
- Schwarzacher, W., *Sedimentation Models and Quantitative Stratigraphy*, Development in sedimentology 19, 382 pp., Elsevier Scientific Publishing Company, Amsterdam, 1975.
- Sinvhal A., and H. Sinvhal, *Seismic Modelling and Pattern Recognition in Oil Exploration*, 178 pp., Kluwer Academic Publishers, Dordrecht, 1992.
- Trivedi, S. K., *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*, 624 pp., Prentice-Hall, Inc., Englewood Cliffs, NJ, 1982.

**Table 3-1.** Markov Chain Properties for the Indian Site

	IPM	UPM	TIM (m) <sup>-1</sup>			OPTM			TPM based on $\Delta z=4$ m.		
			Sand stone	Shale	Coal	Sand stone	Shale	Coal	Sand stone	Shale	Coal
Sandstone	0	.335	--	.101	0	--	1	0	0.681	0.319	0.000
Shale	1	.630	.059	--	.022	.728	--	.272	0.162	0.784	0.054
Coal	0	.035	0	.389	--	0	1	--	0.000	0.800	0.200

**Table 3-2.** Statistics of the Bed Thickness for Three Types of Soil for: (a) the Real Data (Indian Site); and Three Sets of 100 Simulated Images Generated Using: (b) CHSM, (c) DHMC with  $\Delta z=4$ . Meters, and (d) DHMC with  $\Delta z=2$ . Meters

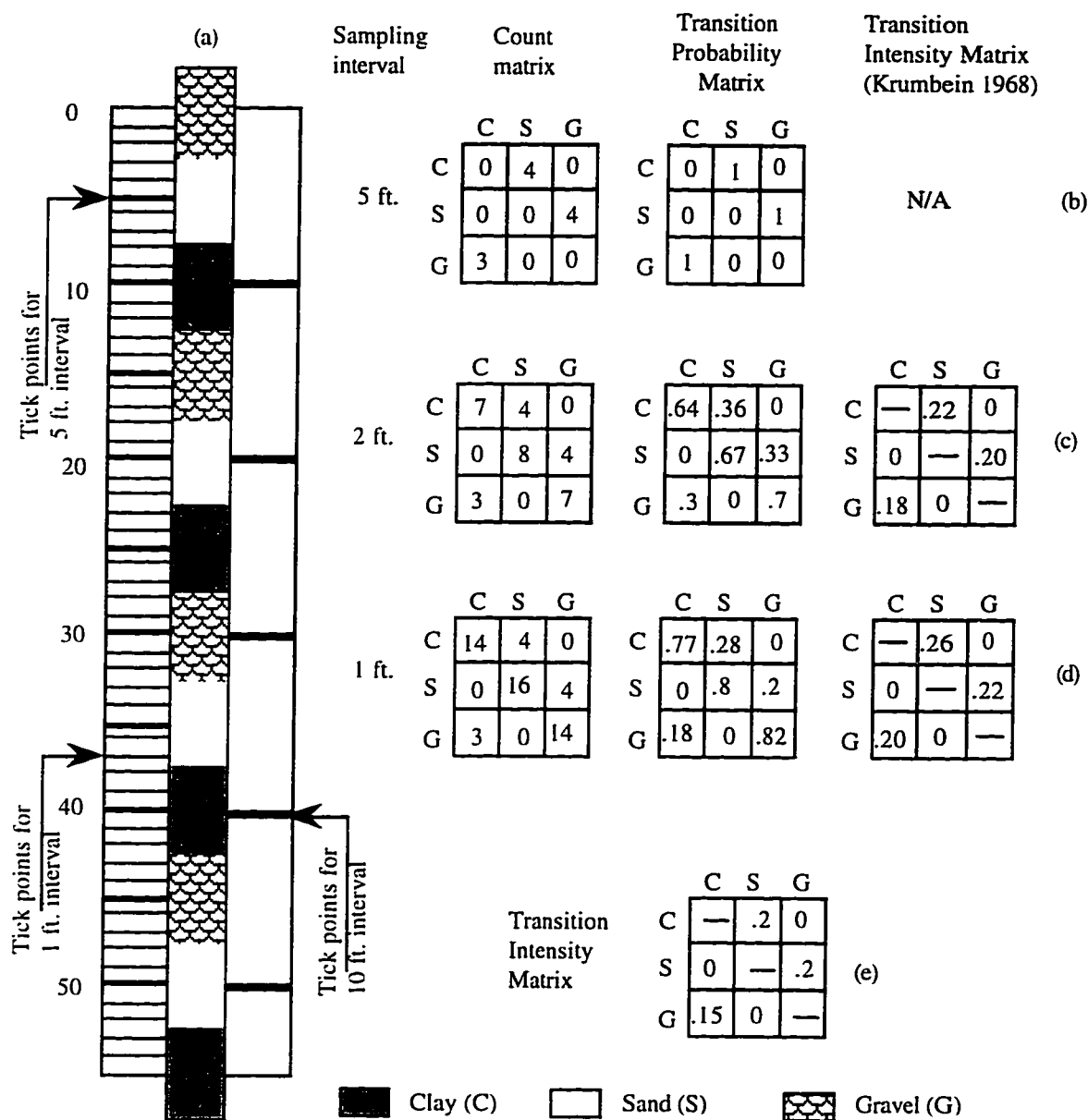
	(a)		(b)		(c)		(d)	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Sandstone	8.50	10.50	9.05	11.4	12.4	9.9	9.84	8.66
Shale	11.76	11.84	12.46	12.6	17.5	14.7	11.93	10.75
Coal	2.56	0.90	2.57	0.97	5.1	2.22	2.6	1.22

**Table 3-3.** The Markov Properties for the Ogden Site

	IPM	UPM	TIM (m) <sup>-1</sup>			OPTM		
			Clay	Silt	Sand	Clay	Silt	Sand
Clay	.8	.444	--	.16	.15	--	.523	.477
Silt	.2	.258	.39	--	.23	.624	--	.376
sand	.0	.298	.07	.28	--	.198	.802	--

**Table 3-4.** Statistics of the Bed Thickness for Three Types of Soil for: (a) The Real Data (Ogden Site), and (b) 100 Simulated Images Based on the CHSM (All Units in Meters).

	(a)		(b)	
Soil type	Mean	Standard deviation	Mean	Standard deviation
Clay	3.44	2.99	3.14	2.74
Silt	1.68	1.34	1.62	1.27
Sand	3.01	2.35	2.80	2.13



**Figure 3-1.** A stratigraphic section illustrating: (a) the drill log and different sampling intervals  $\Delta z$ ; the count matrix, the TPM, and the TIM based on Krumbein (1968) if  $\Delta z$  is (b) 5 ft. (i.e., equal to spacing between transition points); (c) 2 ft.; (d) 1 ft.; (e) the TIM based on continuous sampling. Note how the results in b, c, and d are sensitive to  $\Delta z$ . The results in (e) are not dependent upon  $\Delta z$ .

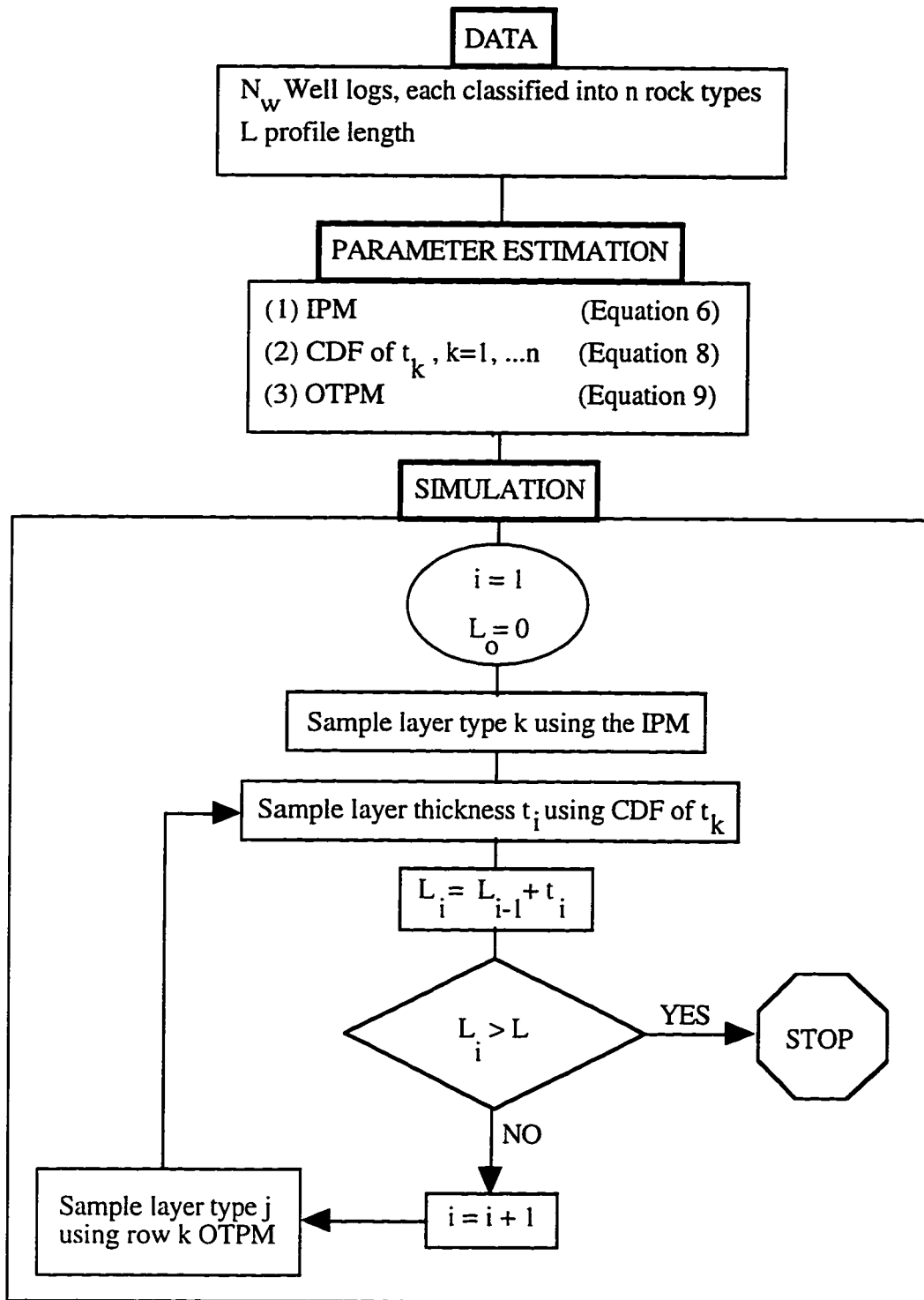
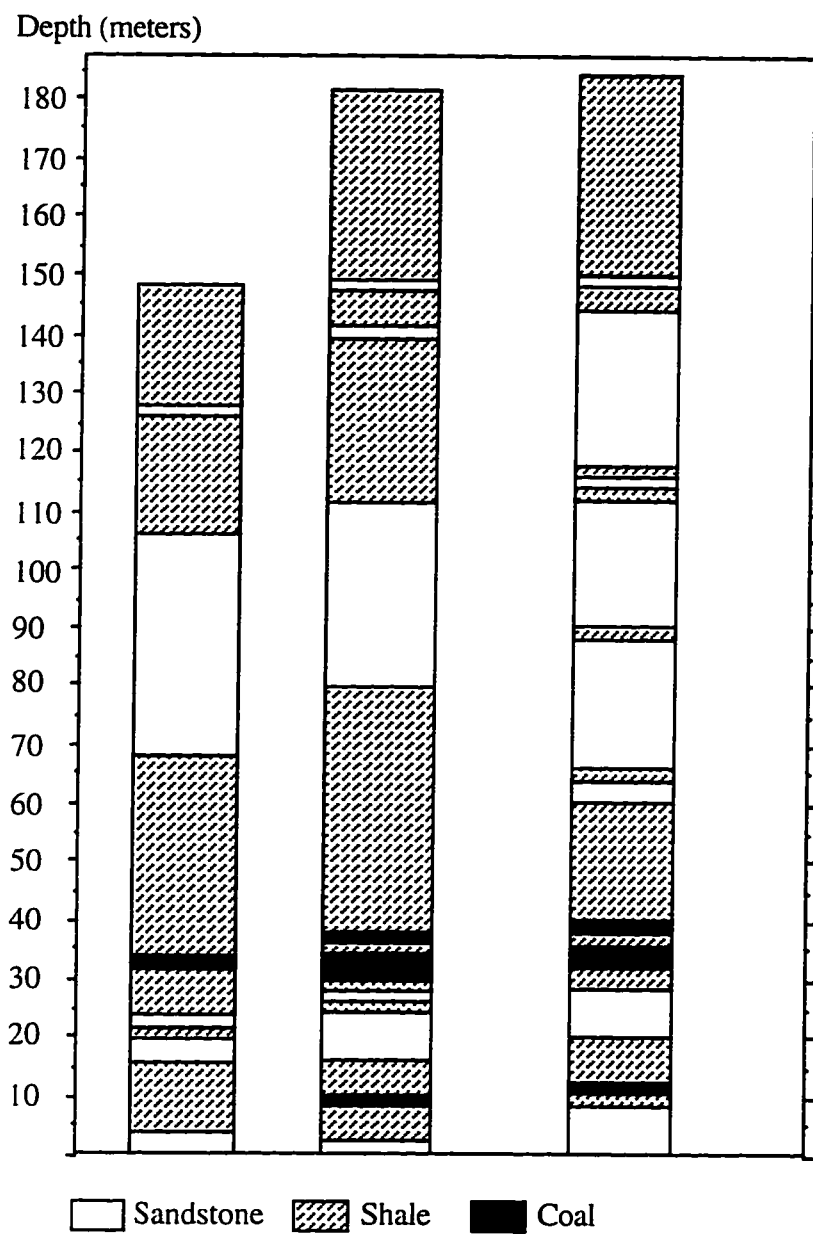


Figure 3-2. Simulation procedures for a CHSM.





**Figure 3-3.** Three bore holes representing a subregion in a sedimentary basin in western India [*Sinvhal and Sinvhal, 1992*].

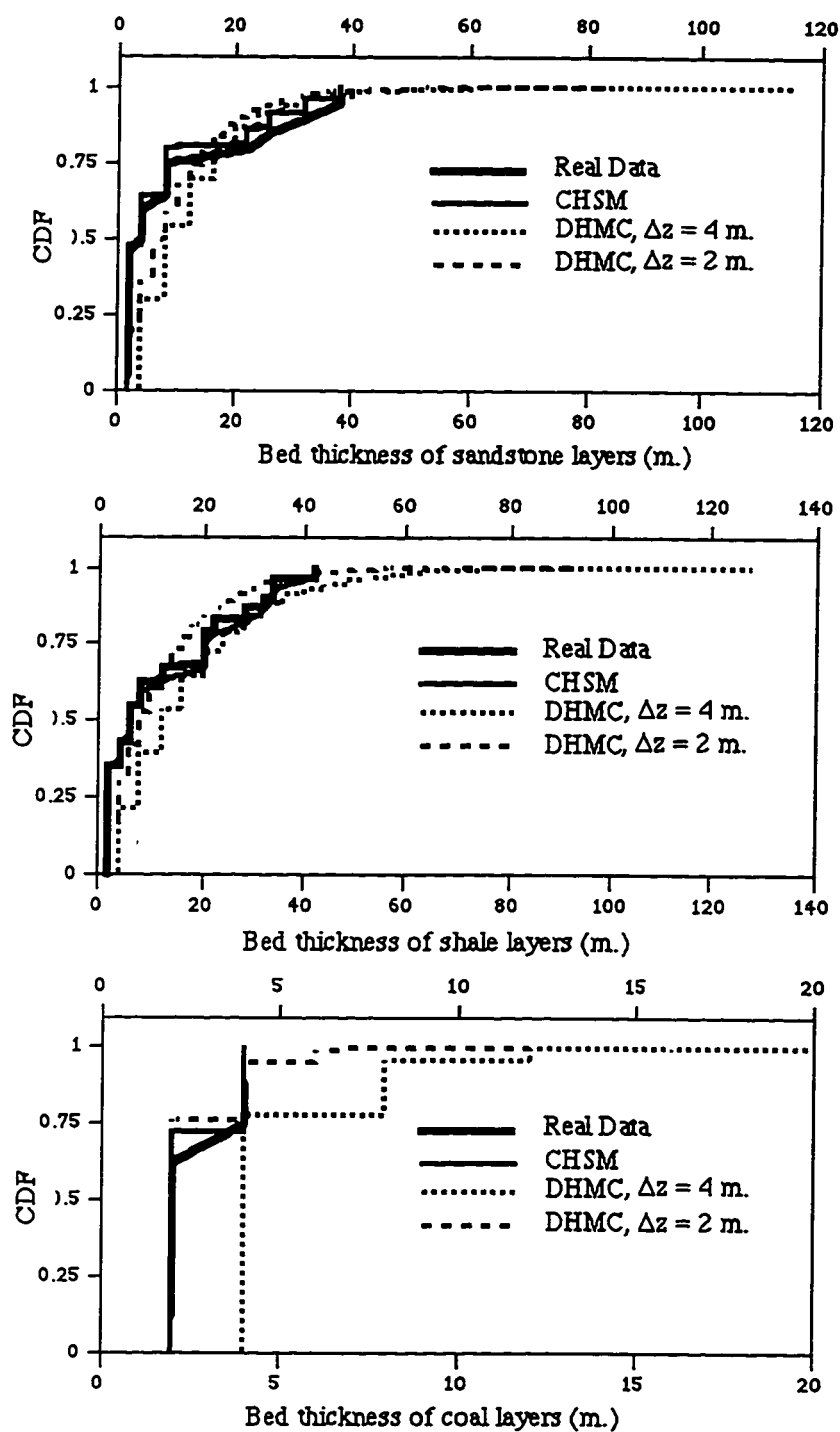
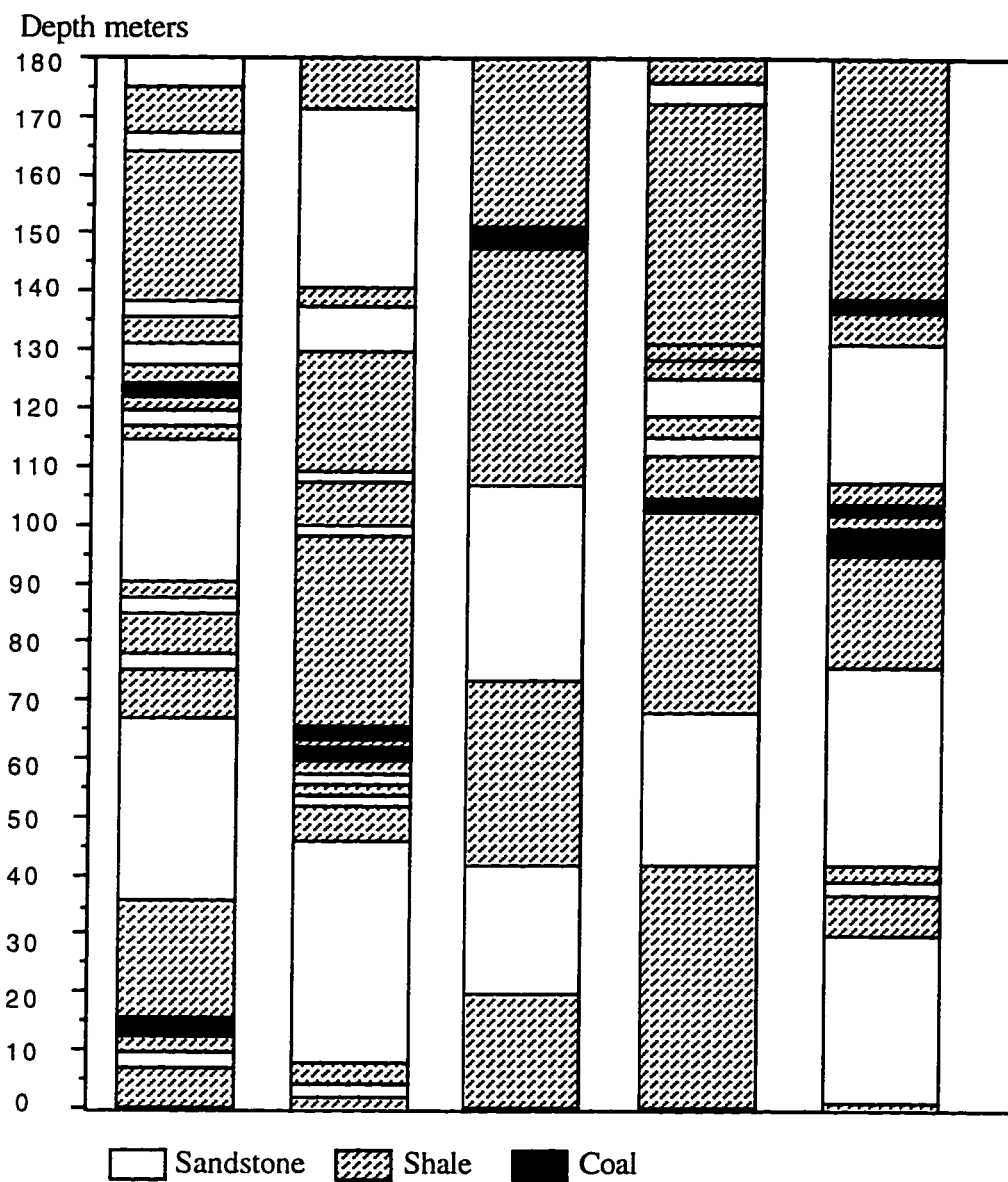
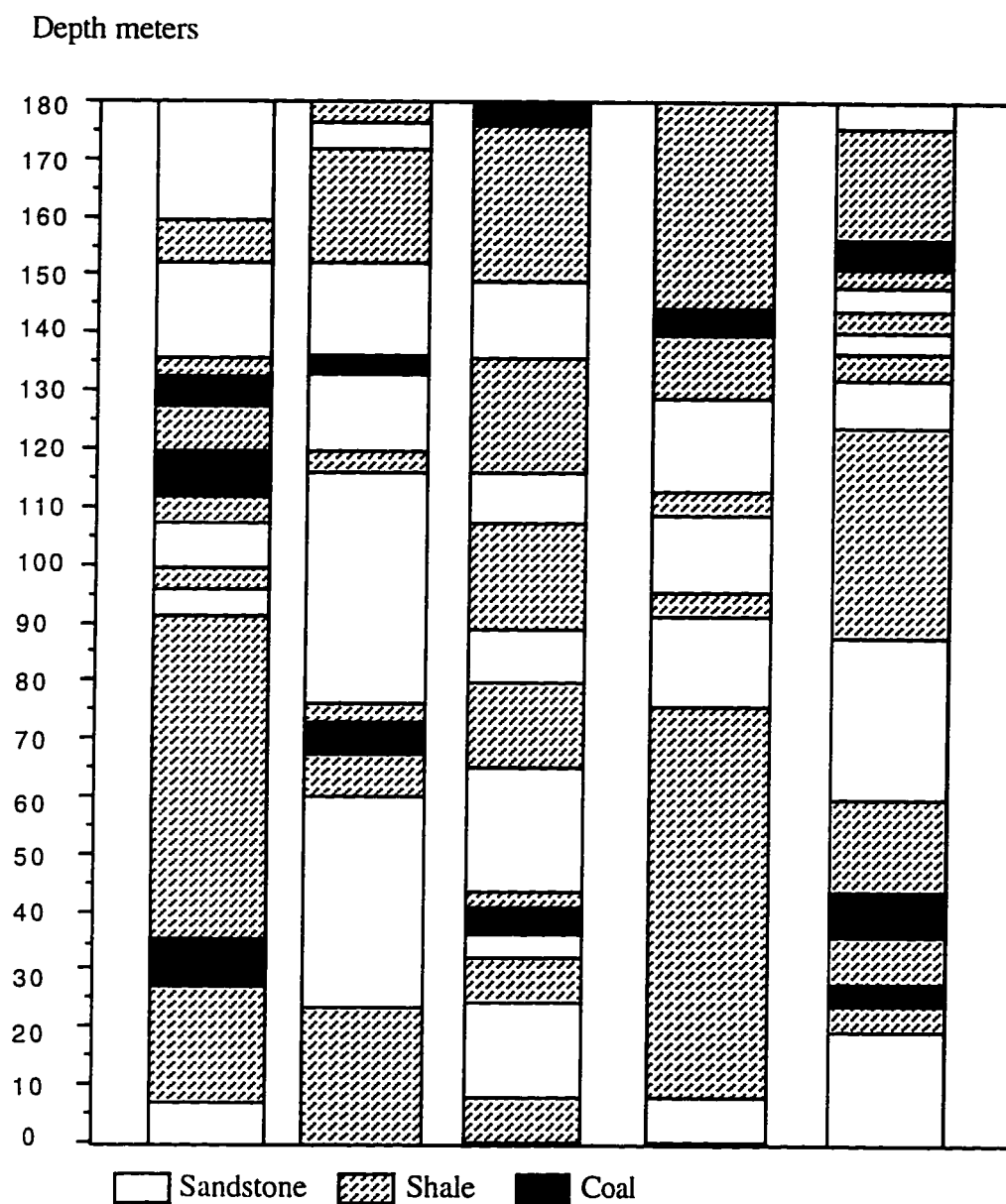


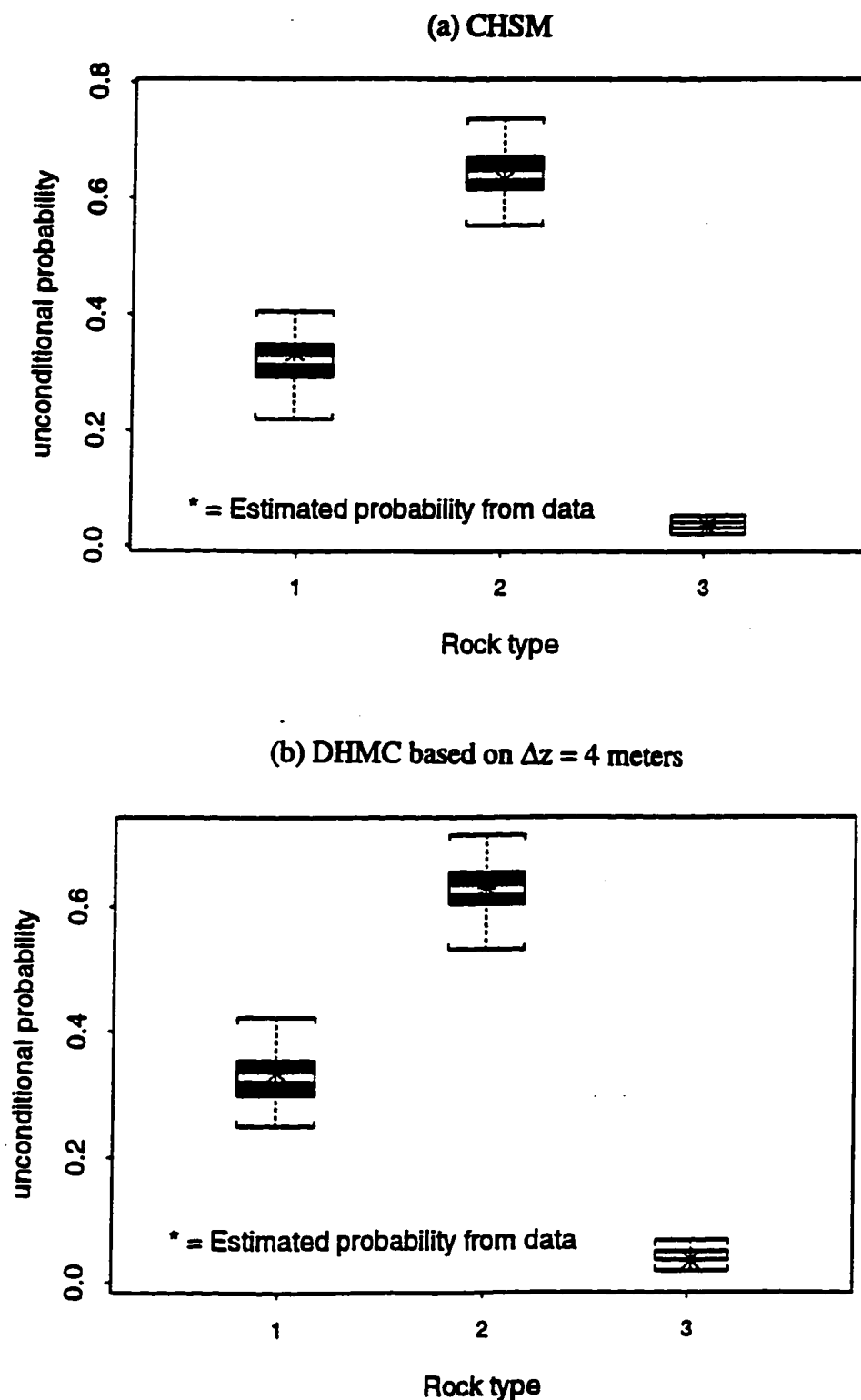
Figure 3-4. Bed thickness empirical CDF for the three rock types at the Indian site.



**Figure 3-5.** Simulated images for the Indian site generated by CHSM model.



**Figure 3-6.** Simulated images for the Indian site generated by DHMC using  $\Delta z = 4$  meters.



**Figure 3-7.** Boxplots for the estimated unconditional probabilities associated with rock type 1 (sandstone), 2 (shale), 3 (coal) at the Indian site (based on 100 realizations).

(a) CHSM

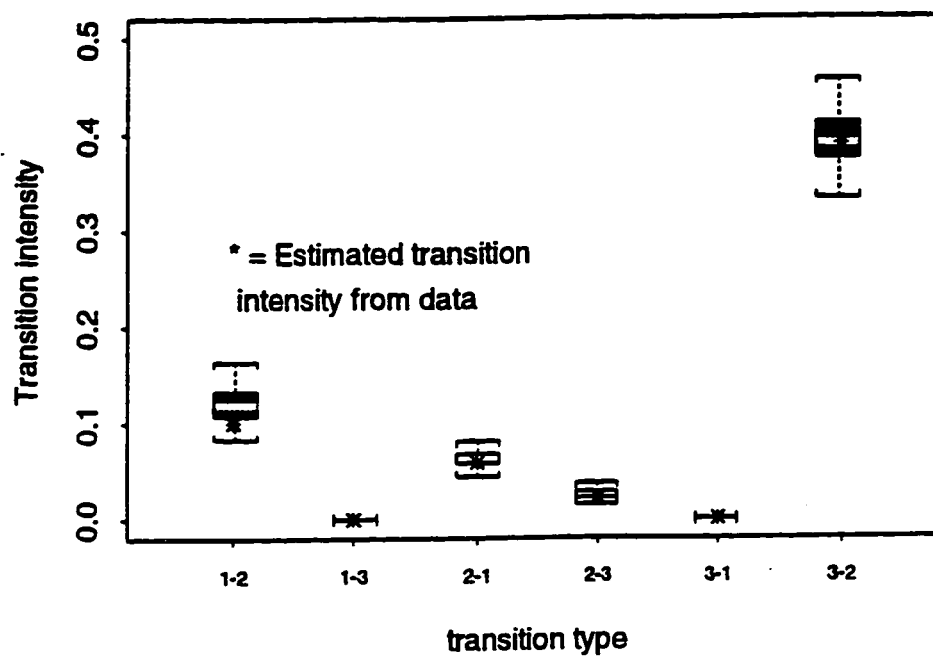
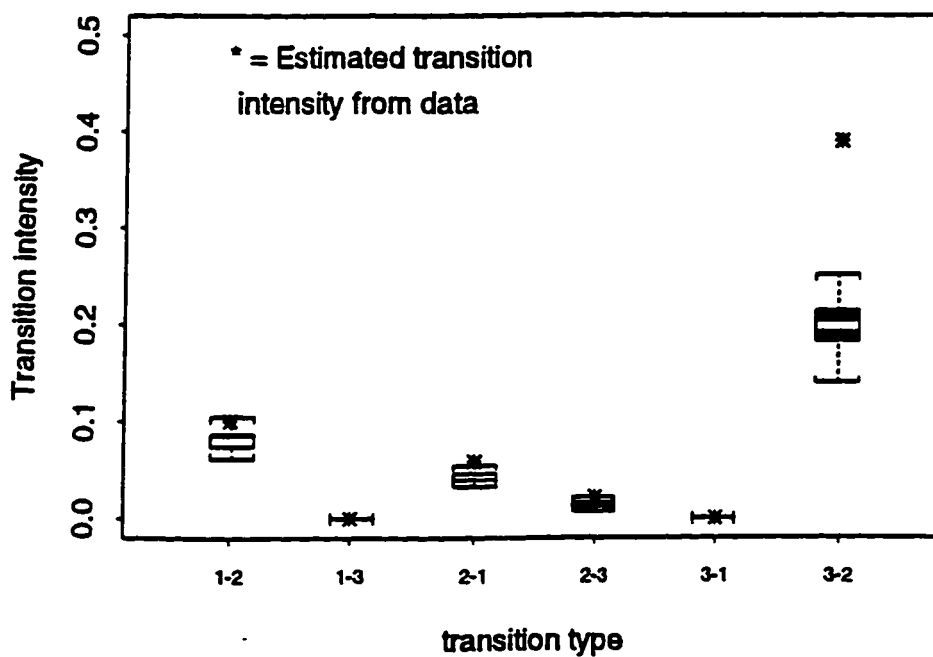
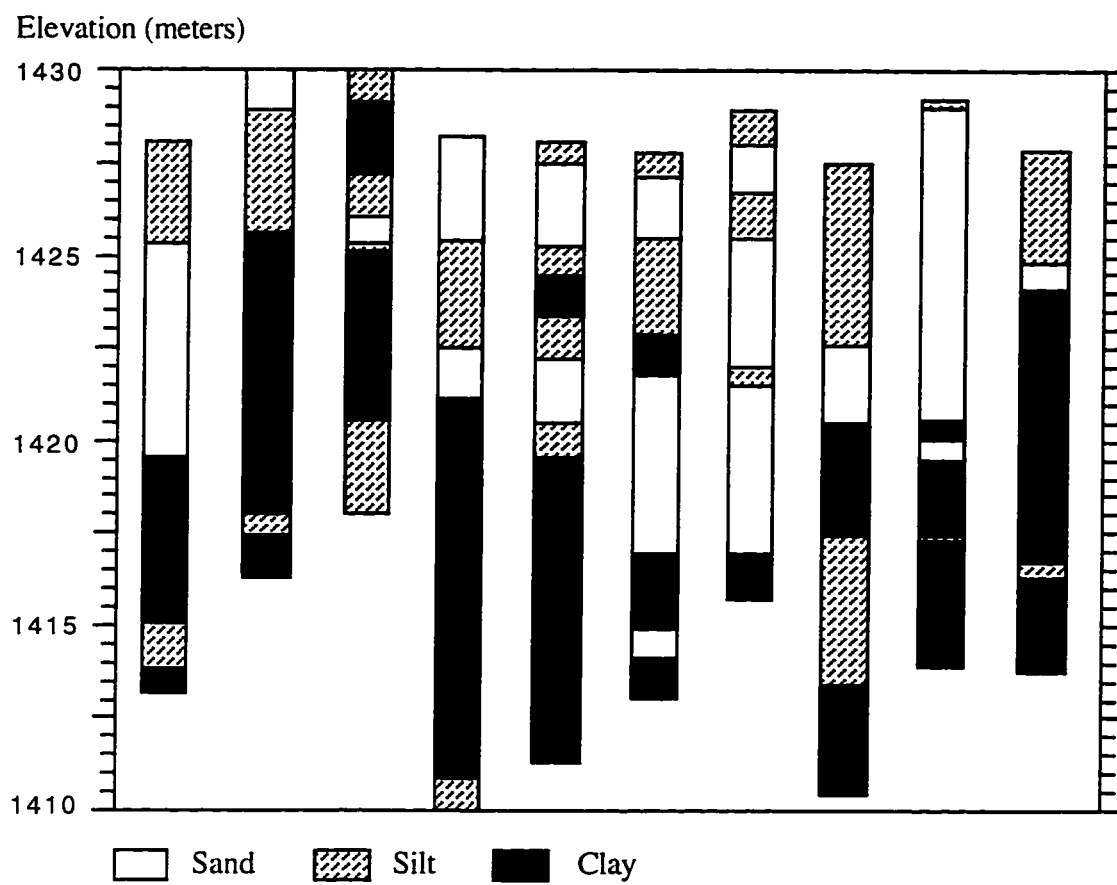
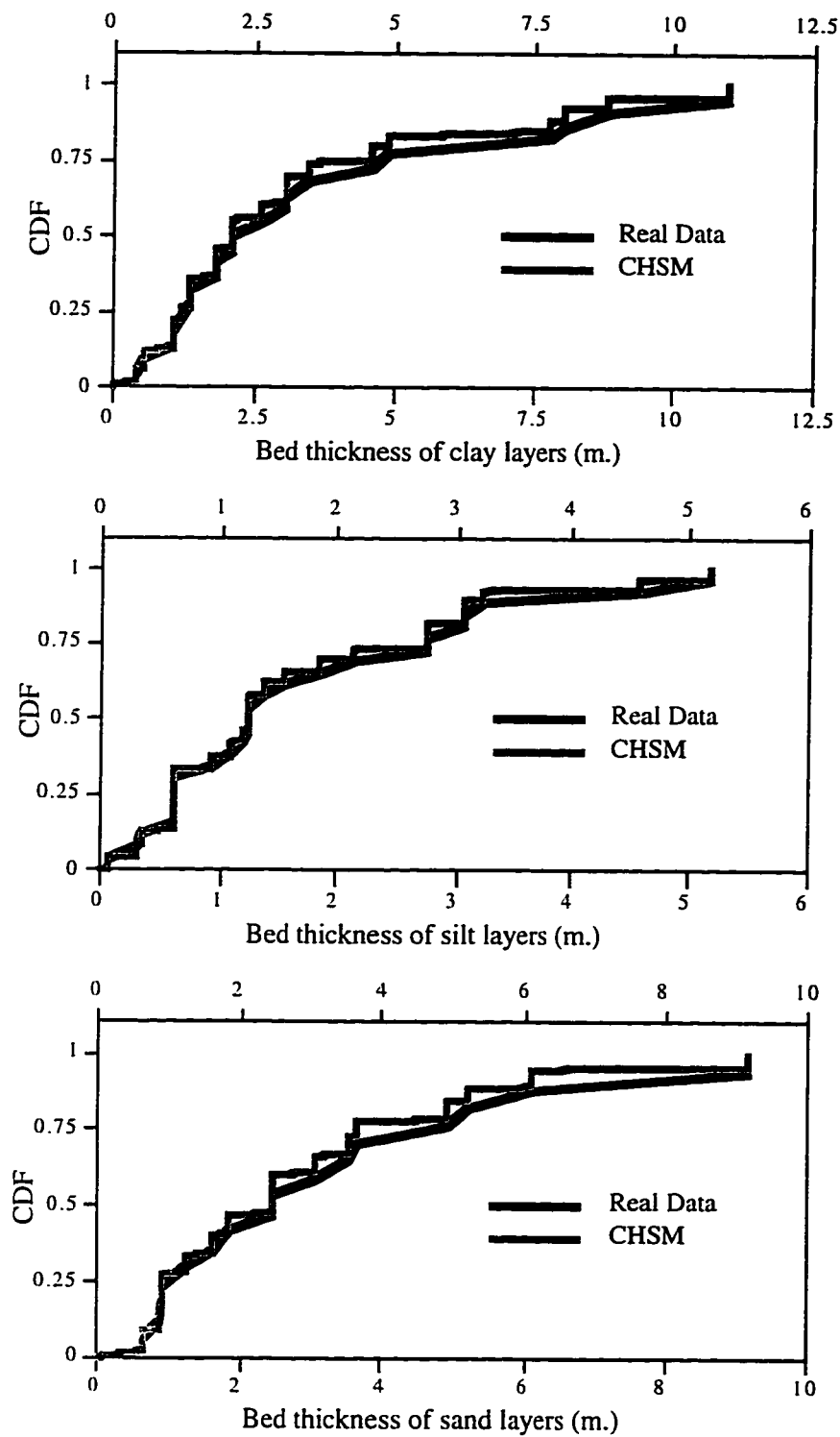
(b) DHMC based on  $\Delta z = 4$  meters

Figure 3-8. Boxplots for the estimated transition intensities from rock type  $i$  to rock type  $j$ , 1 (sandstone), 2 (shale), 3 (coal) at the Indian site (based on 100 realizations).

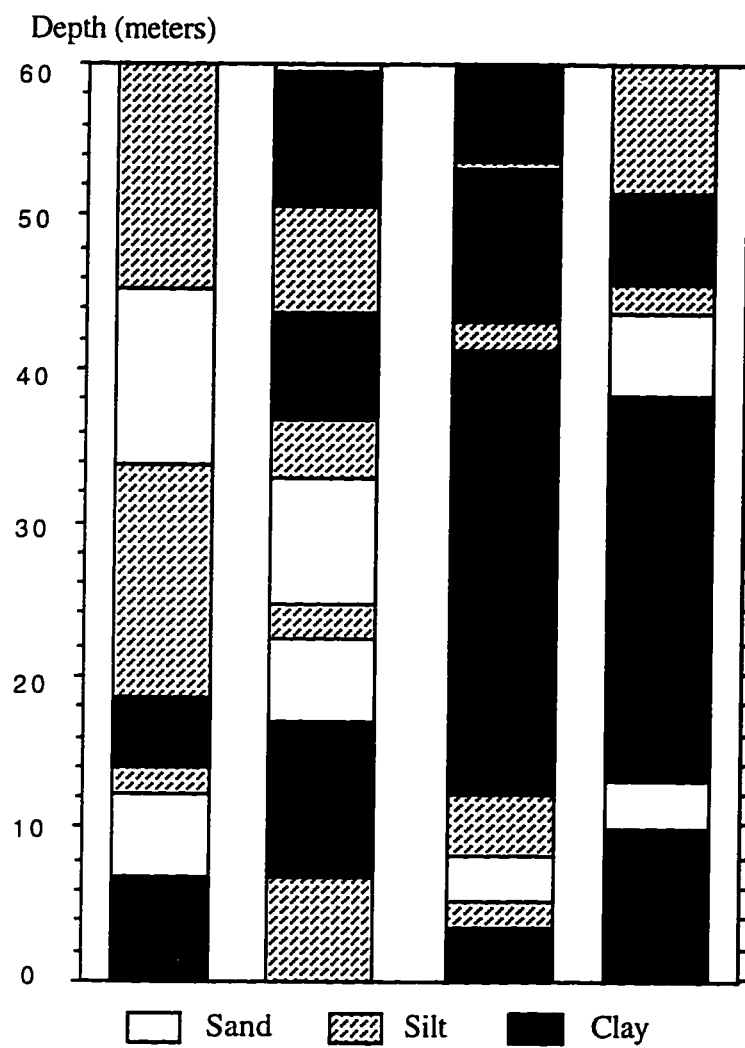


**Figure 3-9.** Bore hole data representing the Ogden Valley aquifer.

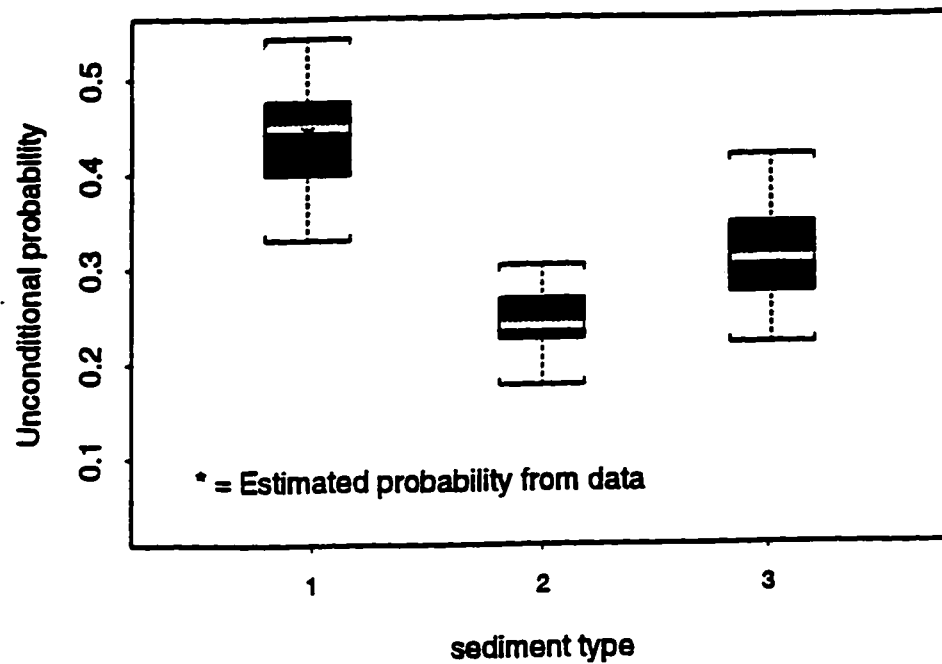


**Figure 3-10.** Bed thickness empirical CDF for the three rock types at the Ogden site.

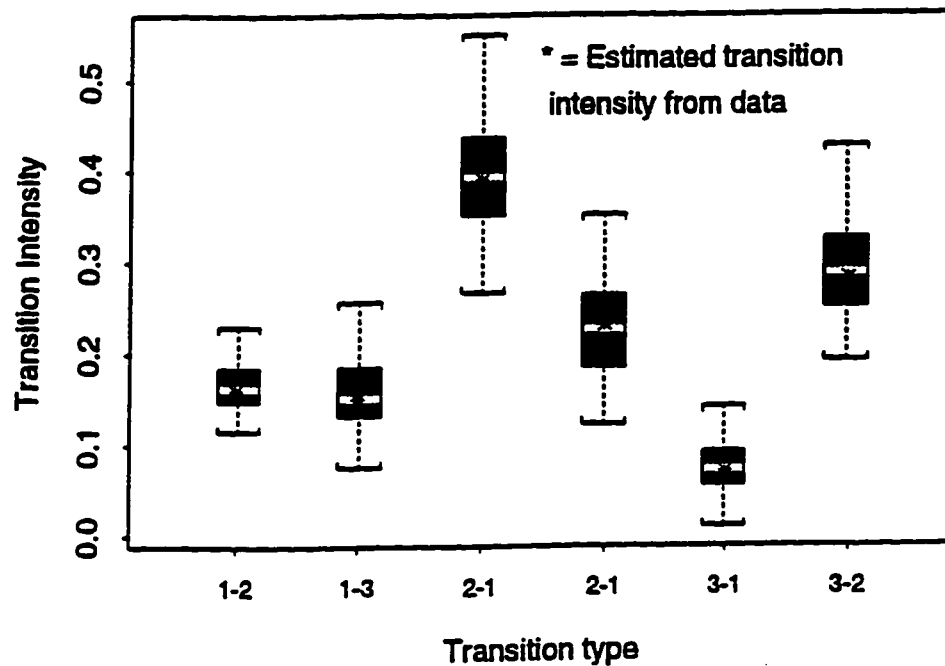




**Figure 3-11.** Simulated images for the stratigraphic sequence at Ogden Valley.



**Figure 3-12.** Boxplots for the estimated unconditional probability for sediment types 1 (clay), 2 (silt), 3 (sand) for the Ogden Valley site (based on 100 realizations).



**Figure 3-13.** Boxplots for the estimated transition intensities from sediment type  $i$  to sediment type  $j$ , 1 (clay), 2 (silt), 3 (sand) for the Ogden Valley site (based on 100 realizations)

CHAPTER 4

CONTINUOUS-PARAMETER NONHOMOGENEOUS  
SEMI-MARKOV MODEL FOR STRATIGRAPHIC  
ANALYSES FROM WELL LOG DATA<sup>1</sup>

Abstract

Well log data are a common source of information for characterizing subsurface environments. A statistical methodology is developed and applied for the interpretation of such data in terms of a multi-state depositional sequence. The well log data is classified into a discrete set of states (e.g., sand, silt, clay), and stratigraphic transitions between these states are described as a continuous- parameter, nonhomogeneous semi-Markov process. The state by state transitions at any vertical horizon are modelled using a transition intensity matrix. Transition intensity is defined as the number of transitions from one state to another per unit depth. This procedure does not require a discretization of the deposition axis as is done with a Markov chain model. The transition intensities are allowed to vary by position in the stratigraphic sequence to account for nonstationarities in the depositional process. A kernel estimator is used to estimate the transition intensity at any depth as a weighted moving average of the number of transitions from one soil type to another and of the associated bed thicknesses. A simulation strategy for simulating pseudo-well logs that have the same statistical properties as the sampled well logs is also developed. An application of the methodology to data from Lake Bonneville sediments in Utah is provided.

---

<sup>1</sup>Coauthored by Alaa Ali and Upmanu Lall.

## Introduction

Alluvial sedimentary basins are a common environment in which to find oil, water, gas, and coal. Such environments are often a consequence of large- and small-scale bed deposition processes, which may be partially random in nature. A large-scale deposition process may lead to the formation of beds that may vary in lithology and thickness. A small scale deposition process may lead to an interruption in such a bed deposition sequence creating smaller scale variations. The understanding of the lithostratigraphic sequence in such environments is important for oil traps exploration, energy extraction optimization from petroleum or geothermal reservoirs, as well as for assessing the potential for vertical contaminant migration in groundwater aquifers. A stochastic modelling approach can be useful to characterize the depositional environment, and to investigate different engineering or management alternatives through Monte Carlo simulations of likely subsurface structure. One such approach is presented here.

Well logs are often a useful source of subsurface data. This information is typically organized into distinct rock or soil types. A Markov chain (MC) analysis is often used [Anderson and Goodman, 1957; Krumbein and Dacey, 1969; Dacey and Krumbein, 1970; Harbaugh and Bonham-Carter, 1970; Bayer, 1985; Sinval and Sinval, 1992] to describe patterns of deposition and to assess the conditional probability of occurrence of different rock types through transition probability matrices. Such an approach may be used to analyze the degree of continuity and the pattern of alternation between different types of soil. Typically, MC models are applied for transitions along the depositional sequence, i.e., in the vertical. In such models, the stochastic process used to represent the depositional sequence is typically assumed to be homogeneous along the profile. Most past MC studies have discretized the profile to represent steps for transitions in the Markov chain. Some problems associated with such a discretization are reviewed by Harbaugh and

*Bonham-Carter* [1970] and *Bayer* [1985]. A semi-Markov process model that obviates the need of domain discretization was presented in Chapter 3.

The continuous-parameter, nonhomogeneous semi-Markov process model presented here considers transitions between different types of soil in the vertical along a continuum. A transition intensity is defined as the number of transitions from one state to another per unit depth. Transition intensities from one soil type to another may be evaluated at any depth horizon and are allowed to vary smoothly over the profile. Kernel methods [*Silverman*, 1986] are used to estimate the transition intensity and its variation with location in the sequence. A strategy for simulating pseudo-well logs using these estimated transition intensities as a function of location is developed. Simulation proceeds from one end of the profile to the other. An initial state is first randomly selected. A bed thickness for this soil type is then resampled from an empirical localized probability density function for bed thickness fitted at that horizon. The next soil type is selected using the transition intensity matrix evaluated at the current elevation. The process is repeated until a desired profile length has been resampled. Application to a high resolution data set from Lake Bonneville sediments in Utah is presented.

### Background

Markov chain (MC) models have been used for subsurface modelling since the 1950's. The occurrence of lithologies is viewed as a stochastic process. The lithology is modelled as a random state variable,  $X(z)$ , which takes discrete values (rock types: 1, 2, 3 ) as deposition progresses. The values of the random variable,  $X$ , are called the states of the MC, while the deposition axis,  $z$ , is called the parameter of the MC. If the random variable  $X$  is considered to occur at discrete points (sampled at a resolution of  $\Delta z$ ) along the  $z$  axis, the chain is called a discrete-parameter MC. On the other hand, if the random

variable  $X$  is allowed to occur at any point, the chain is called a continuous parameter MC.

The Markovian property is stated as:

$$P\{X(z_i) = k \mid X(z_j) = l, X(z_1) = m \dots\} = P\{X(z_i) = k \mid X(z_j) = l\} \quad (z_1 < z_j < z_i) \quad (4-1)$$

where  $z_i, z_j$  are the depths at points  $i$  and  $j$ , respectively.

This property indicates that our knowledge of the state at point  $i$  is enough to infer the state at point  $j$ . First-order dependence is expressed in equation (4-1). Models with higher order dependence can be considered.

In reality, the deposition process varies continuously with time and, hence, with depth. However, a first-order, homogeneous MC with a discrete-parameter space (DHMC) has usually been used to represent lithology.

The use of a DHMC requires a sample discretization,  $\Delta z$ , which represents a unit step in the chain. A consensus for the optimal selection of  $\Delta z$  in a given setting has not emerged [Carr *et al.*, 1966; Krumbein, 1967; Miall, 1973; Sinval and Sinval, 1992]. Once  $\Delta z$  is assumed, the transition probability matrix (TPM) is computed. For more details about the sampling discretization issue, and the TPM computation, the reader is referred to Bayer [1985] and Harbaugh and Bonham-Carter [1970]. Using the DHMC, simulation proceeds in a series of space steps  $\Delta z$  based on the estimated TPM. Successive soil types are randomly sampled in steps of  $\Delta z$  based on the TPM entries. In Chapter 3 it is shown that images generated by DHMC may not preserve the cumulative distribution function (CDF) of bed thickness very well.

### Continuous Parameter Homogeneous Markov

#### Chain and Semi-Markov Models

Given the continuous nature of the deposition process, a continuous-parameter homogeneous Markov chain (CHMC) representation may be more appropriate than the

DHMC. A transition intensity matrix (TIM) is used to describe a continuous-parameter MC and is defined as the first derivative of the transition probability with respect to time, or space [Trivedi, 1982]. In a stratigraphic sequence, an element,  $q_{km}$ , in the TIM was estimated in Chapter 3 as the ratio of the total number of state k to state m transitions to the thickness of state k observed along the entire profile of length L, and is given as:

$$q_{km} = n_{km} / L_k^* \quad (4-2)$$

where  $n_{km}$  is the number of transitions from state k to state m, and  $L_k^*$  is the total length of soil type k from which transitions to other soil types occur.  $L_k^*$  does not include the last layer in each profile from which no transition takes place.

In Chapter 3, we used the above definition and presented a continuous homogeneous semi-Markov model for stratigraphic sequence characterization and simulation that does not call for sample discretization. In this model, the stratigraphic sequence is modelled in two steps:

### Characterization

The random field is characterized using two quantities, a TIM (equation 4-2), and an unconditional probability matrix (UPM) with element  $p_k$  defined as:

$$p_k = L_k / L \quad (4-3)$$

where  $L_k$  is the total length of soil type k, and L is the total length of the stratigraphic section.

Note that the UPM and TIM calculations do not require a discretization of the vertical profile. The quantity  $p_k$  is a measure of lithologic relative frequency (expected value). The quantity  $q_{km}$  is a measure of both the lithologic transition and thickness (continuity).

## Simulation

The goal here is to simulate the random field without discretization such that the TIM, the UPM, and some statistics, e.g., mean and standard deviation, of the bed thickness of each soil type, are preserved. The initial state is first decided randomly according to an Initial Probability Matrix (IPM) defined as:

$$\pi_k = \frac{n_k}{N} \quad (4-4)$$

where  $n_k$ , is the number of well logs starting with soil  $k$ , and  $N$  is the total number of well logs. An off-diagonal transition probability matrix (OTPM) describes the transitions between successive soil types. An element,  $p'_{km}$ , in the OTPM is defined, independently of  $\Delta z$ , in terms of the TIM elements as:

$$p'_{km} = \frac{q_{km}}{M_k}, \quad m=1..n; \quad k \neq m; \quad p'_{kk}=0, \quad M_k = \sum_{j=1 \neq k}^n q_{jk}. \quad (4-5)$$

Instead of considering transitions at discrete distance markers, one considers transitions out of a state to other states (excluding the originating state), after a certain thickness,  $z_j$ , has been randomly deposited in a state  $j$ , following a probability distribution  $f(z_j)$ .

The simulation of a stratigraphic sequence by such a model is similar to that used by *Krumbein and Dacey* [1969] and by *Potter and Blakely* [1967]. Layer thicknesses for a particular soil type are resampled from observed thicknesses for that type of soil, and transitions to other states are modelled by the OTPM. The procedures used for simulating a stratigraphic sequence upwards from the lowest horizon considered are shown in the flow chart in Figure 4-1. In Chapter 3, we show that such a model preserves the UPM and TIM, the relative lithologic frequency, lithologic sequence (regardless of spatial location), and bed thickness, for a stationary stochastic process.

In a nonstationary environment, the transition probabilities or intensities vary with location, and a single lumped UPM and TIM may not be enough to characterize such an



environment. Simulations based on existing Markov models may not preserve the local variation of a UPM and/or TIM in a nonstationary field. Many sites exhibit nonstationarity in the vertical and/or in the horizontal [Miall, 1973; Sinvhal and Sinvhal, 1992]. In such an environment, the site is partitioned into smaller areas where the process is assumed stationary and a DHMC model is applied at each such subsection. To test for nonstationarity, Harbaugh and Bonham-Carter [1970] divided a long stratigraphic sequence into a number of fixed subintervals and computed the transition probability matrix (TPM) for each subinterval. If such TPMs are not the same, the stratigraphic sequence is said to be nonstationary. Also, Anderson and Goodman [1957] presented a statistical test for stationarity. However, a practical procedure for efficiently modelling such nonstationary deposition has not emerged. A goal of this chapter is to develop a continuous nonhomogeneous semi-Markov model for a nonstationary subsurface stratigraphy that recognizes the possibly continuous nature of the underlying nonstationary deposition process and does not suffer from the problems associated with sample discretization. We extend the model presented in Chapter 3 to achieve such a goal.

## Methodology

### Introduction

A nonstationary stratigraphic sequence is first characterized using a moving window or kernel estimation procedure. The UPM and TIM are defined at each point along the profile (i.e., on a continuum) to recognize any local variation of the process. The simulation then proceeds in a similar manner to the procedure for the homogeneous case presented in the previous section. The two steps represent a continuous nonhomogeneous semi-Markov model (CNSM) and are presented below. It is presumed that a number of well logs are available, the environment is homogeneous in the horizontal direction, and that vertical

non-stationarities are of interest. Thus each well log provides a sample of the vertical deposition process, and averaging across well logs is meaningful.

### **Characterization of a Continuous Nonstationary Stratigraphic Sequence**

In a vertically nonstationary environment, local estimates of the TIM elements  $q_{km}(z)$  and UPM elements  $p_k(z)$  varying with  $z$  are desired. One way to localize an estimate at a given point  $z$  is to localize the associated observations to a certain neighborhood of this point. This can be achieved using a "moving window," with a specified length along the stratigraphic sequence, centered at point  $z$ , and computing the UPM and TIM at its center point  $z$  (see Figure 4-2). The window can be selected such that observations closer to its center will have more weight in the estimation than farther observations. The kernel method presented below provides such an estimate.

### **Kernel Estimator**

A nonparametric approach, based on kernel estimation, is developed for local estimates of transition intensity. Kernel estimators are popular for probability density estimation and regression where the researcher does not wish to make prior assumptions as to the functional form of the underlying behavior, desires the estimates to be "local" rather than global and allows the data to have a larger role in the estimation process than it would if used simply through some overall sample moments. Kernel methods have been used to describe the rate of point processes, such as nonhomogeneous Poisson processes [Diggle, 1985; Diggle and Marron, 1988; Solow, 1990]. In Chapter 2, we used the kernel estimator to estimate the unconditional probability of two soil types as a function of location  $(x,y,z)$  from well log data. The reader is referred to Härdle [1989], Silverman [1986], and Scott [1992] for accessible monographs on kernel estimation and to Lall [1995] for a review of hydrologic applications. The kernel estimator is a weighted moving average of a

target function, where a preselected weight function is prescribed and the weighted moving average is taken over an appropriately determined span or bandwidth (Figure 4-2).

At an elevation  $z$ ,  $m_z$  well logs are available. Not all drill logs may be present at some elevations due to variations in length and topographic relief. An averaging interval,  $2h$ , is considered at each drill log  $j$  by identifying soil layers that lie within distance  $h$  from a horizontal level  $z$ . The unconditional probability of soil type  $k$ ,  $p_k$ , at level  $z$  is estimated as:

$$p_k(z) = \frac{1}{h} \sum_{j=1}^{m_z} \int_{z-h}^{z+h} I_{k,j}(\zeta) * K\left(\frac{\zeta-z}{h}\right) d\zeta \quad (4-6)$$

and the transition intensity  $q_{kl}$  at elevation  $z$  is estimated as:

$$q_{kl}(z) = \frac{\sum_{j=1}^{m_z} \sum_{i=1}^{n_j-1} K\left(\frac{z_i-z}{h}\right) * I_{kl}(z_{ij})}{\sum_{j=1}^{m_z} \int_{z-h}^{z+h} I_{k,j}(\zeta) * K\left(\frac{\zeta-z}{h}\right) d\zeta} \quad (4-7)$$

where:

$m_z$  = Number of well logs observed at  $z$

$n_j$  = Number of soil layers at well log  $j$ .

$z$  = Point of estimate.

$z_{ij}$  = Top elevation of soil layer  $i$  at well log  $j$ .

$I_{k,j}(\zeta)$  = Indicator function of soil  $k$  at elevation  $\zeta$  and drill log  $j$ , it is 1 if soil at elevation  $\zeta$  is of type  $k$  and 0 otherwise.

$I_{kl}(z_{ij})$  = Indicator function of transition from soil  $k$  to soil  $l$  at elevation  $z_i$  and drill log  $j$ .

$$K(.) = \frac{15}{16}(1-u^2)^2 \text{ if } |u| \leq 1, 0, \text{ and otherwise, where } u = \frac{z_{ij}-z}{h}, \text{ or } \frac{\zeta-z}{h}$$

$h$  = A bandwidth or averaging interval in the vertical.

The function  $K(.)$  is called a kernel or weight function. It integrates to 1 over  $[z-h, z+h]$ ; hence, the  $p_k$  is a weighted average of the relative frequency of soil of type  $k$  over a moving window centered at elevation  $z$ . The estimator for  $q_{kl}$  is similarly the ratio of a weighted average of the number of  $k$  to  $l$  transitions to a weighted length of soil type  $k$ . We illustrate the estimation of the probability of clay,  $p_c$ , and the transition intensity from clay to sand,  $q_{cs}$ , through a numerical example, using 1 well log, in Figure 4-2. The four shaded areas under the kernel are the weights corresponding to the layers of soil type  $c$ . The three arrows point at the transitions of type  $cs$ . The weight function serves to localize the estimate by giving higher weight to observations near the estimation point than for observations that are farther. The estimate  $p_c$  is obtained as the weighted fraction of the clay thickness as shown. The estimate  $q_{cs}$  is formed by dividing the three weighted transitions  $cs$  by the weighted clay thickness within a bandwidth of the point of estimate.

For a homogeneous situation where the target function is constant, the best estimate will be obtained by taking the whole domain as the averaging neighborhood. On the other hand, if the target function varies rapidly, smaller neighborhoods are desirable. However, if a very small neighborhood is chosen, variance of the estimate increases since very few transitions and layers are encountered. The bandwidth can be selected by minimizing the mean square error (MSE) of some target function. In this study, such a target function involves several transition intensity and probability variables where a simultaneous solution for a single optimal bandwidth under the MSE criterion may not be meaningful. For the purpose of this work, we decided to use the distance to the  $\kappa^{\text{th}}$  nearest transition as an ad hoc method for bandwidth selection. A bandwidth  $h(z)$ , that varies with depth,  $z$ , is chosen

equal to,  $\text{avg\_r}_\kappa$ , the average of the distances to the  $\kappa^{\text{th}}$  nearest transition boundary from the level of estimate at each drill log. Given  $\tau_j$  transitions observed at each drill log  $j$ , the value of  $\kappa$  is taken equal to  $\sqrt{\tau_j}$ . This method leads an adaptive choice of the bandwidth where the bandwidth expands if few observations are found and shrinks if many observations are found. The reader is referred to *Silverman* [1986] and *Scott* [1992] for other methods of bandwidth selection.

### **Simulation of Pseudo-Well Logs**

A strategy for generating pseudo-well logs using a CNSM is presented. The proposed simulation strategy is a generalization to the simulation procedures presented in Chapter 3 and is consistent with the symbolic flow chart presented earlier (Figure 4-1). The simulation of a stratigraphic sequence proceeds upwards from the lowest horizon considered. The TIM is first estimated as a function of  $z$  by estimating each element  $q_{kl}(z)$ ,  $k \neq l$ . The initial state is decided randomly according to an initial probability matrix (IPM). Layer thicknesses for a particular soil type are then generated using an appropriate local probability distribution for bed thickness. The transition to the next state is determined using an off-diagonal transition probability matrix (OTPM), which is derived from TIM. The simulation procedures are similar to those presented in the flow chart in Figure 4-1 except that the all parameters vary with  $z$ . The computation of the IPM is given in equation 4-4, and those of the bed thickness and the OTPM are presented below.

#### **Bed Thickness**

Bed thickness for a soil type  $k$  may be simply resampled from observed thicknesses of that type of soil [*Ali and Lall*, unpublished report, 1996; *Potter and Blakely*, 1967]. Also, in a homogeneous semi-Markov model, bed thickness may be assumed to follow exponential, gamma, or lognormal, distributions [*Schwarzacher*, 1975]. In a

nonstationary environment, if a bed thickness distribution is assumed, a local fitting of the bed thickness is possible by localizing the parameter of the assumed distribution. In the application presented later, the bed thickness of soil  $k$  is sampled using two methods:

*Resampling of windowed observed data.* A local bootstrap [Efron, 1982] procedure can be used to resample bed thickness for soil type  $k$  using the data from the  $m_z$  well logs available at depth  $z$ . All layers of soil type  $k$  that lie within one bandwidth,  $h(z)$ , of  $z$  are first identified. Let us say that there are  $J_k(z)$  such layers, and that the distance of their bottom elevations,  $e_{k,i}(z)$ , from the current level  $z$  is  $r_{k,i}(z)$ . Now arrange the layer indices  $i=1, \dots, J_k(z)$  such that the  $r_{k,i}(z)$  are in increasing order. In other words, record the layer indices in order of distance of their bottom elevations from current horizons. Record the corresponding layer thickness as  $d_{k,i}(z)$ . Now a layer  $i$ , with thickness  $d_{k,i}$ , is resampled with a probability proportional to its rank  $i$ :

$$F(i \leq I) = \frac{\sum_{i=1}^{J_k(z)} \frac{1}{i}}{\sum_{j=1}^{J_k(z)} \frac{1}{j}} \quad (4-8)$$

This metric is based on a locally Poisson approximation to the underlying spatial field and is introduced and discussed by *Lall and Sharma* (1996). The net effect is that the closest layer is chosen twice as often as the next closest layer, three times as often as the next, and so on. This strategy honors the observed data, but does not allow the generation of any bed thickness that were not observed.

*Resampling from a local exponential distribution.* Parametric distributions for the bed thickness,  $d_k$ , may be more attractive where there is only limited data. Here, the bed thickness of soil type  $k$  is assumed to follow an exponential distribution with a local mean parameter that reflects the average bed thickness of soil  $k$  as function of  $z$ . A nonstationary bed thickness distribution for soil type  $k$  is defined at each depth  $z$  as:

$$f(z, d_k(z)) = M_k(z) e^{-M_k(z) d_k(z)} \quad (4-9)$$

where:

$$M_k(z) = \sum_{i=1 \neq k}^n q_{ki}(z) = \frac{1}{\text{Average bed thickness for beds of type } k \text{ at depth } z} \quad (4-10)$$

Recall that  $q_{ki}(z)$  is estimated from equation 4-3. A random bed thickness for soil type  $k$  at depth  $z$  may then be simulated as:

$$d_k(z) = \frac{\log_e(u)}{M_k(z)} \quad (4-11)$$

where:

$u$ : Uniform random variate between (0 and 1)

For  $u$  values close to 0 or close to 1, the bed thickness may be infinitesimally small or extremely thick. For practical purposes, the range of  $d_k(u)$  is truncated to be between  $d_1$  and  $d_2$  where  $d_1$  and  $d_2$  correspond to the smallest and largest bed thickness considered admissible respectively.

#### Off-Diagonal Transition Probability Matrix (OTPM)

A zero-diagonal TPM is used to determine transitions out of a soil type to other soil types at depth  $z$ . An element  $p'_{km}(z)$  is computed as:

$$p'_{km}(z) = \frac{q_{km}(z)}{\sum_{l=1 \neq k}^n q_{kl}(z)} \quad m=1..n; \neq k; p'_{kk}=0 \quad (4-12)$$

Equation 4-11 is interpreted by observing that the probability of transition out of a state at the end of the deposition episode is seen to be proportional to its transition intensity out of the state. Given a current state  $k$ , the next state  $m$  is determined using the transition probabilities  $p'_{km}(z)$  to the  $(n-1)$  states excluding  $k$ , as estimated by equation 4-11.

### Application

The Ogden Valley has an aquifer system that is typical of Lake Bonneville sediments that cover large portions of the state of Utah. The site under consideration is located just west of the Wasatch Mountain Range on the relict Weber Delta. The delta consists of broad plains and terrace, and originates on the western base of the Wasatch Range. Topographically, this site is on a plateau formed by the Weber Delta. The plateau is approximately 90 m above the valley floor. Surface elevations at this site vary from 1400 m above mean sea level along the western side, to 1540 m near the eastern side. Depth to bedrock in the basin ranges from 460 m on the western side to 2300 m on the east. The available data lie in the upper 20 m of unconsolidated geologic material that ranges from very low permeability soil (clay) to a very high permeability soil (sand and gravel). Soils observed along well log profiles had been classified into 14 types. We considered three major classes of soils: 1) low permeability layers (clay), 2) high permeability layers (sand and gravel), and (3) mixed permeability layers (silty sand, silty gravel). The data used for the application were from 10 well logs (see Figure 3-9). The vertical resolution of this data is 1 centimeter. From Figure 3-9, we notice that clay layers tend to be near the bottom, while silt and sand layers tend to be near the top, indicating a possible nonstationarity in the deposition process at least within the observed profile. The CNSM model was applied twice to simulate these data. First (application 1), the bed thickness was resampled from the observed data. In the second application, the bed thickness was assumed to follow an exponential distribution. For each application, 90 realizations were generated where each realization consisted of 10 pseudo logs (see Figures 4-3 and 4-4). For each realization, the following statistics are calculated for each soil type: mean and standard deviation of bed thickness, profile-averaged UPM and TIM, and UPM and TIM as function of  $z$ . The UPM



and TIM for the simulations are computed in each case using the same methods as for the original well log data. Boxplots showing the mean value and the tenth and the ninetieth percentiles from the simulations are presented in each case.

Comparative results for simulations from the two applications of the CNSM model are presented in this section. "Bulk" properties for the full profile are compared first, followed by an interpretation of the nonstationary nature of the profile.

A visual comparison of one realization of 10 pseudo-well logs selected at random from the 90 generated for each application is presented in Figures 4-3 and 4-4. Comparing with the 10 observed well logs in Figure 3-9, we observe that the pseudo well logs appear plausible with a higher tendency for sand near the top and clay near the bottom of the profile as was the case with the real data. In the realizations selected here, clay is present in all the well logs, whereas one of the real well logs did not have any clay. Hereafter, we index the soil types as (1) clay, (2) silt, and (3) sand.

From a geologic standpoint, it is of interest to see whether the first two moments of the bed thickness for each soil type over the profile are well preserved. These statistics are presented in Figure 4-5 for the two applications. Each boxplot shows the average value from the simulations, and the 10th and 90th percentiles of the statistic of interest as the edges of the box. The value of the statistic for the 10 real well logs is also shown. For application 1, the results (hatched boxplots in Figure 4-5) for the mean and standard deviation of the bed thickness (empirical resampling) are consistent with the statistics for the real data. In application 2, the average bed thickness (Figure 4-5) simulated for clay and sand using the exponential distribution appears reasonable, while the average bed thickness simulated for silt is too low. For all three types of soil, the standard deviation (Figure 4-5) of bed thickness from the 90 simulations is biased lower relative to the statistic from the original sample. The exponential distribution is specified fully by the mean value of the

process. Consequently, the simulations suggest that the exponential distribution may not be a good model for the bed thicknesses for this data set.

The above observations are reinforced when one examines the unconditional probabilities for each soil type and each transition intensity averaged over the profile. The comparative results from the two applications are presented in Figure 4-6. The results for the unconditional probabilities ( $p_1$ ,  $p_2$ ,  $p_3$ ) for each type of soil are quite similar for the two applications. This is consistent with the results for reproducing the mean bed thickness. However, the bias and the variability of the transition intensities ( $q_{12}$ ,  $q_{13}$ ,  $q_{21}$ ,  $q_{23}$ ,  $q_{31}$ ,  $q_{32}$ ) for the simulations from application 1 (resampling real bed thicknesses) is considerably smaller relative to those using the exponential distribution. This is consistent with the bias in the standard deviations of bed thickness observed earlier and indicates that the bed continuity and transition frequencies are incorrectly preserved when the exponential distribution is assumed.

Further insights into the differences between the two applications are possible by examining how the unconditional probabilities and transition intensities vary with depth. These results are provided in Figures 4-7 and 4-8. In each figure a solid line is used to show the variation of the statistic of interest with depth as estimated from the 10 real well logs, a dotted line is used to show the average of the same statistic across the 90 realizations of 10 well logs each, and dashed lines show the 10<sup>th</sup> and 90<sup>th</sup> percentiles from the simulations. The averages and the percentiles are computed for each point along the profile across the simulations. The 10 real well logs show that the estimated unconditional probability of clay decreases with elevation (above mean sea level); and the unconditional probabilities of silt and sand increase with elevation. These trends are reproduced by both sets of simulations with lower bias apparent for application 1 (empirical resampling). Once again, the major differences between the two methods are in how the transition intensities

are reproduced. Here, the performance of the application 1 procedure is dramatically better than when the exponential distribution is used for describing bed thickness (application 2). The bias and variance of the simulated  $q_{kl}$  values in application 2 is very large wherever the corresponding  $p_k$  is small. See for example the results for  $q_{31}$  near elevation 1418, and  $q_{12}$  or  $q_{13}$  near elevation 1430, in Figure 4- 8. It is not totally clear whether this is an edge effect (1418 and 1430 are the endpoints of the sample) or it is associated with limited effective sample size for fitting the exponential distribution in an area where  $p_k$  is low. Since this problem is not observed with application 1 (Figure 4- 7), one suspects that it is due to the choice of the exponential distribution to represent bed thicknesses in this application.

### Summary

A stochastic model for characterizing subsurface soils and for simulating pseudo well logs from well log data was presented here. The key contributions of this work are:

(1) A procedure for computing the unconditional probabilities and transition intensities for multiple types of soils from well log data is provided. This procedure considers nonstationarity of the deposition process in the vertical. This is a new development for a potentially difficult problem. No a priori discretization of the domain for model definition is necessary. Kernel estimators with a variable bandwidth (distance to  $k^{\text{th}}$  nearest transition) are employed.

(2) A semi-Markov framework is used to generate the pseudo-well logs. The parameters of this process are allowed to vary smoothly along the profile in the vertical. Two procedures for resampling the thickness of the beds for each soil type were investigated. For the application presented, resampling from the observed bed thicknesses that lie within the window used for estimating local probabilities and transition intensities

worked better than assuming the exponential distribution. We recommend this as a robust and intuitively pleasing choice that avoids the issue of specification of a parametric distribution, and honors the observations.

(3) The methods presented are nonparametric, i.e., no prior assumptions are made as to the underlying probability distribution (except the assumption of the local exponential distribution for the bed thickness in method. They do have parameters, and the assumption that a semi-Markov process is appropriate for describing the deposition process. However, the resulting procedure is still quite adaptable and should work well for a variety of situations.

Extensions to consider nonstationarity in the horizontal in addition to the vertical are possible. The real limitation here is the availability of a sufficiently large number of well logs at the site. Further research into improved methods for bandwidth selection for the kernel estimator used is also needed. Our investigations show that the choice used here works reasonably well and is not too sensitive to small perturbations in specification. It has the advantage of adapting the degree of averaging or smoothing to the neighborhood, rather than prescribing a fixed window size. However, investigation of its mean square error efficiency across the statistics of interest, and improvements that may result from such an analysis need to be performed.

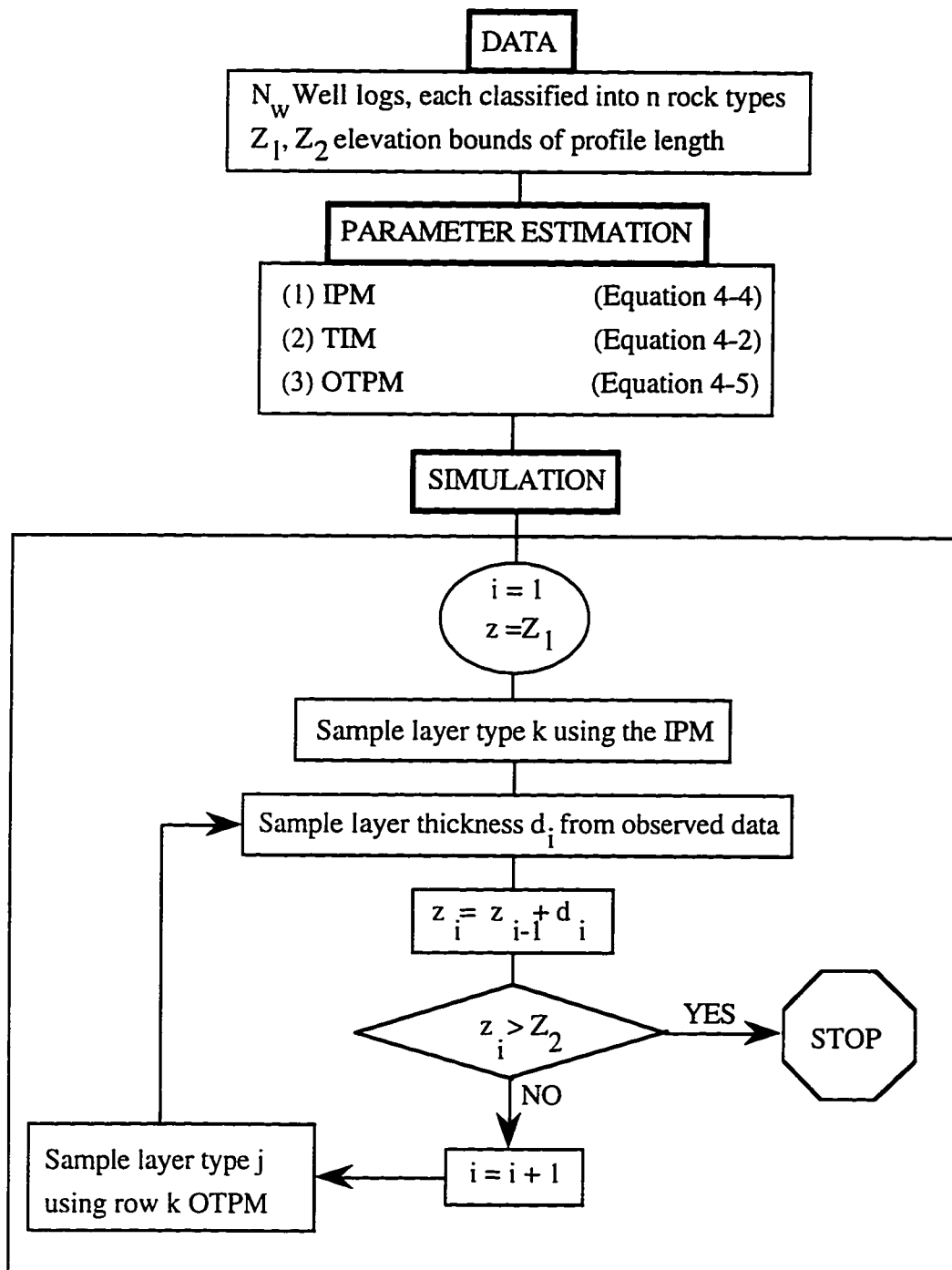
## References

- Anderson, T. W., and L. A. Goodman, Statistical inference about Markov Chains, *Annals of Mathematical Statistics*, 28, 89-110, 1957.
- Bayer, U., *Pattern Recognition Problems in Geology and Paleontology*, 229 pp., Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1985.
- Carr, D. D., A. Horowitz, S. V. Harbar, K. F. Ridge, R. Rooney, W. T. Straw, W. Webb, and P. E. Potter, Stratigraphic sections, bedding sequences, and random processes, *Science*, 154, 1162-1164, 1966.

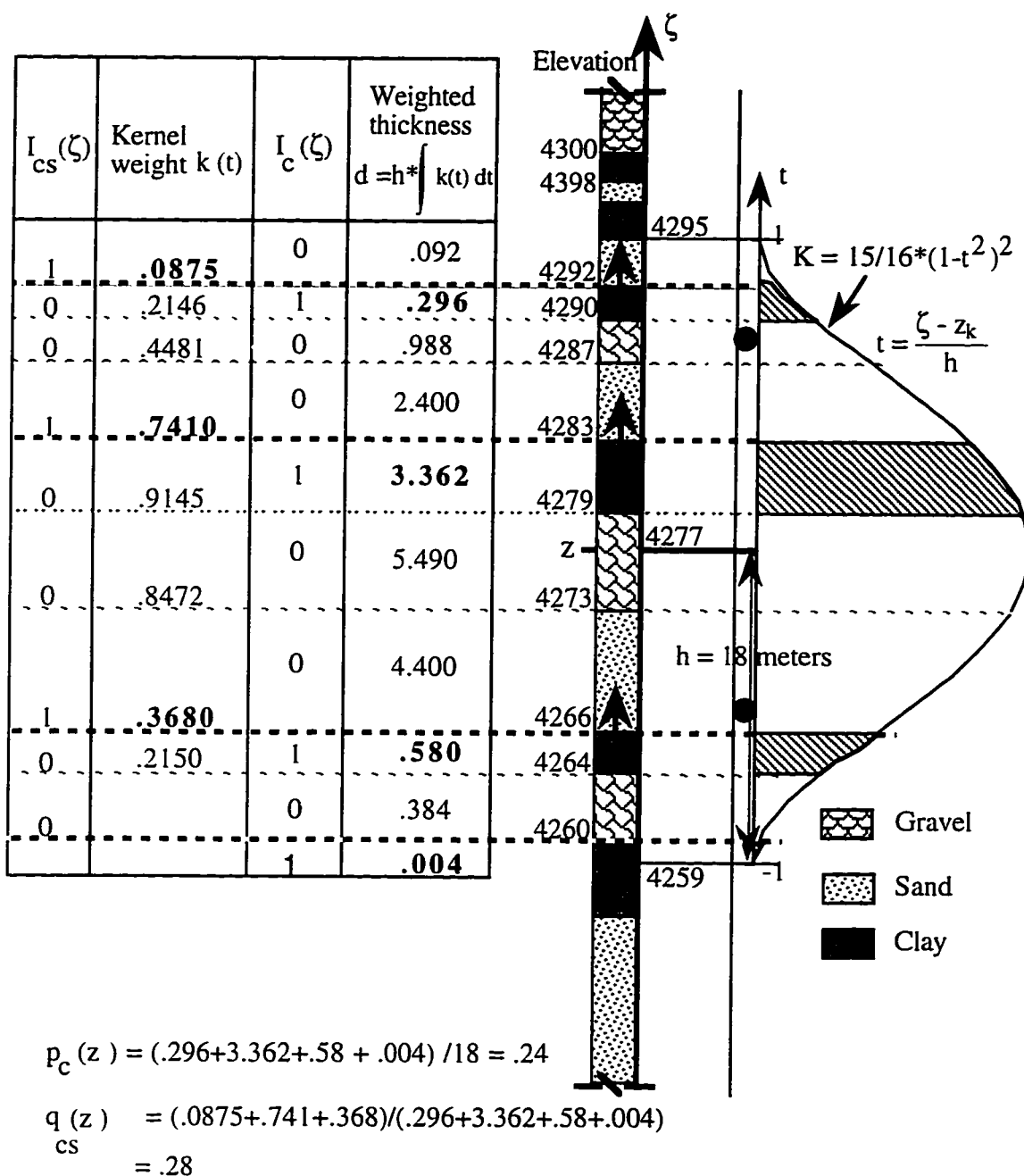
- Dacey, M.F., and W. C., Krumbein, Markovian models in stratigraphy, *J. Int. Assoc. Math. Geol.*, 2, 175-191, 1970.
- Diggle, P., A kernel method for smoothing point process data, *Appl. Stat.*, 34(2), 138-147, 1985.
- Diggle, P., and J. S. Marron, Equivalence of smoothing parameter selectors in density and intensity estimation, *J. Amer. Stat. Assoc.*, 83(403), 793-800, 1988.
- Efron, B., *The Jackknife, the Bootstrap, and Other Resampling Plans*, 92 pp., Soc. Indust. Appl. Math., Philadelphia, Pa., 1982.
- Harbaugh, J. W., and G. Bonham-Carter, *Computer Simulation in Geology*, 575 pp., Wiley-Interscience, New York, 1970.
- Härdle, W., *Applied Nonparametric Regression*, 333 pp., Cambridge University Press, Cambridge, Mass., 1989.
- Krumbein, W. C., FORTRAN IV computer programs for Markov chain experiments in geology, *Computer Contribution 13*, 38 pp., Geological Surveys of Kansas, 1967.
- Krumbein, W. C., and M. F., Dacey, Markov chains and embedded Markov chains in geology, *J. Int. Assoc. Math. Geol.*, 1, 79-96, 1969.
- Lall, U., Nonparametric function estimation: Recent hydrologic applications, *Reviews of Geophysics, US National Report 1991-1994*, 1093-1102, 1995.
- Lall, U., and A. Sharma, A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resour. Res.*, 32( 3), 679-690, 1996.
- Miall, A. D., Markov chain analysis applied to an ancient alluvial plain succession, *Sedimentology*, 20, 347-364, 1973.
- Potter, P. E., and R. F., Blakely, Generation of a synthetic vertical profile of a fluvial sandstone body, *J. Soc. Pet. Eng.*, 243-251, 1967.
- Schwarzacher, W., *Sedimentation Models and Quantitative Stratigraphy*, Development in sedimentology 19, 382 pp., Elsevier Scientific Publishing Company, Amsterdam, 1975.
- Scott, D. W., *Multivariate Density Estimation, Theory, Practice, and Visualization*, 317 pp., John Wiley and Sons, New York, 1992.
- Silverman, B. W., *Density estimation for Statistics and Data Analysis*, 175 pp., Chapman & Hall, New York, 1986.
- Sinvhal A., and H. Sinvhal, *Seismic Modelling and Pattern Recognition in Oil Exploration*, 178 pp., Kluwer Academic Publishers, Dordrecht, 1992.

Solow, A. R., The nonparametric analysis of point process data, The freezing history of Lake Konstanz, *J. Climate*, 4(1), 116-119, 1990.

Trivedi, S. K., *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*, 624 pp., Prentice-Hall, Inc., Englewood Cliffs, NJ, 1982.

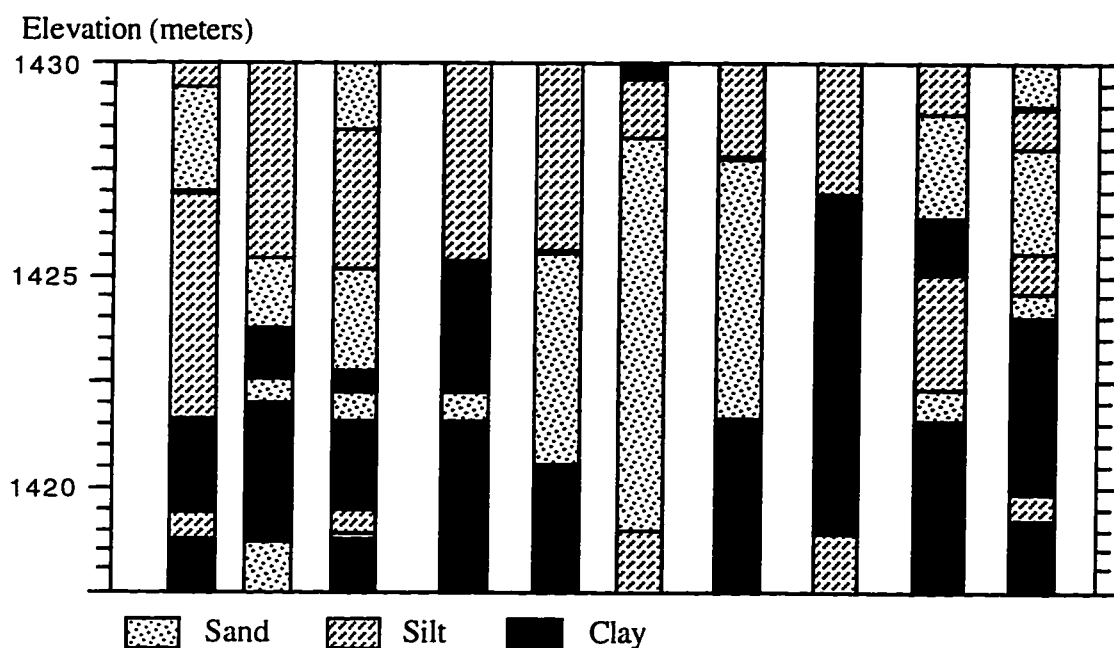


**Figure 4-1.** Simulation procedures for a homogeneous semi-Markov model.

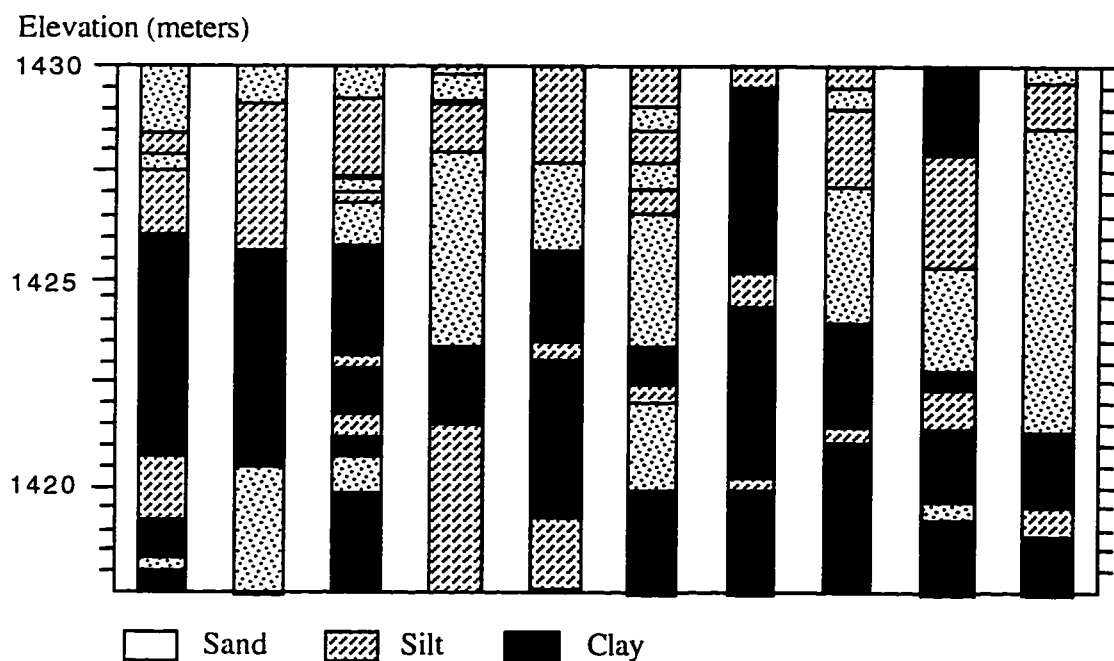


**Figure 4-2.** Kernel estimation of the transition intensity in the vertical. A bandwidth of 18 m, equal to distance to the  $k^{\text{th}}$ , ( $k=9$ ), nearest transition is used. Calculations for two quantities are presented: 1) the unconditional probability of clay is the dashed area under the kernel normalized to the bandwidth; and 2) the transition intensity from clay to sand is the ratio between the sum of kernel weights at the transition points and sum of weighted thickness,  $d$ , of soil of type clay.

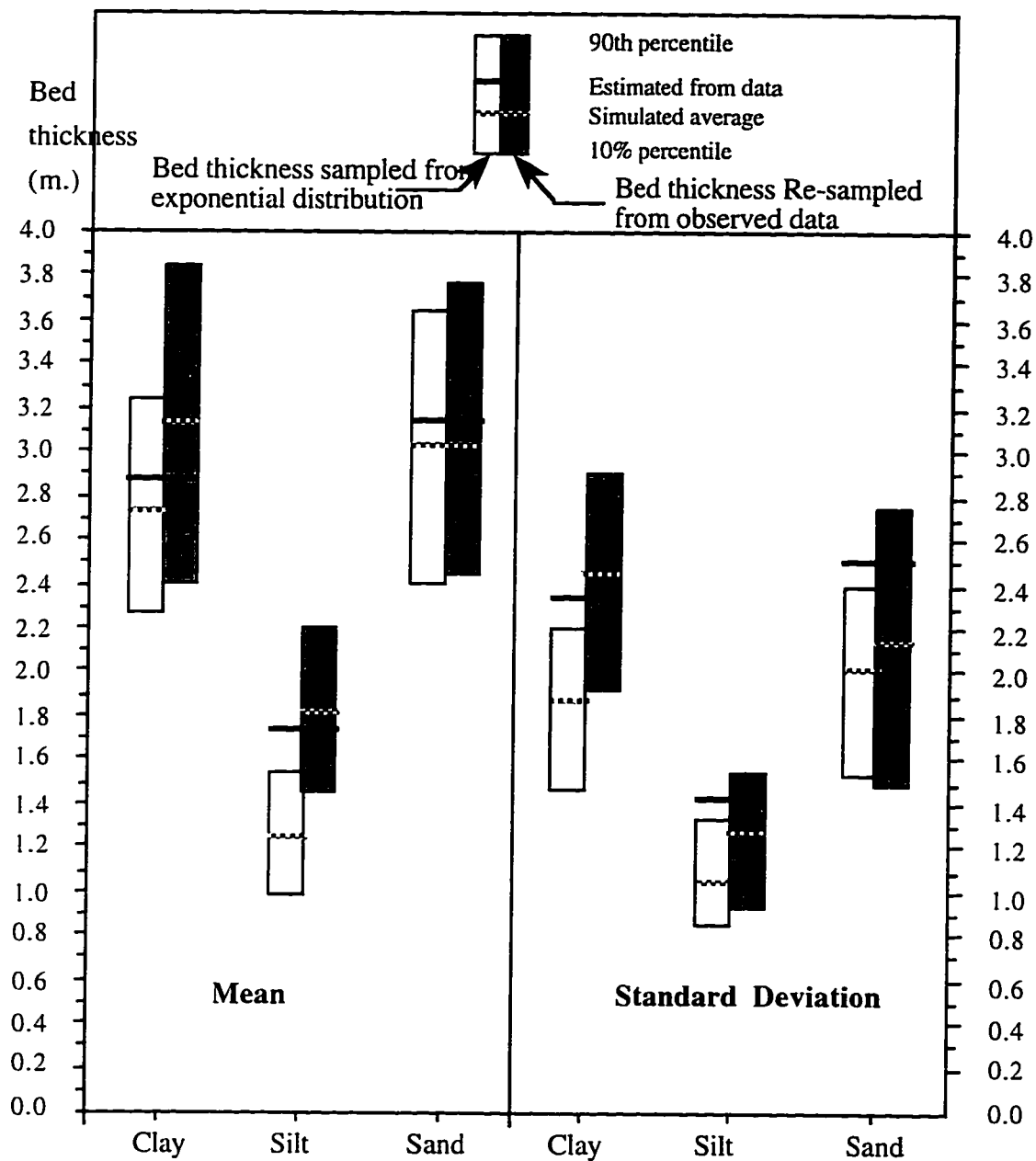




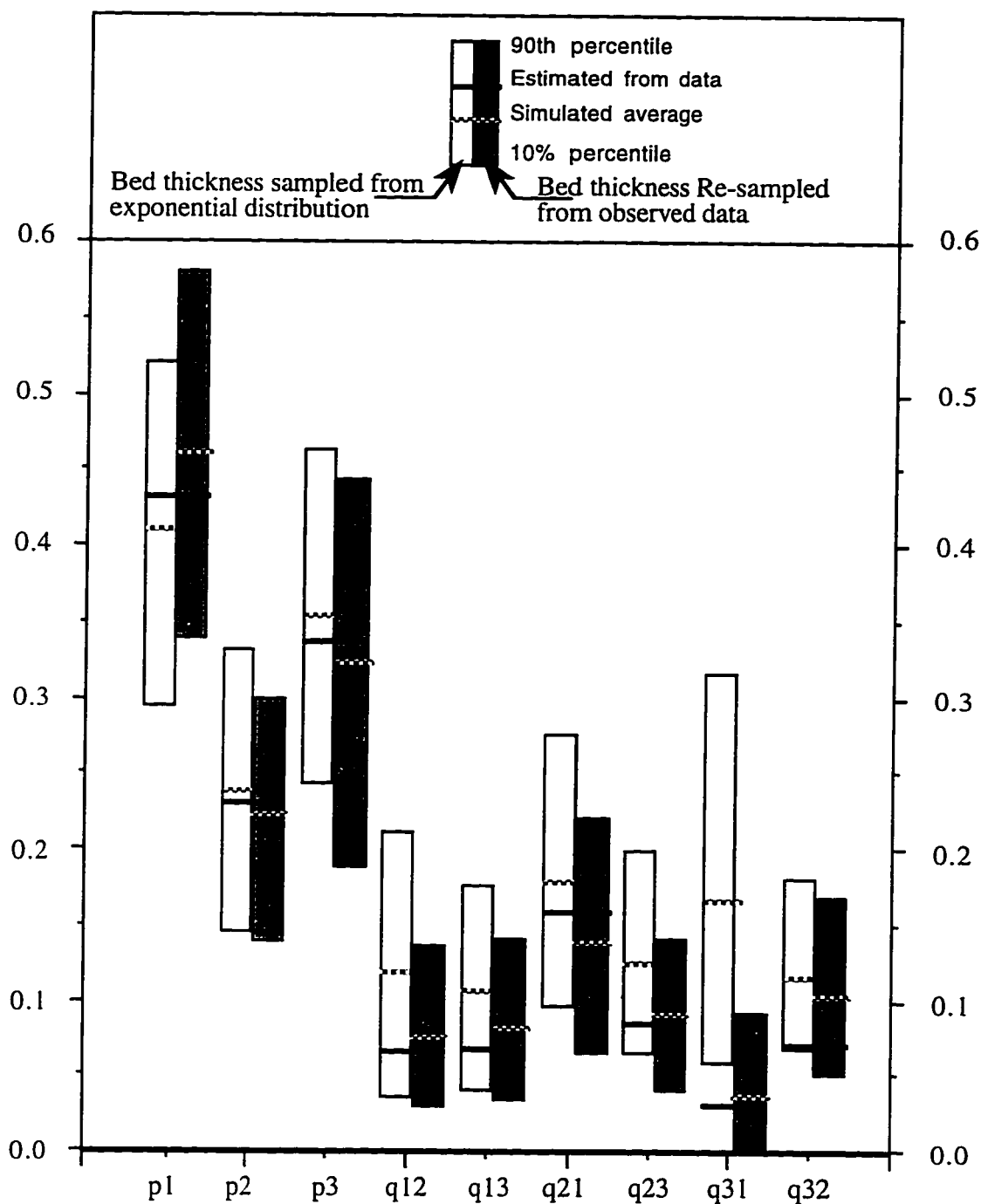
**Figure 4-3.** Simulated well logs based on an application of the CNSM model to the data from Figure 3-9. Bed thickness is resampled from the observed data.



**Figure 4-4.** Simulated well logs based on an application of the CNSM model to the data from Figure 3-9. Bed thickness is resampled from an exponential distribution.



**Figure 4-5.** Boxplots for the bed thickness statistics (in meters) of the three soils from 90 CNSM realizations based on: 1) direct resampling from data, 2) an exponential distribution. Resampling from observed data is substantially better in preserving the statistics, particularly the standard deviation.



**Figure 4-6.** Boxplots for the probabilities (p) and transition intensities (q) averaged over the entire profile for clay (1), silt (2), and sand (3) from 90 CNSM realizations based on: 1) Resampling from observed data and 2) an exponential distribution. The results are generally comparable for the two methods. Resampling from the observed data is substantially better where there are differences.

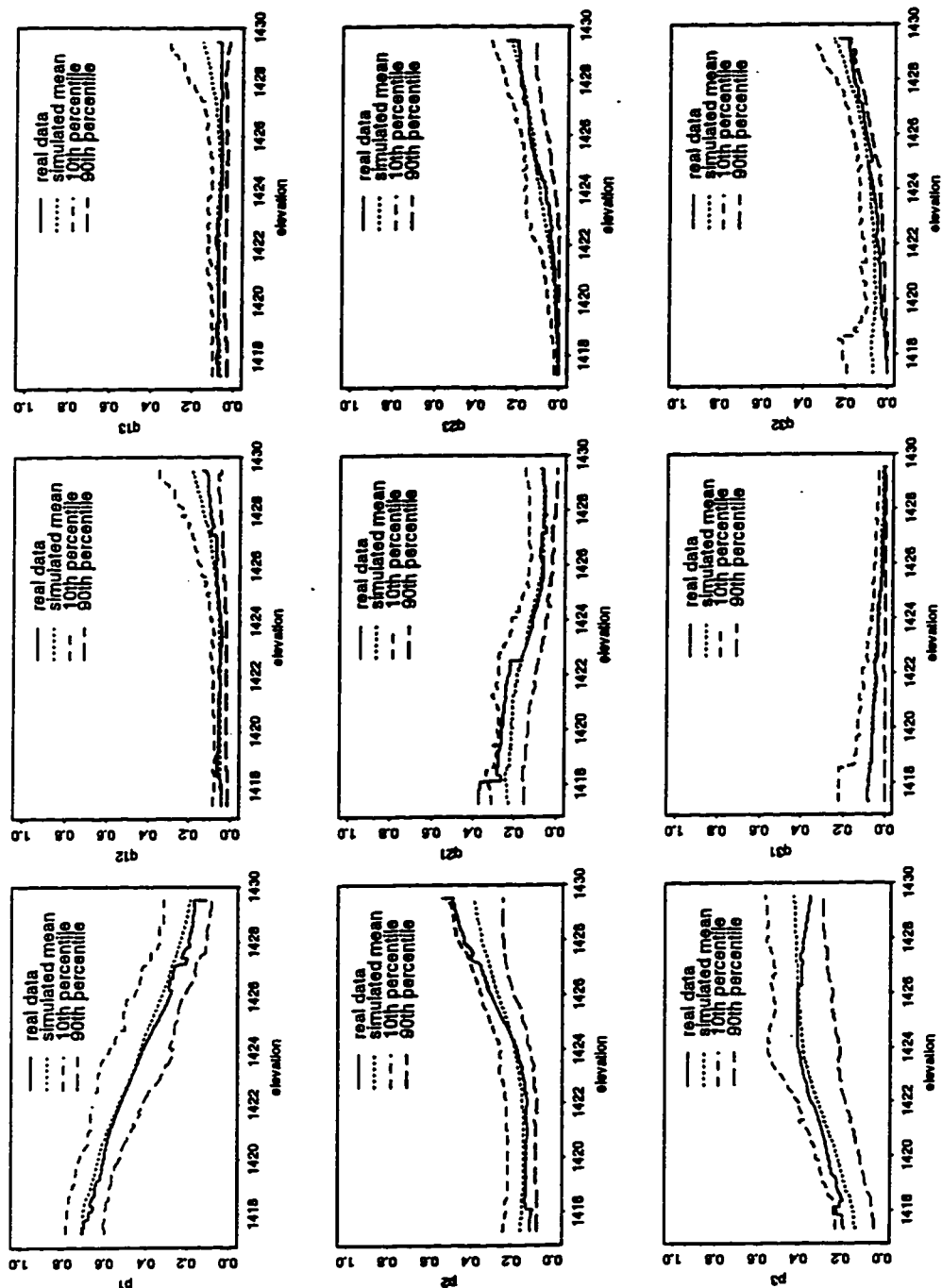


Figure 4-7. The UPM and TIM elements as a function of elevation for the real and pseudo-well logs. Bed thickness is resampled from the real data.

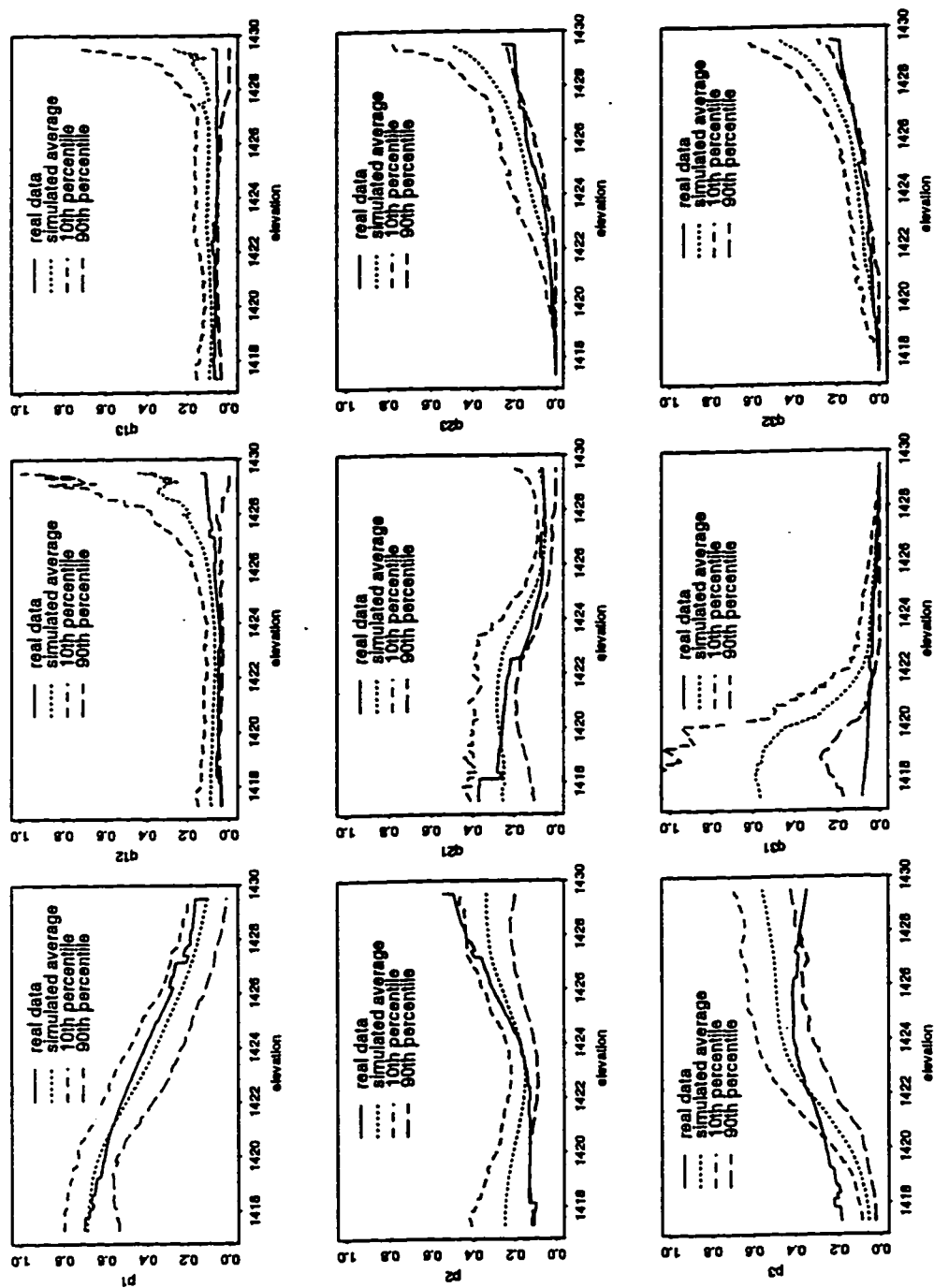


Figure 4-8. The UPM and TIM elements as a function of elevation for the real and pseudo-well logs. Bed thickness is sampled from an exponential distribution.

CHAPTER 5  
A K-NEAREST NEIGHBOR SIMULATOR OF PSEUDO-BORE-  
HOLE LOGS FOR SUBSURFACE CHARACTERIZATION<sup>1</sup>

Abstract

Subsurface characterization is important for investigations of groundwater contamination as well as petroleum extraction potential. Bore hole or drill logs are a common source of localized stratigraphic information. Sampling and interpolation uncertainties complicate the characterization of the subsurface from these data. Geostatistical methods (Kriging based) have been used to map subsurface attributes from such data, and to generate conditional simulations that honor observed strata. These methods often require an a priori discretization of the bore hole data along each vertical section into strata types, as well as assumptions as to the statistical homogeneity of the underlying random field that generated the discretized data.

A method for generating likely realizations of subsurface soils by sampling pseudo-drill logs on to a horizontal grid by bootstrapping (resampling) drill logs is presented here. Entire drill logs are resampled on to the grid locations, thus obviating the need for a prior discretization of the sample space. The bootstrapping algorithm used can be interpreted in terms of a nonhomogeneous random field description of the data with the necessary probability distributions estimated implicitly using a nonparametric (k-nearest neighbor) probability density estimate for the sampling distribution of the drill logs (real or pseudo) that lie in a neighborhood of a grid point at which resampling is needed. Example applications to a synthetic and to a real data set are provided. Extensions to incorporate other sources of information, as well as potential applications of the method, are discussed.

---

<sup>1</sup>Coauthored by Alaa Ali and Upmanu Lall.

## Introduction

The spatial variation of subsurface attributes (e.g., hydraulic conductivity or porosity) is of interest as problems of groundwater contamination are studied and an identification of preferential pathways for flow and transport is needed. Such preferential pathways may occur at pore or fracture scales, or as macro-scale connected features (e.g., leakage zones between shallow, contaminated aquifers and deeper, confined, water supply aquifers). The latter are of interest here. In sedimentary, alluvial environments, the soil deposition and erosion processes can be spatially heterogeneous. Structured heterogeneities may arise as a consequence of low frequency, cyclical climatic variability superposed on prior topography and soil conditions. Spatial randomness in the subsurface soils is present as a consequence of variability of the many interacting factors that shape the subsurface environment. A limited sampling of the subsurface environment is available through surface exposure mapping, point measurements (e.g., bore holes), and areally averaged or bulk tests (e.g., pump tests). Given these factors, stochastic methods are considered useful for subsurface characterization from various data sources.

Bore hole (drill) log data are commonly acquired at sites where groundwater contamination or oil exploration is of interest. They are also inventoried by regulatory agencies (e.g., State Water Rights, environmental protection, and energy and mining Entities). These data provide a nearly continuous sampling of the subsurface in the vertical. However, essentially only a point section of the aquifer is sampled laterally at each bore hole site. Existing pump test methodologies only provide estimates of areally averaged hydraulic parameters, and hence are of limited value for preferential pathway identification. Tracer tests can be more useful in this regard, since there is a possibility of identifying travel times between different sections of an aquifer. Often, shallow aquifers may be

contaminated, and a regulatory agency is concerned about the likely locations where leakage of this water to deep, confined, water-supply aquifers may occur, under localized head gradient reversal. Drill log data are useful in this context since they provide direct information on variation in subsurface soils in the vertical. These are the primary data source considered in the work presented here. The analysis of potential hydraulic connections between any two points in the aquifer would be feasible if one sampled the subsurface with an arbitrarily large number of drill logs on a regular grid or lattice. This would provide information on soil variation along the vertical at a large number of locations. Since acquiring such a data set is economically and practically infeasible, one needs a procedure to provide such a representation given existing, scattered, bore hole data.

A goal of the work reported here is to provide a methodology for generating candidate realizations of  $np$  pseudo-drill logs on a regular horizontal (two dimensional) lattice (i.e., a rectangular grid of sites), given a set of  $n$  drill logs at scattered locations over the geographical region of interest. Each such realization provides a plausible representation of detailed, subsurface characteristics that can be analyzed to identify potential connected pathways or other features. Analysis of the attributes of multiple realizations allows one to make probabilistic statements about parameters (e.g., travel time or degree of connection between any two points) of interest.

Geostatistical methods, such as indicator Kriging, have been used in the past to analyze bore hole logs [*Journel*, 1989; *John and Dreiss*, 1989] and to generate simulations of subsurface attributes on a two-dimensional lattice [*Deutsch*, 1992]. Typically, subsurface soils are categorized into two (e.g., 1 for high conductivity and 0 for low conductivity) or more categories; each drill log is discretized in the vertical at some resolution; assuming stationarity of the stochastic process that generated this discretized data, a variogram is selected to represent the associated spatial correlation structure; and an image (e.g., a value



of soil type at each lattice cell) is generated using conditional simulation of the associated spatial random field. Difficulties with such an approach include 1) a lack of clear understanding of the consequences of the a priori discretization of the sample space (the spatial coordinates, as well as the soil types); 2) the validity of the assumption of stationarity; 3) the identifiability of an isotropic or anisotropic variogram from the data; and 4) the weakness of spatial correlation (i.e., large nugget) at the spatial scales of interest. The last point often relates to the lack of statistical homogeneity across the site.

This chapter is one of a series reporting work on analyzing bore hole logs using nonparametric density/intensity estimators [*Lall*, 1995] for other hydrologic applications) as an alternative to the Kriging-related methodology. Kernel intensity and regression estimators (these are weighted moving averages centered about a point of estimate) were used in Chapter 2 to estimate the unconditional probability of observing a high conductivity soil at a given location in the aquifer from bore hole logs converted to a binary sequence (1 for high conductivity and 0 for low conductivity) of breakpoints (depths to each interface between 0 and 1). A binary soil type classification, but no discretization of the drill log is needed, and "memory" or spatial dependence between nearby locations is limited to that induced by a nonhomogeneous Poisson process whose rate parameter represents the likelihood of occurrence of the soil type of interest. Consequently, this method may not be suitable for generating realizations that preserve spatial continuity of soils. Chapters 3 and 4 explore homogeneous and nonhomogeneous semi-Markov process representations of transitions in the vertical between two or more soil types (states), and allow the associated state transition intensities to vary with spatial location. Once again, the drill logs are not discretized, but the soil types are, and kernel intensity and regression methods are used to estimate model parameters as a function of spatial location. The variation of the transition intensities with location in the aquifer is useful for insights on the persistence of a particular

type of soil and, hence, on connectivity. Realizations of the subsurface are provided by sampling from the semi-Markov process. A problem with this procedure is that reliable estimates of the state transition probabilities are difficult to obtain when the number of observed transitions and drill logs is small and the number of soil types considered is large. This is the classical curse of dimensionality as the dimension of the state space increases.

The simulation procedure presented here does not require an a priori discretization of either the soil types or the drill log vertical coordinate (entire drill logs are resampled to the lattice vertices); is similar to the k-nearest neighbor bootstrap for resampling time series [Lall and Sharma, 1996; Rajagopalan and Lall, unpublished report, 1996]; and is motivated by a representation of the drill log data as a vector-valued (consider each drill log  $i$ , to be a sequence of  $m$  blocks of soils of type  $j$  of thickness  $d_{jm}$ ) nonhomogeneous random field  $\{\psi(\mathbf{x}), \mathbf{x} \in \mathbf{I}\}$  where the set  $\mathbf{I}$  is assumed to be a regular, discrete lattice in  $\mathbf{IR}^2$ , the horizontal plane. The simulation algorithm employed is discussed after a brief discussion of the problem addressed. Theoretical considerations involved in the design of the algorithm as well as some implementation details are then discussed. Applications to two data sets (a synthetic and a real one) follow.

### Problem Definition

Consider the situation depicted in Figure 5-1. Given  $n$  existing drill logs, a stochastic realization of pseudo logs at a regular lattice of  $n_p$  sites needs to be generated. The resampling method should be such that 1) each pseudo-log is representative of that location and 2) spatial continuity of soils is preserved in a statistical sense across realizations, and is representative for a given realization. Spatial continuity implies dependence of the process at one point on its values in some locale about that point. It is desirable to eschew a prior discretization of the sample space in this process. This may still be necessary for mapping

or interpreting attributes, but may not be needed for generating sequences. The method used should adapt to a variety of situations, specifically to statistical heterogeneities and provide a general framework with minimal modeling assumptions.

*Ripley* [1981] observes that Markov processes are most used where the process is defined in one dimension (e.g., time) since the order property (e.g., future/past) is fundamental to Markov theory. In higher dimensions (e.g., on the plane), a Markov process can still be used if one considers dependence of the process  $\zeta(\mathbf{x})$  on the process value at neighbors of  $\mathbf{x}$ , that lie within a certain distance  $\rho$  of  $\mathbf{x}$ . If the data were obtained on a lattice, construction of such models is feasible. Here, the original drill logs need not be regularly spaced. Clearly, the Markov property implies *local* dependence of the process. In a nonstationary environment, the nature of this dependence may change with locale. For the problem discussed here, the process  $\psi(\mathbf{x})$  is vector valued with integer and continuous arguments (  $m$  *ordered* blocks of soils, with soil types indexed by  $j$ , and with thickness  $d_m$ ) defined at each spatial location  $\mathbf{x}$ . This is a difficult problem to address using traditional statistical methods.

The bootstrap [*Efron*, 1979] is a sampling with replacement technique that is useful for developing empirical confidence limits for sample statistics whose sampling distributions are difficult to evaluate analytically. An advantage of this method is that it provides samples of the underlying process using the available data, with a minimum of additional assumptions as to model structure. A limitation is that historically it was applicable only for independent and identically distributed data. Moving blocks of bootstraps have been developed (*Hall*, 1985, 1988; *Kunsch*, 1989) to deal with the case of dependent or spatial data. While these methods can provide proper estimators of confidence limits for statistics of the full sample, they do not preserve spatial continuity for or across realizations. The  $k$ -nearest neighbor conditional bootstrap of *Lall and Sharma* [1996] is motivated by

Markovian considerations, and aims to preserve the Markovian dependence structure. Here, we propose to adapt this scheme to the task of resampling a drill log on to a lattice location using information "local" to a neighborhood of the candidate lattice location. Hereafter, consider resampling a drill log,  $i$ , to be synonymous with resampling the vector field  $\psi(x_i)$  associated with it.

### Simulation Algorithm

Spatial imaging software (e.g., Spyglass™) often provides the user with the choice of filling each grid cell with the value of the data point that is closest to it ("nearest neighbor fill"). This is equivalent to sampling the closest real drill log to each location on the horizontal lattice. Clearly, this is unsatisfactory, since one obtains just one realization; does not recognize the spatial uncertainty associated with the information recorded at each site; and makes very limited use of the possible Markovian dependence of the underlying process. Nevertheless, this procedure illustrates that the process values at a geographical nearest neighbor may be considered representative of that point.

Next, consider resampling drill logs on to the lattice using the ordinary bootstrap. This is tantamount to choosing one of the drill logs with probability  $1/n$ , and placing it at the desired lattice location. A number of realizations of the spatial random field can now be generated. The spatial correlation structure as well as the spatially variation in the mean and other statistics of the field will however not be preserved.

One can use importance resampling [Johns, 1988; Hammersley and Handscomb, 1964] to correct this deficiency of the bootstrap. In this case, the probability with which a drill log is resampled on to a lattice location will need to depend on some criteria or weights that are applied to the original data in order to accord greater weights to drill logs that are more representative of that locale. A logical criterion would be to specify a weight (or

importance) that decreases with the distance of the existing drill logs from the candidate location. Further, one can restrict the number of nearest neighbors,  $k$ , of the lattice location that are considered. This will localize the area from which drill logs are resampled at the current location. This scheme can generate realizations, like the bootstrap, with local structure that is representative of the random field since one of the  $k$ -nearest neighbor drill logs is selected.

Lateral spatial continuity (vertical continuity is always preserved since an entire drill log is resampled) in a given realization can be preserved by considering both the original drill log data and previously filled/resampled lattice locations as candidates to be one of the  $k$ -nearest neighbors. However, this device can reduce the variation across realizations, if the lattice locations are always simulated in the same order, since there will be a tendency to fill the locations with the prior nearest neighbor lattice drill log. This problem is solved by randomizing the sequence in which the  $np$  lattice locations are filled for each realization. Once the  $np$  locations are filled, the domain is resimulated by visiting the lattice locations through a different random sequence. The data in this case are the  $n$  real drill logs, and  $np$  pseudo logs. The purpose of the second cycle is to eliminate a possible bias due to a certain random sequence, and to provide a better spatial continuity across the site. More cycles may be performed for a better continuity convergence. From our experience, the number of cycles required depends on the sparsity of drill logs.

The weight function, and the number of nearest neighbors,  $k$ , remains to be specified. *Lall and Sharma* [1996] considered a Markov time series model of order  $p$ , that is specified through the conditional density function  $f(Z_t|Z_{t-1}, Z_{t-2}, \dots, Z_{t-p})$ , where  $Z_t$  is the stochastic process of interest, and  $\{Z_{t-1}, Z_{t-2}, \dots, Z_{t-p}\}$  is a state space vector  $\theta_t$  it is conditioned on. For the drill log resampling strategy described above, the equivalent conditional density function is  $f(\psi(x)|J_k(x))$ , where  $\psi(x)$  is the vector-valued random field

at a lattice location  $\mathbf{x}$ , and  $\mathbf{J}_k(\mathbf{x})$  is the vector containing the spatial locations of the  $k$ -nearest neighbors of  $\mathbf{x}$ . This can be thought of as a nonhomogeneous (the density function of  $\psi$  is location dependent) Markov process of order 0. *Lall and Sharma* [1996] assumed that if one considered the  $k$ -nearest neighbors of  $\theta_t$  in state space to be locally Poisson distributed, then a resampling kernel (a weight function that recognizes how much weight to assign to each state space neighbor in considering it as representative of the current state space vector  $\theta_t$ ) could be derived as:

$$K_j = \frac{1/j}{\sum_{j=1}^k 1/j}, \quad j=1 \dots k \quad (5-1)$$

This is a probability mass function, where decreasing probabilities are assigned to further neighbors (increasing index  $j$ ). Since there may be two or more neighbors at the same distance, it is useful to randomly permute the indices of lattice locations and drill logs prior to distance calculations. In our context, this assumption implies that the drill log locations that are the  $k$ -nearest neighbors of the current lattice location, are presumed to be randomly (Poisson) distributed with a local mean. This is a tenable assumption, since 1) the sequence in which the lattice locations is filled is random and 2) even if there is clustering of filled sites and real drill logs, only a local density of sampling locations is of interest.

*Lall and Sharma* [1996] and *Diggle and Matern* [1981] recommend  $k \approx \sqrt{n}$ , as a rule of thumb, if a formal choice based on cross validation is not used. This rule of thumb is used here. One can readily change  $k$  to match desired statistics from simulations.

The simulation algorithm is summarized in the flowchart in Figure 5-2.

## Applications

The methodology presented is applied to two case studies. The first one is a control situation, where boreholes are sampled from an environment where a prior probability distribution of sand occurrence is assumed. The second uses borehole data from a site in Ogden, Utah. All dimensions mentioned in this section are in meters.

### Control Situation

The purpose of this section is to answer the question: Given boreholes sampled from an environment with prescribed probability distribution of a given soil type, do KNN generated realizations, using such boreholes, honor such a distribution? A binary aquifer representation [soil type under consideration (say sand) / soils of other types, (say clay)] is considered adequate to answer this question. However, this is not a feature or limitation of the KNN model as it can be used for multi-type soil representation.

The control setting is a sedimentary environment with sand occurrence prescribed by the probability distribution:

$$\lambda(x,y,z) = \sqrt{\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2} \text{ if } \left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2 \leq 1, \text{ and} \quad (5-2)$$

$$= 1 \text{ else.}$$

where  $-250 \leq x \leq 250$ ,  $-250 \leq y \leq 250$ , and  $0 \leq z \leq 60$ .

A cross-sectional view of this environment through isolines of  $\lambda$  is shown in Figure 2-4. The values of a, b, and c are 250, 250, and 60 m, respectively. This environment was sampled by 60-m deep boreholes at 144 equally spaced locations. At each pseudo-borehole, sand (assigned 1) or clay (assigned 0) units were randomly generated at a 1-m vertical discretization. The 1-m section is assigned sand if  $\lambda(x,y,z) > u$ , and clay if  $\lambda(x,y,z) \leq u$ . Here, u is a uniform random number between 0 and 1, and  $\lambda(x,y,z)$  is

computed from equation 5-2, and  $(x,y,z)$  are the coordinates of the center of the 1-m section.

A set of 144 boreholes is generated from the environment described above. This data set is used to generate 100 realizations of sand/clay (or 0/1) on a 10 m by 10 m by 1 m grid in the x, y, and z directions, respectively. Such realizations are averaged over the grid nodes producing an estimate,  $\tilde{\lambda}$ , for the true probability distribution,  $\lambda$ .

## Results

The estimated contours of  $\tilde{\lambda}$  from the 100 realizations are shown together with the true isolines of  $\lambda$  in Figures 5-3 and 5-4. The estimated  $\tilde{\lambda}$  value appear to be consistent with the underlying model. Compared to similar results obtained by the kernel estimator (Figures 2-5, and 2-6), these results show rougher picture due to the unsmoothed variation in the vertical.

## Ogden Valley Aquifer System

The Ogden Valley has an aquifer system that is typical of Lake Bonneville sediments that cover large portions of the state of Utah. The site under consideration is located just west of the Wasatch Mountain Range on the relict Weber Delta. The delta consists of broad plains and terrace, and originates on the western base of the Wasatch Range. Topographically, this site is on a plateau formed by the Weber Delta. The plateau is approximately 90 m above the valley floor. Surface elevations at this site vary from 1400 m above mean sea level along the western side, to 1540 m near the eastern side. Depths to bedrock in the basin ranges from 460 m on the western side to 2300 m on the east. The available data lies in the upper 20 m of unconsolidated geologic material that consists of



silts, clays, gravels, and sands. The geological units of interest at this site are the Pleistocene Provo formation, and the Pleistocene Alpine Formation.

Groundwater is found in the shallow alluvial deposits, in the sand and gravel deposits of the Provo Formation, and in the sand lenses within the underlying Alpine Formation, which is predominantly clay. The uppermost zones of ground water are locally discontinuous and exist under unconfined conditions. The groundwater in the sand and silt layers of the upper portions of the underlying Alpine Formation usually exists under confined or semiconfined conditions.

The base of the unconfined aquifer (Provo Formation) rests unconformably on the Alpine Formation. The relatively low permeability clay deposits impede the downward migration of ground water contaminants and enhance lateral migration along the upper surface of the clay in the downdip direction. Localized pockets of groundwater and dense non-aqueous phase liquid, DNAPL, have been identified in sand and silt lenses within the clay matrix to depths of 30 m below land surface [Rich, 1995]. Even though the permeability of the Alpine Formation is relatively low, these sand seams may allow for enhanced fluid migration in their primary direction of orientation.

The concern at the study site was deep migrating free DNAPL that was detected in the aquifer system. Site characterization of the subsurface is important in identifying possible travel paths and zones of DNAPL. A binary classification of the soil type for result interpretation is used here. This is not an inherent limitation of KNN and is not used in the KNN procedure. This is done here so as to compare with semi-Markov analysis application. We chose to assign 1 to soils of permeability higher than silt, and 0 otherwise. However, any other threshold may be selected.

The site is bounded by 570000 m, 570280 m east, and 89125 m, 89430 m north. Forty-four irregularly spaced bore holes were available for this application. Bore hole

depths ranged from 5 to 36 m. The subdomain used for this investigation has coordinate axes rotated  $17^{\circ}$  counterclockwise with an origin located at 570150 m east, and 89160 m north. A layout for the borehole locations as well as the subdomain is shown in Figure 5-5.

## Results

One hundred realizations were generated within a three-dimensional grid with dimensions of 108, 195, and 10 m (divided into  $36 \times 65 \times 30$  pixels of dimensions:  $3 \times 3 \times .33$  m in the x, y, and z directions, respectively). Figure 5-6 shows a probabilistic image, average across realizations, of sand/clay distribution. This figure shows identifiable areas with high/low conductivity zones embedded in opposite type (zones B and C). Also it shows high probability of observing sand in the northern, top part of the image (zone A). These features show up in individual realizations as well, in the general area, but vary in their manifestation in each realization (Figure 5-7.).

Tests of the ability of the model to preserve statistical attributes of the original data field are compared. The nonhomogeneous semi-Markov process model developed in Chapter 4 is applied to both the raw data, and to each realization generated. The transition intensities and the unconditional probabilities of each types of soil are estimated and compared (see Figure 5-8). We see that the variation in the unconditional probability of sand/clay with elevation is reproduced very well by the simulations. The estimates of the transition intensities from clay to sand and sand to clay exhibit increased bias and variance near the ends of the domain. This may reflect a bias in the estimation of the transition intensities using the kernel method rather than a problem with knn, since we resample the full drill log with knn, and do not actually have a boundary estimation problem. There are very few transitions from sand/clay or clay/sand in these data. Alternately, this may reflect a need to change k.

Figure 5-9 shows drill log data along axis y-y in Figure 5-5. Compared to the image in Figure 5-7, these drill logs show reasonable consistency with the feature identified in the image. Although these drill logs show quite a bit of variety, and they are not the same logs being resampled at all nearby locations, they seem to evolve in a consistent manner with the probabilistic trend observed in the image.

### Discussion and Conclusions

A method for generating likely realizations of subsurface soils by bootstrapping  $k^{\text{th}}$  nearest drill logs was developed. A probability mass function (kernel function), equation 5-1, was used to resample the nearby drill logs where decreasing probabilities are assigned to farther neighbors. The simulation proceeded in a sequential manner such that at location  $I(i)$ , a pseudo log is sampled from  $n$  real drill logs and  $i-1$  pseudo logs. Application to a control situation demonstrated the efficacy of the  $k$ -NN resampler in reproducing the probability distribution of a given soil type. The results of Ogden Valley aquifer show that the  $k$ -NN resampler preserves the unconditional probability of bed thickness and the transition intensity in the vertical from sand to clay and from clay to sand. Also, observation at drill logs along a certain line (axis y-y in Figure 5-5) shows reasonable consistency with the horizontal trend in the probabilistic image (Figure 5-6.).

One of the problems encountered in the implementation of the  $k$ -NN simulation model is the possibility of having drill logs that start at different elevations and are of different lengths. This is a problem that may be faced if there are significant terrain variations, and the sampling program is not designed specifically for the investigation. Changes in elevation may be accommodated by moving the top of the resampled drill log to match either 1) the surface elevation of the lattice point, or 2) the general dip and strike at the site as one moves from the original location to the lattice location. One may restrict the

neighbors to include only those that have drill logs of adequate length (or pick up the lower piece from a nearby longer drill log). None of these solutions is entirely satisfactory. Another strategy would be to first extend all drill logs to the same (or desired length) using the simulation methods developed in Chapter 4. However, this may not provide results that differ from using the longer drill logs to fill in the lower pieces.

The k-NN simulator is computationally fast. Typical cpu times to generate a realization with 70200 lattice nodes and 2340 drill logs were of the order of 30 seconds on a DEC 3000 computer. The generation of a large number of realizations for analysis is thus feasible. The realizations generated can be classified into a set of soil types (this information carries over with the original drill logs) and then displayed using a three-dimensional modeling package such as Spyglass Dicer. Sections of the site can be highlighted and animation can be used to provide subsurface "tours." Each such realization can be used with a scheme such as in Chapter 2 to identify connected preferential pathways in the subsurface, as well as to compute any other statistics of interest (e.g., variograms, layer thicknesses, layer transition probabilities) and confidence intervals associated with those statistics.

Extensions of this approach to consider other sources of information, e.g., seismic well logs, or seismic interpretations, need to be developed. These were not pursued here, largely because of a lack of ready availability of such data. In concept such data can be readily considered through inclusion in the state parameter vector (currently just the location coordinate) in the process of searching for the k-nearest neighbors of the lattice location. Similarity for the purpose of importance resampling probability estimates would then be defined in terms of the ancillary information as well as location attributes. A higher order Markov random field model (e.g., the model would have the subsurface parameter of interest depend on the parameter values at neighboring locations, instead of just on the

locations) could also be considered. An extension of the k-nn simulator to consider such dependence is feasible through an appropriate specification of the state parameter vector. However, this may not be a practical solution given data limitations, unless stationarity or statistical homogeneity were assumed.

### References

- Deutsch C. V., Annealing techniques applied to reservoir modeling and the integration of geological and engineering (well test) data, Ph.D. dissertation, 306 pp., Stanford University, Stanford, Calif., 1992.
- Diggle, P. J., and B. Matern, On sampling designs for the estimation of point-event nearest neighbor distributions, *Scandinavian J. Stat.*, 7, 80-84, 1981.
- Efron, B., Bootstrap methods, Another look at the jackknife, *Annals of Statistics*, 7, 1-26, 1979.
- Hall, P., Resampling a coverage pattern, *Stochastic Processes and Their Applications*, 20, 231-246, 1985.
- Hall, P., On confidence intervals for spatial parameters estimated from nonreplicated data, *Biometrics*, 44, 271-277, 1988.
- Hammersley, J. M., and D. C. Handscomb, *Monte Carlo Methods*, 60 pp., Methuen, London, 1964.
- John, N. M. and S. J. Dreiss, Hydrostratigraphic interpretation using indicator geostatistics, *Water Resour. Res.*, 25(12), 2501-2510, 1989.
- Johns, M. V., Importance sampling for bootstrap confidence intervals, *J. Amer. Stat. Assoc.*, 83, 709-14, 1988.
- Journel, A. G., *Fundamental of Geostatistics in Five Lessons*, 40 pp., Stanford Center for Reservoir Forecasting, Stanford University, Stanford, Calif., 1989.
- Kunsch, H. , The jackknife and the bootstrap for general stationary observations, *Ann. of Stat.*, 17, 1217-1241, 1989.
- Lall, U., Nonparametric function estimation: Recent hydrologic applications, *Reviews of Geophysics, US National Report 1991-1994*, 1093-1102, 1995.
- Lall U., and A. Sharma, A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resources Research*, 32( 3), 679-690, 1996.

Rich, N., DLOG3D application at operable unit 2 Hill Air Force Base, M.S. thesis, 93 pp., Utah State University, Logan, 1995.

Ripley, B., *Spatial Statistics*, 252 pp., John Wiley and Sons, New York, 1981.

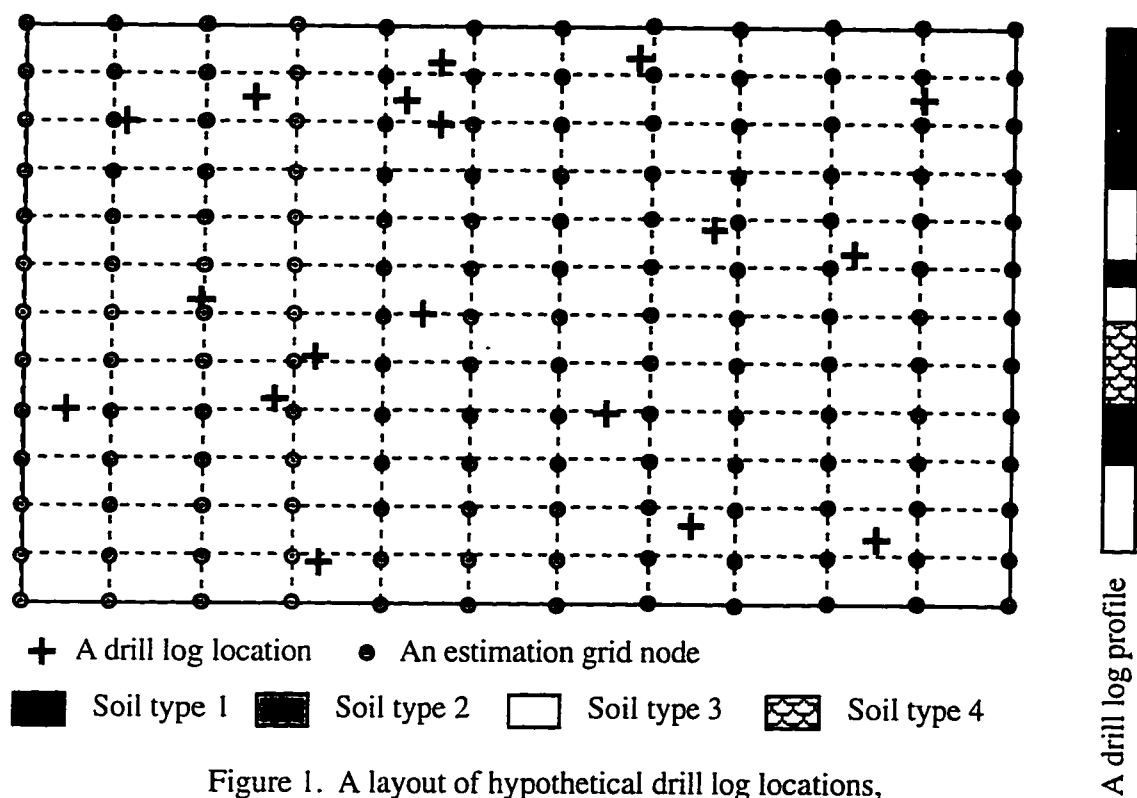
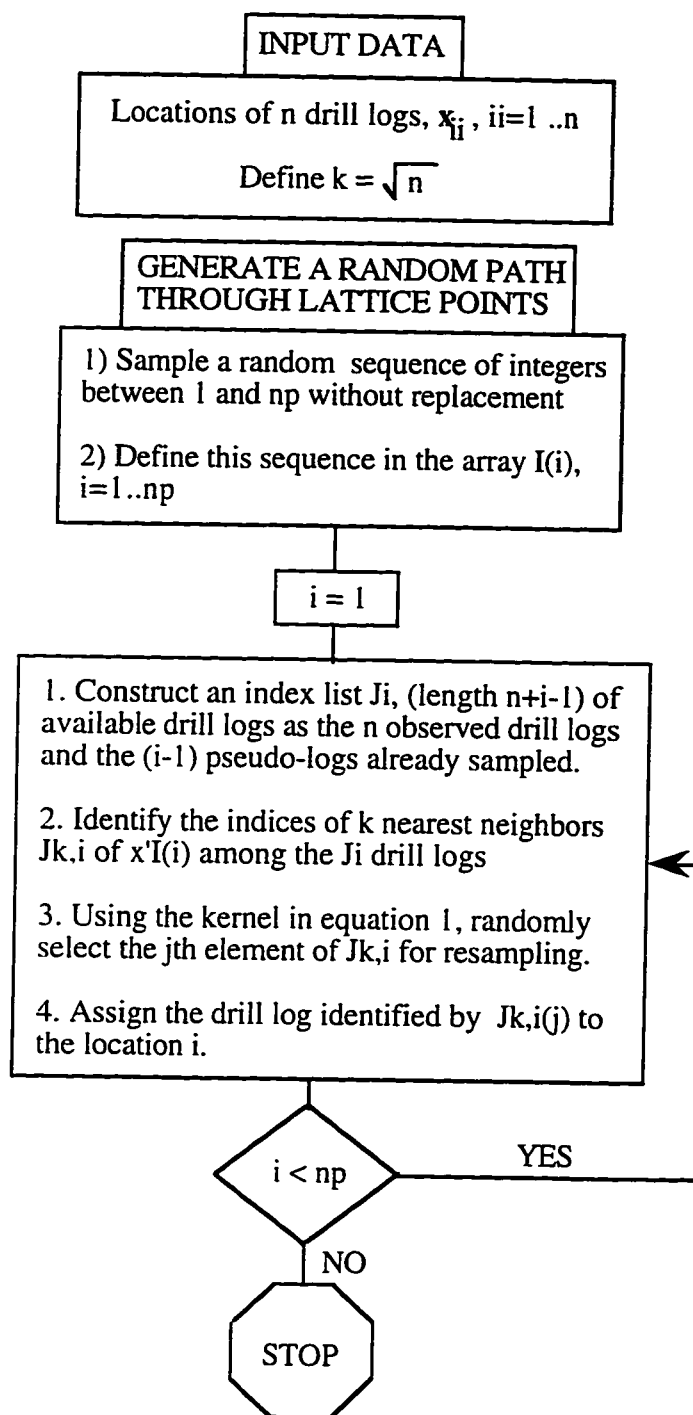


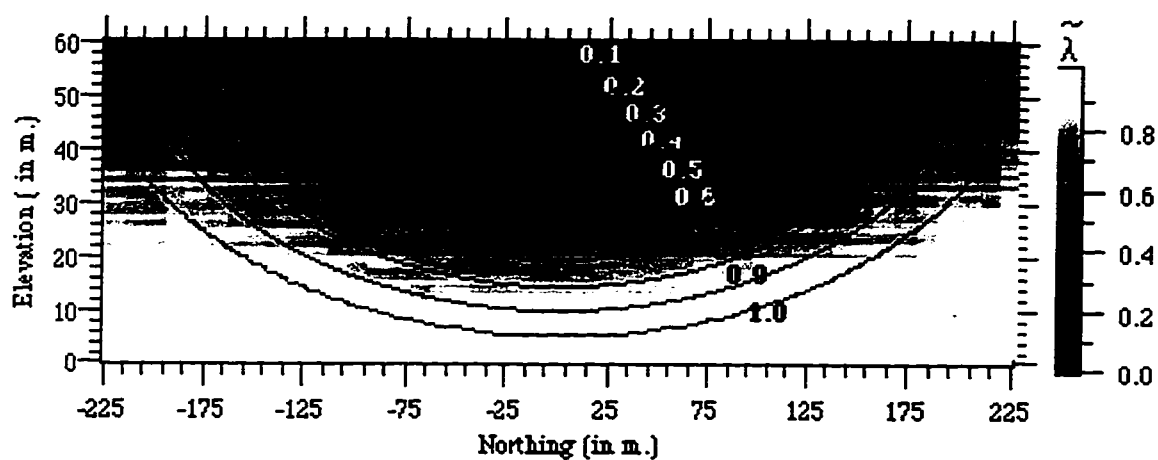
Figure 1. A layout of hypothetical drill log locations, the estimation grid, and a hypothetical drill log profile.

**Figure 5-1.** A layout of hypothetical drill log locations, the estimation grid, and a hypothetical drill log profile.

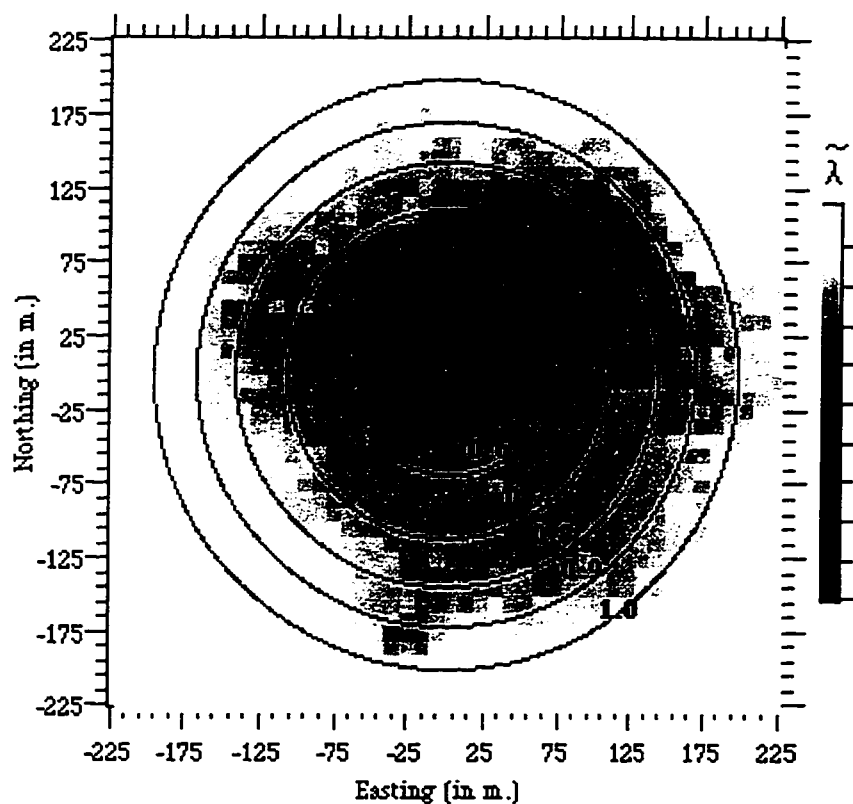


**Figure 5-2.** Simulation procedures for a realization using KNN method.

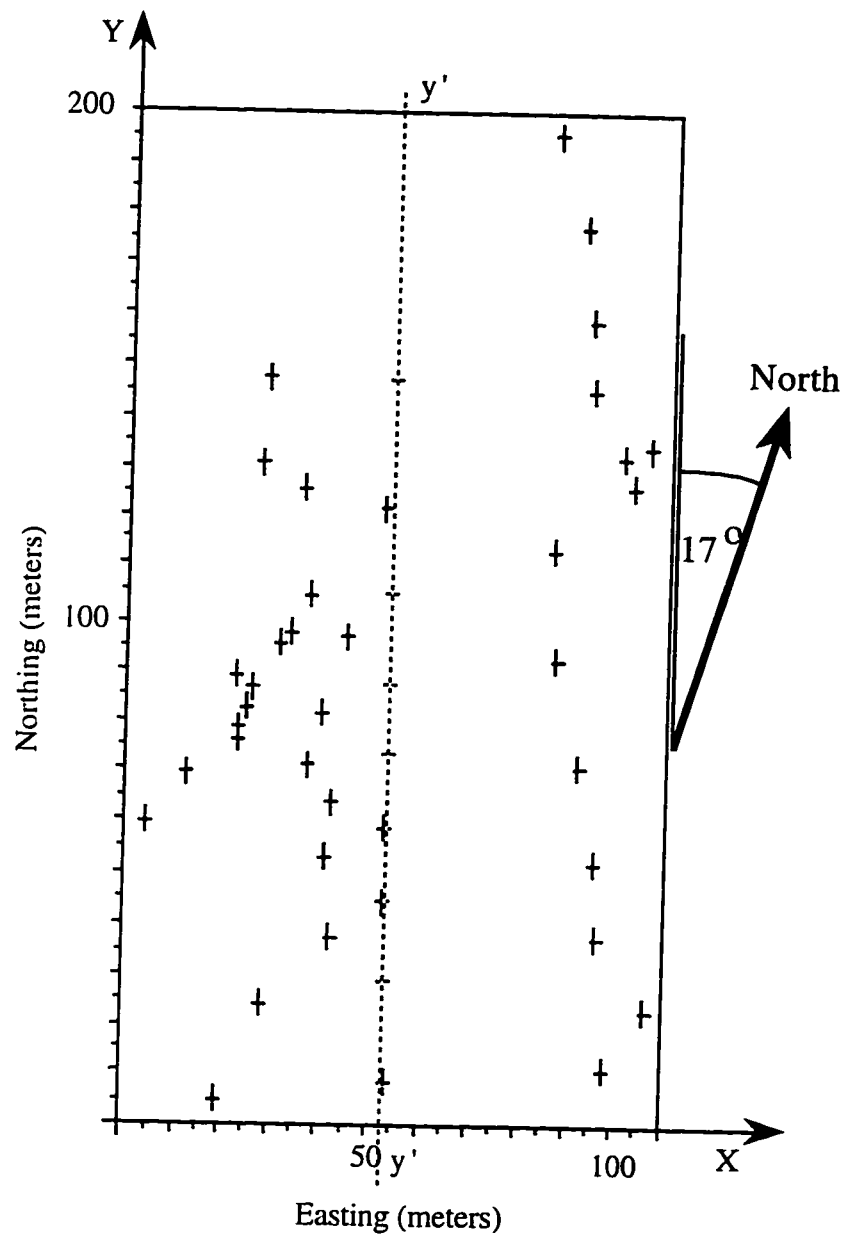




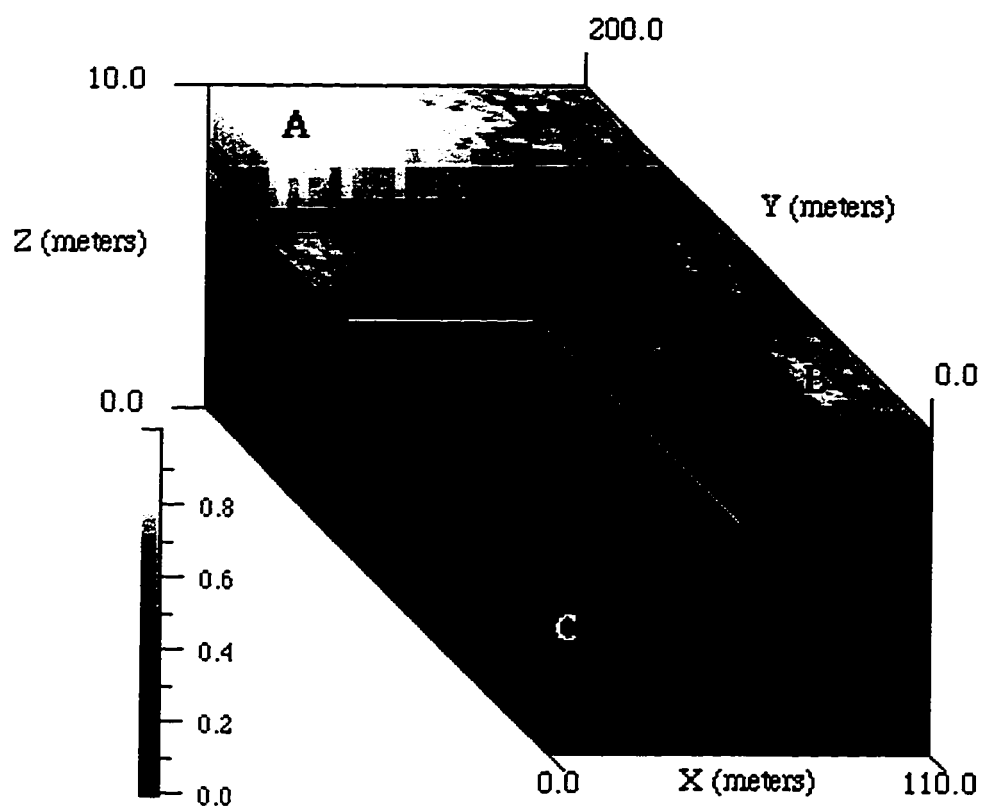
**Figure 5-3.** Elevation-Northing of the probability of sand occurrence for synthetic data (contour image), and the true function (solid lines) at the site center (Figure 2-4, section A-A').



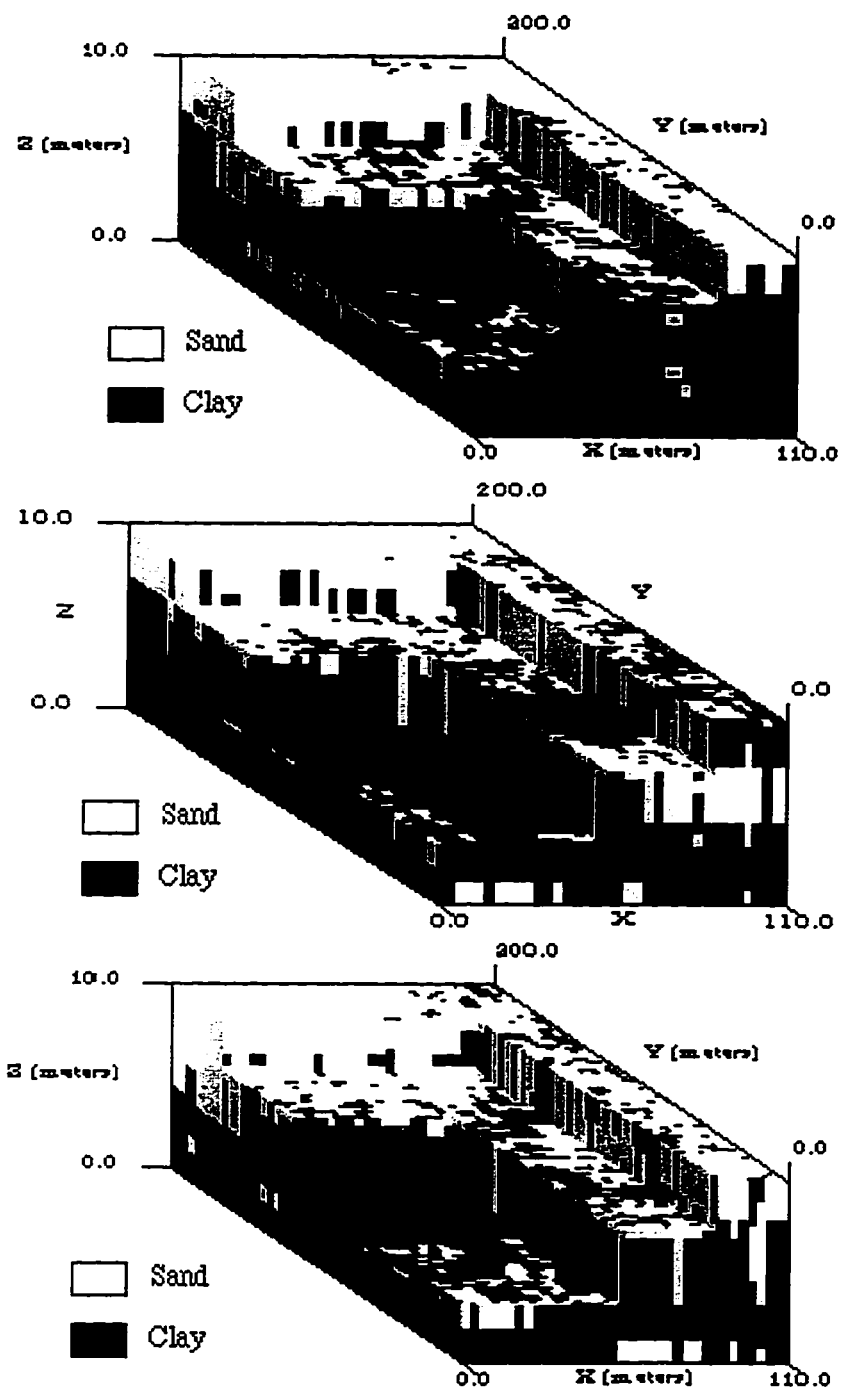
**Figure 5-4.** Northing-Easting of the probability of sand occurrence for synthetic data (contour image), and the true function (solid lines) at elevation = 30 m (Figure 2-4, section B-B').



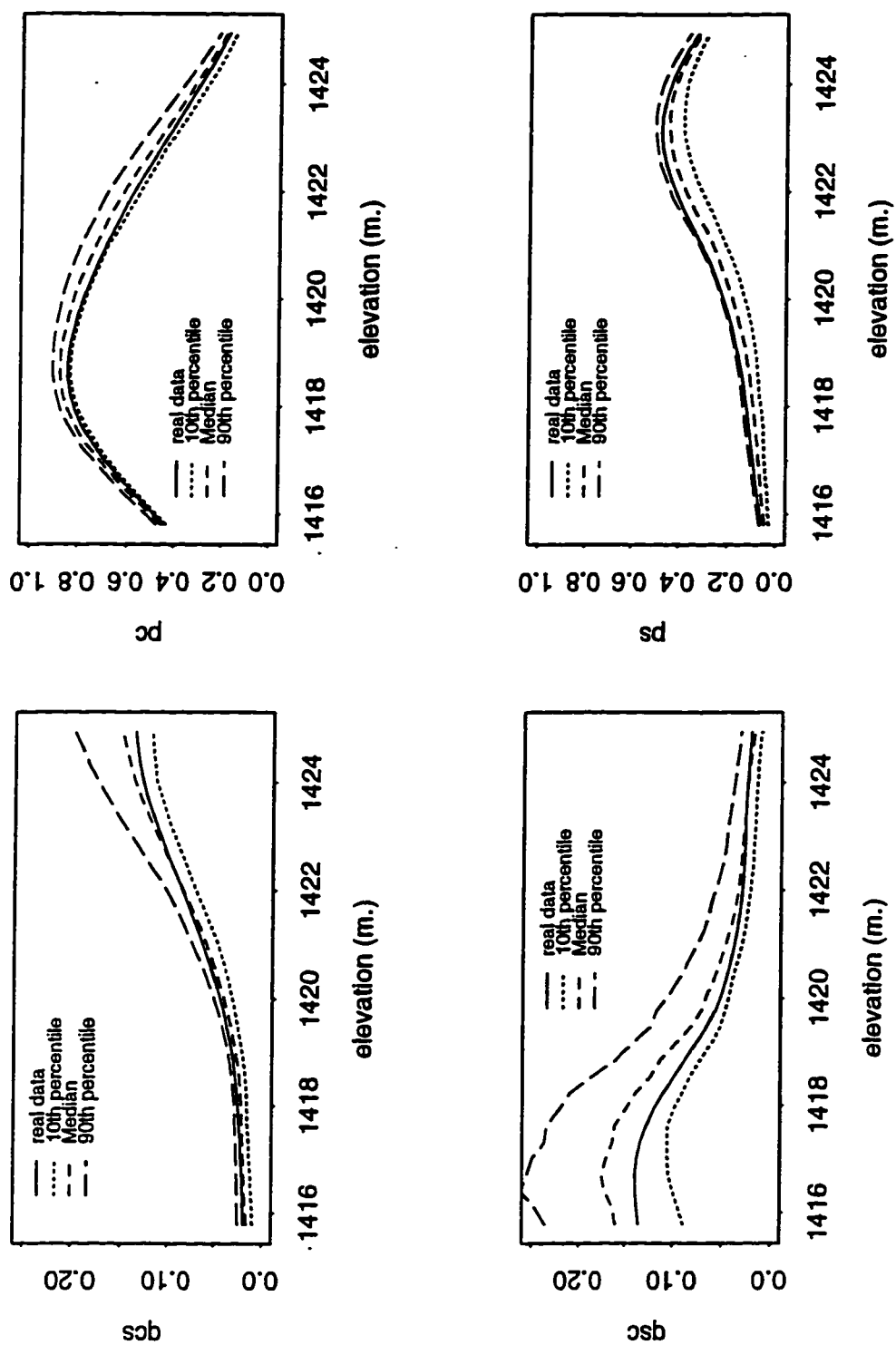
**Figure 5-5.** Plan view of the drill log data layout and locations of estimate. Origin is located 1870560 m. East, 292510 m. North. Coordinate axes have been rotated 17 counter clock wise.



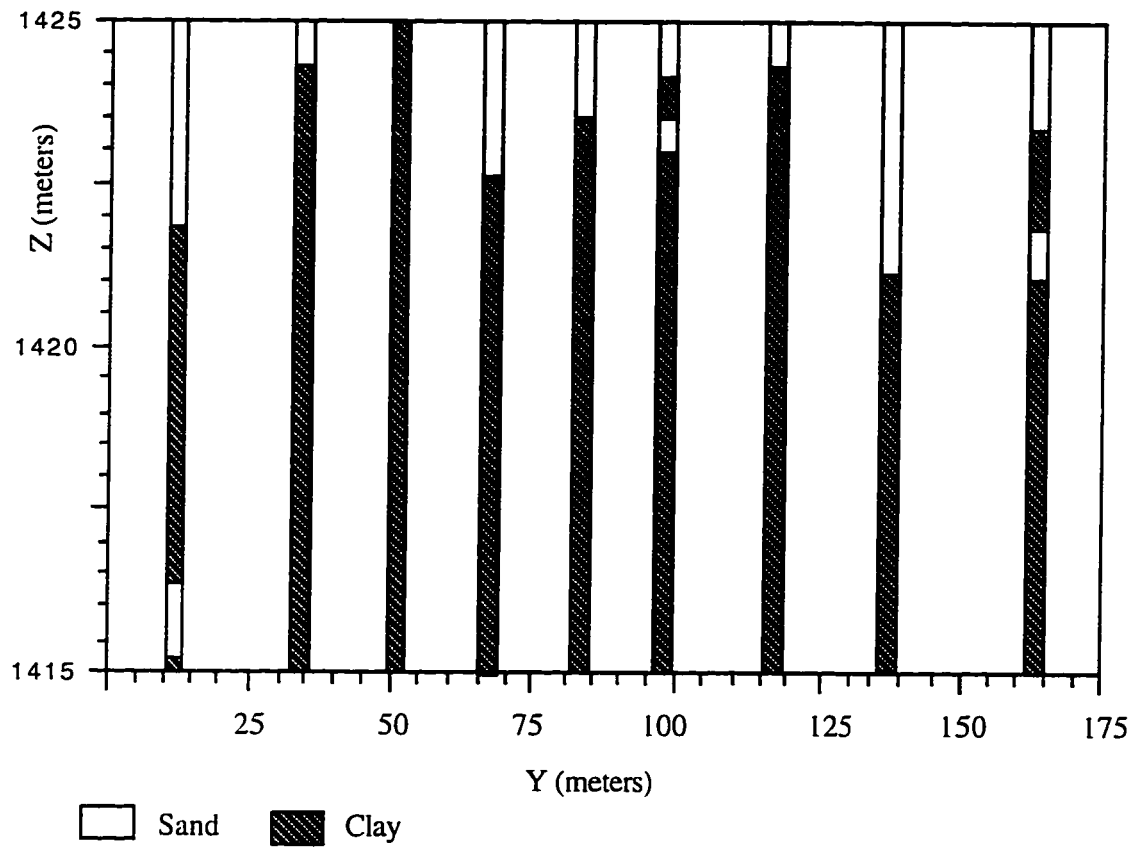
**Figure 5-6.** Three-dimensional cut out of the aquifer system. A, B, and C are probabilistic zones that indicate the likelihood of occurrence of geologic features.



**Figure 5-7.** Three realizations of the aquifer system which honor the probabilistic image presented in Figure 5-6.



**Figure 5-8.** The transition intensities, and unconditional probabilities as a function of elevation for the real data and the simulated images.



**Figure 5-9.** Bore holes along axis y-y' in Figure 5-5.

CHAPTER 6  
IDENTIFYING POTENTIAL PREFERENTIAL PATHS  
FOR SUBSURFACE TRANSPORT USING  
SIMULATED ANNEALING<sup>1</sup>

Abstract

Connected zones of high hydraulic conductivity in subsurface environments can constitute preferential pathways for contaminant transport and fluid flow. The identification of such pathways is important for well head protection, groundwater monitoring and capture system design for groundwater remediation. Since it is not practical to exhaustively sample the subsurface, the existence and probable extent and location of such paths must be inferred from available data. Drill or bore hole logs are the primary source of subsurface data considered. A stochastic simulator is used to generate realizations of subsurface soil attributes on a lattice. Measures of connectivity between any two locations in the aquifer are introduced. These measures consider travel time between lattice locations under an assumption of unit hydraulic gradient along each candidate-connected path. A preferential pathway between any two points in the aquifer is then one with the highest degree of connectivity (lowest travel time). A simulated annealing algorithm is then used to identify potential preferential pathways between any two prescribed points in the aquifer for each realization. This algorithm seeks superior solutions to global optimization problems that have a large solution space. Simulated annealing proceeds by generating a sequence of random connected paths between the two points of interest. The objective function (connectivity) is evaluated for each candidate path. A threshold probability is defined for accepting or rejecting alternate solutions that may have higher objective function values than

---

<sup>1</sup>Coauthored by Alaa Ali and Upmanu Lall.

the best solution obtained at a given iteration. The candidate path is then randomly perturbed, and the new objective function evaluated. The procedure is iterated with the perturbation length, and the acceptance probability reduced as iterations progress until the solution converges to a near optimal solution. The algorithm was applied to a synthetic data set and to data from a site in Ogden, Utah.

### Introduction

Spatial variability in soil type (hydraulic conductivity) strongly influences contaminant transport and flow direction. One way of investigating such an influence in a heterogeneous soil block in a large-scale problem is to find an equivalent homogeneous soil block with single (effective) hydraulic conductivity. This approach does not account for local interconnectedness between zones of high conductivity, ignores the impact of the potential presence of preferential pathways, and, hence, may not lead to reasonable analyses for contaminant travel time distribution. Preferential pathways for flow in porous media due to interconnected zones of high hydraulic conductivity, e.g., sand lenses or fractures, are important to modeling contaminant transport and aquifer pollution control [Smith and Schwartz, 1984; Fogg, 1986; Hestir and Long, 1990; Schafer-Perini and Wilson, 1991]. Such paths are of concern when considering the transport of chemicals moving within the groundwater because they create the potential for early arrival of a chemical at a pumping well location (Figure 6-1). The identification of these pathways is important to federal and state agencies for contaminant risk assessment and management decisions in well head protection and water rights programs. They are also important for designing extraction systems for oil reservoir management.

In this study, the focus is on the identification of such pathways for a sedimentary environment where the soil is classified into two or more classes and the soil information is



resampled from bore hole data on a three-dimensional lattice. A probabilistic representation of the field is incorporated and simulated annealing is used to identify potential preferential pathways between different points in the aquifer. Such an analysis can then be used for 1) better groundwater monitoring, 2) treatment or interception system design, and 3) well head protection zone delineation. Often, such an analysis is of interest also to identify potential interconnections between shallow contaminated aquifers and deep water source aquifers.

Existing measures for analyzing connectivity do not reflect the complex patterns of high conductivity lenses [Fogg 1986; Journel and Albert, 1988]. Also, methods for investigating interconnectedness do not provide explicitly defined spatial connectivity patterns across a site (Silliman and Wright, 1988). In this study, we will explicitly identify preferential pathways and an associated measure of the travel time.

The paper is organized into four sections: 1) a review of existing methods for identifying preferential pathways or interconnectivity, 2) a formal definition of the problem solved here, 3) a solution methodology, and 4) applications to a synthetic and field data sets.

## Background

The purpose of this section is to review selected concepts and estimation methods of interconnectedness and preferential pathways that exist in the literature.

In a heterogeneous aquifer with high degree of contrast in hydraulic conductivity (e.g., sand/clay sequences), a few well connected lenses of high hydraulic conductivity can completely change the spatial pattern of the hydraulic conductivity and velocity fields. Fogg [1986] showed that the effective permeability in the vertical increases by a factor of  $10^2$  if the sand lenses are interconnected. Also, he found that the degree of interconnectivity

between sand lenses directly increases the flow rate by a factor  $10^3$  to  $10^4$ . Suppose we had perfect information on the attributes of such an aquifer, what are some measures of connectivity? How does one define or identify a connected path?

Various efforts have been made to identify high conductivity features, and/or estimate a measure of connectivity by: 1) developing a geologic representation for the aquifer system and 2) using such a representation for flow simulation [Fogg, 1986; Scheibe and Freyberg, 1990; Webb and Anderson, 1996]. These studies succeeded, to a certain extent, to provide a realistic representation of the aquifer system but did not quantify the interconnectedness across the aquifer system. Fogg [1986] used geologic information to reconstruct the geologic facies of the Wilcox aquifer system near the Oakwood Salt Dome. He considered lateral interconnection between channel-fill sand bodies where the sand fraction exceeds 20 percent (thickness-percent maps of channel-fill sand bodies were available). Such a measure provides a qualitative idea about the interconnectedness, but does not necessarily correspond to possible complex connectivity patterns nor does it provide a quantification for the effective degree of interconnectedness. In another study, Webb and Anderson [1996] focused on the reconstruction of the internal architecture of the aquifer system and on the identification of braided channels using topographic maps and some more detailed surficial geometrical measurements for model calibration. Although particle tracking simulations show reasonable consistency between flow pathlines and the braided channels, the effective degree of interconnectivity was not quantified and the global interconnection across such high conductivity channels was not explicitly identified. Also, the uncertainty associated with such a technique is hard to determine since the data used for calibration did not include important data such as borehole lithology and/or seismic data.

In general, a perfect representation of the porous media is difficult to obtain or calibrate because of the inadequacy of the information/data available. The sources of information on the subsurface may be drill logs, seismic data, outcrop exposures, field observations, trench studies, and/or topography maps. Drill logs are usually the most commonly available source of data. The drill logs are usually at spatially scattered locations. This leads to two problems: 1) the information is not available everywhere and 2) the inference of soil type across the aquifer can only be made with some uncertainty. Given drill log data, how can one infer a measure of connectivity and identify preferential pathways across a site? To answer this question, a stochastic approach is necessary to first develop a likely representation of the subsurface. Some methods that can be used are Indicator Kriging [Johnson and Dreiss, 1989], semi-Markov models (Chapters 3 and 4), and  $k^{\text{th}}$  nearest neighbor method (Chapter 5). The purpose of these models is to make use of available data to generate a realization of the aquifer system where the spatial structure as revealed from the data is preserved. An appropriate representation is treated as a three-dimensional image in which connectivity between desired locations is investigated. The process can be repeated with multiple realizations to generate a probability distribution of likely interconnectivity between a source (origin) and sink (destination).

A measure of connectivity was proposed by *Journel and Albert* [1988]. They defined a bivariate cumulative density function (CDF) as the probability of simultaneous occurrence of two high permeability values, of separation distance  $h$ , that is greater than, or equal, specified threshold  $z$ , and is expressed as:

$$F(h; z) = \frac{1}{N(h)} \sum_{l \in N(h)} j(u; z) j(u+h, z), \quad j(u; z) = \begin{cases} 1, & \text{if } u \geq z \\ 0, & \text{if } u < z \end{cases} \quad (6-1)$$

where  $N(h)$  is the number of pairs with separation distance  $h$ , and  $z$  is the threshold of hydraulic conductivity. This is a global measure that implicitly considers the data to be

statistically stationary and, hence, ignores the local heterogeneity involving discrete boundaries and discontinuity. A recent study [Webb and Anderson, 1996] shows that such a measure failed to indicate underlying discrete structure, while particle tracking simulations showed flow occurrence along high hydraulic conductivity path.

*Silliman and Wright* [1988] investigated the paths of high hydraulic conductivity in porous media. In this study, a Monte Carlo technique was used to generate uncorrelated and correlated second-order random fields. Every grid generated by the Monte Carlo technique is searched for continuous paths that connect one face of the grid to the opposite face and along which the conductivity of each element exceeds a cutoff conductivity  $K_c$ . Their investigation focused on the identification of a maximum value of  $K_c$  at which there is at least one connected path of high  $K$ . Such a path connects all grid elements of hydraulic conductivity greater than  $K_c$ . Therefore, there are no constraints on the geometry of the path except that it connects two opposite faces. In this study, we question the following: 1) the connected path as defined does not necessarily correspond to a unique path of a flow particle, 2) ignoring large areas of  $K$  values higher than  $K_c$  which are surrounded by pixels of  $K$  values slightly less than  $K_c$ , and 3) the exhaustive search algorithm of all potential grid paths may be time consuming, and hence reliable identification of such paths may not be feasible.

In this chapter, we provide a procedure to identify the preferential paths and develop a measure of connectivity specific to locations (origin "O" and destination "D") in an aquifer.

### Problem Definition

Given a site and drill log data, we would like to probabilistically assess the degree of connectivity between any pair of locations of interest and identify the associated potential preferential pathways. As indicated in the previous section, this procedure entails 1) the

generation of stochastic realizations of a gridded representation of the aquifer system, and 2) the identification of preferential paths between the points of interest for each realization. First, we need to define a metric for connectivity. Such a metric needs to recognize the length of the path, and the resistance to the motion along that path. Given a three-dimensional lattice, each cell has a known soil type  $j$ , ( $j=1, \dots, J$ ), and hence hydraulic conductivity ( $K_j$ ) and effective porosity ( $\phi_j$ ). If we consider flow in a specific direction, e.g., vertical, the equivalent hydraulic conductivity of the aquifer in the direction normal to flow is given as:

$$\bar{K} = \sum_{i=1}^n L_i / \sum_{i=1}^n \frac{L_i}{K_i}$$
 where  $L_i$ ,  $K_i$  are respectively the thickness and hydraulic conductivity of the  $i^{\text{th}}$  layer traversed.

Clearly, the  $\bar{K}$  is maximized as the denominator is minimized. One can analogously consider  $\bar{K}$  to be a measure of connectivity along a path in the lattice, where  $L_i$  can be thought of as the cell widths and  $k_i$  the hydraulic conductivity of each cell. A preferential pathway is then one for which  $\bar{K}$  is maximized. Equivalently, we can think of this process in terms of a pseudo travel time under a fixed (e.g., unit) hydraulic gradient along a flow path. Preferential pathways would be those with the shortest travel times between the two points of interest. The identification of the preferential pathway between any two points O and D (e.g., a contamination source and a well) for a given realization can then be cast as an optimization problem. The objective of the optimization problem is to search for a connected path between O and D that has the minimum travel time compared to all other connected paths that may be found between O and D.

The lattice is composed of  $N=n_x*n_y*n_z$  cells "or pixels" using a discretization ( $\delta x*\delta y*\delta z$ ) of the aquifer. A node on the lattice and the associated cell are identified through an index  $i=1, \dots, N$ , and an associated position vector  $\vec{x}_i$ . A connected path  $\pi$  between any two points O and D on the lattice is defined through an ordered set  $J_{OD,\pi}$  of

node locations  $i$  such that the  $k^{\text{th}}$  members of the set is one of the six neighbors (see Figure 6-2) of the  $(k-1)^{\text{th}}$  member of the set. The set has  $mm$  members, representing the number of nodes traversed. The length of the connected path is then  $\sum_{k=1}^{mm-1} d_{\theta(k-1), \theta(k)}$ , where  $d_{\theta(k-1), \theta(k)}$  is the Euclidean distance between successive element centroids along such a path,  $\theta(k)$  is the index of element  $k$ , and  $mm$  is the total number of elements in  $J_{OD, \pi}$ . The travel time  $T_{OD}$  along the path  $\pi$  is 
$$\sum_{k=1}^{mm-1} \frac{d_{\theta(k-1), \theta(k)}}{2} * \left( \frac{\phi_{\theta(k-1)}}{K_{\theta(k-1)}} + \frac{\phi_{\theta(k)}}{K_{\theta(k)}} \right)$$

Given  $N$  cells in the lattice, and a reasonable separation between  $O$  and  $D$  (e.g., they are not adjacent), the number of distinct connected paths that can be formed is very large (several orders of magnitude of  $N$ ).

Each such path corresponds to a value of the decision variable for the optimization problem of interest here. This is a vector-valued (elements of  $J_{OD, \pi}$ ) decision variable with integer arguments (node location  $i$  included in  $J_{OD, \pi}$ ). Such combinatorial problems are considered "NP-hard." A global optimal solution to such a problem is difficult to guarantee without exhaustive enumeration and within usual computational constraints. We use a technique (simulated annealing) that is successful in obtaining good solutions to such a problem.

In summary, we seek solutions to the identification of potential preferential pathways by 1) generating stochastic realizations of the aquifer system on a lattice, 2) defining a metric for connectivity of high conductivity materials, and 3) identifying the connected paths between selected "origin/destination" pairs on the lattice that have the potential for being preferential paths between those pairs of points, using simulated annealing.

### Problem Solution Strategy

As stated earlier, a realization of the aquifer is needed for the interconnectedness

investigation. A typical step in any stochastic simulator is to divide the domain space into subunits with a certain level of discretization. A preliminary study showed that the probability density function (pdf) of the travel time is very sensitive to the level of discretization. Unless such a discretization reflects the natural system, the pdf of the travel time may not be meaningful. Here, we use the  $K^{\text{th}}$  nearest neighbor, KNN, simulator described in Chapter 5 to generate each lattice realization from drill log data. This simulator presumes each drill log to be the outcome of a spatial stochastic process modeled as a Markov Random Field. The random field is resampled using a KNN probability density estimator. Entire drill logs are resampled as pseudo-logs onto horizontal lattice coordinates. The vertical continuity of soil type is thus explicitly preserved. The horizontal continuity/structure of soils is preserved implicitly by conditioning resampling a drill log at each lattice location using the real as well as the pseudo-drill logs that are the KNN of the current lattice location. Given a realization, and an O, and D, the simulated annealing technique used to identify preferential pathways between O and D is now briefly described.

### Introduction to Simulated Annealing Technique

Recall that the objective is to identify a connected path between O and D that has the minimum travel time compared to other connected paths that may be found between O and D. Due to the large configuration space of alternative connected paths, one cannot exhaustively search for optimal path. There are several methods to find optimal, or near optimal, solutions for such large-scale problems. Iterative improvement [Kirkpatrick, 1984] is a neighborhood search algorithm for the optimal solution. In this method, a possible alternative configuration is locally searched in the neighborhood of the current configuration  $\pi$ . If the new configuration results in a decrease of the cost function (i.e., travel time), the new configuration is favored over the old one. In this method, the system

usually can get stuck in a local minimum since it accepts only the moves leading to better solutions, which may not necessarily lead to a globally optimum solution. Simulated annealing is an improved neighborhood search algorithm where the local minima problem is treated.

Simulated annealing technique is a structured random search algorithm that proceeds sequentially evaluating candidate solutions searching for a global optimal one. Such a technique is presented in the following steps:

- 1) Generate a candidate solution (connected path between O and D).
- 2) Evaluate the corresponding objective function (travel time).
- 3) Perturb the solution with a certain perturbation length,  $m$ , and evaluate the new objective function.

- 4) Accept, or reject the new solution according to the Metropolis Criterion:

$$P_{\text{accept}} = \begin{cases} 1 & \text{if } \Delta t \leq 0 \\ e^{-\left(\frac{\Delta t}{C}\right)} & \text{if } \Delta t > 0 \end{cases} \quad (6-2)$$

where  $\Delta t$  ( $t_{\text{new}} - t_{\text{old}}$ ) is the difference between the travel time along the new path and that along the old path, and  $C$  is a control parameter.

The use of the above acceptance probability criterion allows the acceptance of unfavorable solution ( $t_{\text{new}} > t_{\text{old}}$ ) to give the system the chance to get out of local minima.

- 5) Steps 3 and 4 are repeated while the perturbation length  $m$  and the control parameter  $C$  are lowered according to a certain schedule.

- 6) The process is terminated when the objective function asymptotically approaches a certain value.

Input parameters for such a technique are presented in Figure 6-3. A detailed description of the basic elements of simulated annealing process is given below.



### Generation of Candidate Solution (Connected Path)

A system configuration is represented by a connected path between O and D. Such a path may be constructed in a deterministic manner possibly resulting in a biased path and, hence, a possibility of local minima. On the other hand, a purely random construction may not represent a realistic flow path. Therefore, a structured algorithm is developed to construct a random path (see flow chart in Figure 6-4).

A connected path between the origin and destination points is constructed using  $n$  segments of random paths connecting  $n-1$  randomly selected points in space. The number of elements in segment  $k$  is  $mm_k$ , and the total number of elements in the connected path is  $mm = \sum_{i=1}^n mm_i$ . An element  $j$  in segment  $k$  has index  $\sum_{i=1}^{k-1} mm_i + j = I(k,j)$ . A segment  $k$  connects a point of index  $\sum_{i=1}^{k-1} mm_i = I(k,0)$  (considered local origin with position vector  $\vec{x}_{I(k,0)}$ ), and a point of index  $\sum_{i=1}^k mm_i = I(k,mm_k)$  (considered local destination with position vector  $\vec{x}_{I(k,mm_k)}$ ). The construction of a segment between  $[\vec{x}_{I(k,0)}, \vec{x}_{I(k,mm_k)}]$  is restricted by: 1) the path is allowed to visit any location in the domain, 2) the path moves in unit steps of  $\delta x$ ,  $\delta y$ , or  $\delta z$ , and 3) the path is potentially directed towards location  $\vec{x}_{I(k,mm_k)}$ . A probabilistic framework is developed to meet the three restrictions as follows:

1) At any point  $j$ , of location  $\vec{x}_{I(k,j)}$ , the path is considered to move in six potential local directions  $(x, -x, y, -y, z, -z)$ . The strategy here is to give higher weights for directions of movement towards the destination  $(-x, -y, -z$  in Figure 6-2). The new location  $\vec{x}_{I(k,j+1)}$  is identified once the local direction is selected. Given  $\Delta x, \Delta y, \Delta z$  as the cartesian offsets of point  $\vec{x}_{I(k,j)}$  from point  $\vec{x}_{I(k,mm_k)}$  and  $\theta_x, \theta_y, \theta_z$  are the corresponding directional angles (see Figure 6-2), a candidate direction is sampled with the following probability:

$$p_u = \frac{q_u}{Q} \quad (6-3)$$

where:

$$Q = q_{-x} + q_x + q_{-y} + q_y + q_{-z} + q_z$$

$p_u$  = is the probability of picking a pixel in direction  $u$ ,  $u \in \{x, y, z, -x, -y, -z\}$

$$q_{-x} = 1 + \cos(\theta_x), q_x = 1 - \cos(\theta_x)$$

$$q_{-y} = 1 + \cos(\theta_y), q_y = 1 - \cos(\theta_y)$$

$$q_{-z} = 1 + \cos(\theta_z), q_z = 1 - \cos(\theta_z)$$

2) The new location  $j+1$  is now considered as location  $j$ . Repeat step 1 for a new location and so on until point  $k$  is reached.

This probability potentially guides the connected path towards the destination without bias to any direction.

### Cost Function

The objective (cost) function is the travel time of a particle along the connected path in a constant gradient field. The travel time described here is only a conceptual definition which can be used as a measure for the actual travel time. The total travel time is the sum of the travel times,  $\delta t$ , between the  $n$  pixels along the connected path. The travel time,  $\delta t_{i, i+1}$ , between the centroids of pixel  $i$  and pixel  $(i+1)$ , is given as:

$$\delta t_{i, i+1} = \frac{1}{2} \left( \phi_r \frac{\delta_u}{K_{ru}} + \phi_s \frac{\delta_u}{K_{su}} \right), \quad (6-4)$$

the total travel time for connected path  $\pi_j$  is:

$$T(\pi_j) = \sum_{i=1}^{mm-1} \delta t_{i, i+1} \quad (6-5)$$

where:

$r, s$  = Soil types of pixels  $i$ , and  $i+1$ , respectively.

$\phi_r, \phi_s$  = The effective porosities for soil types  $r$ , and  $s$ .

$K_{ru}, K_{su}$  = The permeabilities of soils type  $r$  and  $s$  in direction  $u$ , respectively.

$\delta_u$  = The pixel length in direction  $u$ .

$mm$  = Total number of pixels along the random path (length of the random path).

### System Rearrangement

The connected path is rearranged by causing a perturbation of some length ( $m$  pixels) as follows (Figure 6-5, flow chart in Figure 6-6):

- 1) Randomly select a starting point  $x_i$  for perturbation along the original path.
- 2) From point  $x_i$ , count  $m$  pixels along the connected path and define an ending point  $x_j$  (see Figure 6-5).
- 3) Create a new random path between  $x_i$  and  $x_j$  as presented earlier.
- 4) Estimate the travel time along the original path,  $t_{old}$ , and along the perturbed one,  $t_{new}$ .
- 5) The new configuration is accepted or rejected according to Equation 2.

### Annealing Schedule

The annealing schedule is a group of parameters that control the progress of the process. The simulated annealing starts by defining initial values for the perturbation length  $m_0$  and the control parameter  $C_0$ . A stepwise reduction is applied, at a certain rate  $\alpha_1$ , to the perturbation length  $m$ . For each  $m$  step, a stepwise reduction is applied at a certain rate  $\alpha_2$  to the control parameter  $C$ . For a control parameter step, a number of accepted rearrangements is specified, and for a perturbation length step, a specific number of control parameter steps are tried. The description of these parameters is given below.

#### Initial Values for Perturbation Length and Control Parameter

The perturbation length  $m$  and control parameter  $C$  are two parameters of the simulated annealing process. The perturbation length,  $m$ , controls the degree of freedom for searching alternate paths. It is defined as a certain number of connected pixels along the path. A large perturbation of the connected path is initially desired to ensure a maximum

coverage for the configuration space. Our experience shows that too long a length results in excessive computational time, and too short a length may result in local minima as a significant portion of the configuration space is not covered. The control parameter  $C$  controls the rate of acceptance and rejection of candidate solutions. We followed the criterion of *Kirkpatrick* [1984]. An initial value of  $C_0$  is selected such that 80 percent acceptance is initially obtained. The procedures of estimating  $m_0$  and  $C_0$  are presented below (see Figure 6-7).

1) Generate a number, say  $n_{path}$ , of connected paths between  $O$  and  $D$ , and estimate the associated length  $mm$ , and consequently estimate an average  $mm_0$ .

2) Estimate  $m_0$ , the initial perturbation length, as a portion of  $mm_0$ . From our experience, an initial value  $m_0$  ranging between  $1/2$  and  $1/8$  of  $mm_0$  does not result in local minima. The value of  $m_0$  is chosen to be  $1/6$  of  $mm_0$ .

3) Generate a connected path between  $O$  and  $D$ .

4) Given initial perturbation length  $m_0$ ,  $C_0$  is estimated by first generating a certain number of perturbations ( $np$ ) of the initial connected path, and estimating the corresponding travel times  $t_1, t_2, \dots, t_{np}$ . The absolute differences  $\Delta t_i = |t_{i+1} - t_i|$  are then calculated and consequently the corresponding average:  $\Delta t_{avg} = \frac{1}{np-1} * \sum_{i=1}^{np-1} \Delta t_i$  is obtained.  $C_0$  is obtained by substituting  $\Delta t_{avg}$  for  $\Delta t$ , and setting  $P_{accept} = .8$  in equation 2.

### Decrement Coefficients

The control parameter  $C$  and perturbation length  $m$  are lowered at certain rates  $\alpha_1, \alpha_2$  ( $C_i = \alpha_1 * C_{i-1}$ ,  $m_i = \alpha_2 * m_{i-1}$ ) (flow chart in Figure 6-6). To speed up the convergence rate, our experience shows that it may be better to have decrement coefficients exponentially varying with the number of iterations. Decrement coefficients at iteration  $i$  are expressed in

terms of those at iteration  $i-1$ ,  $(\alpha)_i = (\alpha)_{i-1}^\beta$ , where  $\beta (\leq 1)$  is the increasing rate of the decrement coefficient. Figure 6-8 shows a plot of  $\alpha$  versus  $i$  at different values of  $\beta$ . Initially, the value of  $\alpha$  can be small, e.g., .5, resulting in rapid decrements initially, which allows more freedom in selecting high initial parameter values. Our experience suggests .8, and .5 as initial values for  $\alpha_1$ ,  $\alpha_2$ , respectively. The value of  $\beta$  is desired to be small enough for fast convergence, but large enough to avoid getting stuck in local minima. A  $\beta$  value of .5 resulted, for some cases, in local minima. Increasing  $\beta$  to .7 showed reasonable results. To assure a fast convergence, the upper bounds of  $\alpha_1$  are set to .99, and for  $\alpha_2$  is set to .9. A lower bound for  $m$  is set to 3 pixels.

#### Algorithm Convergence

In this problem,  $C$  is reduced after a certain number of transitions are accepted, and  $m$  is reduced after a certain number of  $C$  reductions take place (flow chart in Figure 6-6). In this study we suggest a value of  $L$ , MC length, to be  $100*m$ , where  $m$  is the perturbation length. Another type of thermal equilibrium is also considered for the number of decrement steps of the parameter  $C$  before the perturbation length  $m$  is lowered. Our experience shows that such a number may be taken as the difference between the total length of the random path and the perturbation length (i.e.,  $mm-m$ ). This means that the larger the perturbation length, the smaller the selection space along the random path.

#### **Termination Criterion**

If the average travel times at five successive  $C$  steps are the same, the system is considered frozen and the annealing process cannot cause more cooling and the process is terminated. From the results, the final and the average travel times at each time step in the final stages of the process are almost the same, indicating no significant variation within such a time step.

## Applications

The algorithm developed above is applied to two problems. The first problem is a control situation, where we know the exact geologic setting and, hence, the flow particle behavior. The second problem is a case study of the Ogden Valley aquifer where we have a realization of the aquifer lithology distribution.

### Control Situation

A control situation with a simple setting was constructed to test the algorithm developed in this study. The purpose of this demonstration is to show that a preferential pathway can be identified in a dominantly clayey environment with a thin sand layer connected across the site. Figure 6-9(a) shows a 150\*75\*25 m clay matrix with two sand layers. A unit pixel in this grid is of dimensions ( $\delta x=3$ ,  $\delta y=3$ ,  $\delta z=.8$  m). The first sand layer of thickness .8 m, is located at elevation 21 m and is interrupted by the clay matrix in two spots. Each clayey spot has maximum and minimum widths of 21 and 7 m, respectively. The second sand layer is vertical and is located at Easting = 146 m. (see Figure 6-9(b)). The environment is assumed isotropic with  $(K/\phi)_{\text{sand}}=10^{-1}$ , and  $(K/\phi)_{\text{clay}}=10^{-4}$  cm/sec. The goal is to identify the preferential pathways between an Origin (6, 15, 21) m, and a Destination (144, 75, 3 m). The optimal paths shown in Figure 6-9(b) has a theoretical minimum travel time of 140 days.

### Results

The first step in the simulated annealing process is to select reasonable values for  $C_0$ ,  $m_0$ , initial values for  $\alpha_1$ ,  $\beta_1$ ,  $\alpha_2$ , and  $\beta_2$  that lead to a correct optimal path identification. The increase of one or more of these parameters decreases the potential of local minima on

the expenses of the time needed for convergence. Reasonable values can be selected by trying the solution with several combinations of these values. We have fixed the values of  $\alpha_1(.8)$ ,  $\beta_1(.7)$ ,  $\alpha_2(.9)$ , and  $\beta_2(.7)$  and changed  $C_0$  and  $m_0$ . A value of  $C_0$  that allowed for only 50 percent initial acceptance rate, and/or a value of  $m_0 = mm/10$ , resulted in a fast convergence with a local minimum where the connected path got stuck in one of the thick bases of the clayey areas. The system with  $m_0 = mm/6$  and  $C_0$  leading to 80 percent initial acceptance rate converged to a near optimal solution (see Figure 6-9b). The control parameter  $C$  and perturbation length  $m$  are lowered in steps. For each  $C$  step, iteration, the system accepts  $100*m$  rearrangements, and for each  $m$  step, the system proceeds with  $(mm-m)$   $C$  steps. The progress of the annealing process is described in terms of several attributes of the cost function and the control parameter  $C$ . Figure 6-10 shows a plot of the cost function at the end of each  $C$  step versus the control parameter on a log-log scale. The travel time minimization process is mostly gradual except at  $C=30$  days/gradient where there is a dramatic drop. At such a time step, if the travel path finds a sand pixel, a sudden decrease of the travel time (30 days) occurs leading to a series of acceptances, and hence, a rapid drop of the travel time within this time step. Figure 6-11 shows the travel time averaged over the  $C$  step vs  $C$  step. The plot in this figure shows similar observations as in Figure 6-10 except that the local variations of the average travel time is much less than that of the final travel time.

In this application, a proper selection for the parameters led to a convergence to a global minimum. Such parameters generated high energy (similar to the hydraulic gradient) initially to drive the particle through the sand layers in the dominating clay continuity. In a setting, where the average continuity of clay layers is smaller, such parameter values may lead to reasonable convergence. However, the use of such a set of parameters in another situation may not guarantee an optimal solution.

### **Ogden Valley Aquifer System**

The Ogden Valley has an aquifer system that is typical of Lake Bonneville sediments that cover large portions of the state of Utah. The site under consideration is located just west of the Wasatch Mountain Range on the relict Weber Delta. The delta consists of broad plains and terrace, and originates on the western base of the Wasatch Range. Topographically, this site is on a plateau formed by the Weber Delta. The plateau is approximately 90 m above the valley floor. Surface elevations at this site vary from 1400 m above mean sea level along the western side, to 1540 m near the eastern side. Depths to bedrock in the basin ranges from 460 m on the western side to 2300 m on the west. The available data lie in the upper 20 m of unconsolidated geologic material that consists of silts, clays, gravels, and sands. The geological units of interest at this site are the Pleistocene Provo formation and the Pleistocene Alpine formation.

Groundwater is found in the shallow alluvial deposits, in the sand and gravel deposits of the Provo Formation, and in the sand lenses within the underlying Alpine Formation, which is predominantly clay. The uppermost zones of groundwater are locally discontinuous and exist under unconfined conditions. The groundwater in the sand and silt layers of the upper portions of the underlying Alpine Formation usually exists under confined or semiconfined conditions.

The base of the unconfined aquifer (Provo Formation) rests unconformably on the Alpine Formation. The relatively low permeability clay deposits impede the downward migration of groundwater contaminants and enhance lateral migration along the upper surface of the clay in the downdip direction. Localized pockets of groundwater and non-aqueous phase liquid (DNAPL) have been identified in sand and silt lenses within the clay matrix to depths of 30 m below land surface [Rich, 1995]. Even though the permeability of the Alpine Formation is relatively low, these sand seams may allow for enhanced fluid



migration in their primary direction of orientation.

The concern at the study site was deep-migrating DNAPL that was detected in the aquifer system. The DNAPL moved downward to the water through the unsaturated zone within a few days. DNAPL continues moving below the water table through the high K zones until it rests on a sufficiently low K layer. The preferential paths of high K represent a very crucial factor in the continuation of the DNAPL migration to the deep aquifer. The identification of such paths is then seriously important for DNAPL removal and isolation. Also, it is important for assessing the potential of DNAPL migration to the deep aquifer for the well head protection program.

The site is bounded by 570000 m, 570280 m east, and 89125 m, 89430 m north. Forty-four irregularly spaced boreholes were available for this application. Borehole depths ranged from 5 to 36 m. The subdomain used for investigation has coordinate axes rotated  $17^\circ$  clock wise with easting, and of origin 570150 m east, and 89160 m north. A layout for the borehole locations as well as the subdomain is shown in Figure 6-12. The KNN method, presented in Chapter 5, is applied to simulate such a domain using the available borehole data. Zones with soil permeability higher than silt are considered high conductivity zones. We refer the reader to Chapter 5 for more details, and access to several realizations generated using the KNN technique. Figure 6-13 shows one realization that is used for the preferential path identification problem. Such a realization has dimensions of 108, 195, and 10 m and divided into  $36 \times 65 \times 30$  pixels of dimensions:  $3 \times 3 \times .33$  m in x, y, and z directions, respectively. The average porosity is assumed .25 for all types of soils. The hydraulic conductivity K is assumed  $2.5 \times 10^{-2}$  cm/sec for sand and  $2.5 \times 10^{-5}$  cm/sec for clay. The quantity  $K/\phi$  is thus  $10^{-1}$ ,  $10^{-4}$  cm/sec for sand and clay, respectively.

### **Preferential Pathway Identification**

The annealing algorithm was applied to the simulated image shown in Figure 6-13. This figure shows an area of high hydraulic conductivity soils near the bottom of the region. We are interested in a possible interconnection between the upper unconfined aquifer (Provo Formation) and thus a high hydraulic conductivity area. Preferential pathways with short travel time are searched for. For demonstration purposes, a line source is considered at the top along the  $x$  axis at  $y=0$ . (See Figure 6-13.) The destination point D is located near the bottom (at elevation = 1m.) within the high hydraulic conductivity area (see Figure 6-13). Details about the results of this application are provided below.

### **Results**

In this application, we present 1) analysis of the simulated annealing process in terms of the variation of the cost function with the control parameter; 2) the probability density functions, pdf, of preferential pathways leaving points 1, 2, 3, 4, 5, and 6 along the source line towards point D; 3) a realization of such preferential paths; and 4) several realizations of the preferential path 1-D. Details about the results are provided below.

At the end of each control parameter step, the final cost and the average cost are recorded. Figure 6-14 shows a plot for the final cost with the control parameter where a large local variation of the final cost with a gradually varying trend is observed. Figure 6-15 shows a similar trend for the average cost, but with less local variation. This is obvious because the averaging process results in less local variability. In general, the local variability decreases with the decrease of the control parameter as the system becomes less chaotic. Most of the solution improvement is observed at  $C > 1000$ . This shows the significance of the controls parameter initial value in the contribution of solution improvement.

The simulated annealing algorithm is used to generate six sets of 100 realizations of optimal pathways shown as 1-D, 2-D, 3-D, 4-D, 5-D, and 6-D in Figure 6-17. The corresponding probability density function, pdf, for each set is estimated and presented in Figure 6-16. The first observation is that most of the pdf's exhibit multimodality, indicating different collections of preferential paths, where each collection belongs to a narrow range of travel time, and possibly unique zones of preferential paths. For a given pdf, the highest mode corresponds to the shortest travel time, indicating a high tendency of flowing through paths of the shortest travel time. There are four major modes (a, b, c, and d) observed in the pdf's and they correspond to  $t=2, 75, 120$ , and  $150$  days, indicating four major collections of preferential pathways between the points along the line source and point D. Note that passing through one or two clay pixels significantly increases the travel time.

For path 1-D, mode "a" is observed with a high frequency but modes "b," "c," and "d" and other modes are observed with very low frequencies. This indicates the likely presence of a preferential pathway with a very low flow resistance between points 1 and D. For path 2-D, mode "a" is observed at a relatively less frequency, but modes "c" and "d" are observed at much higher frequency (compared to those observed at point 1). A realization of preferential pathways connecting the six points to point D is shown in Figure 6-17. Also, 10 realizations of optimal paths connecting point 1 to point D are also shown in Figure 6-18. Nine of these paths have travel time ranges between 650 and 750 days/gradient.

### Discussion and Conclusions

The main objective of the work presented was to identify preferential pathways and to provide a measure for the travel time between two locations in dual porous media. The

KNN method was first used to provide simulated image of the aquifer system. The simulated annealing method was then utilized to search for the least resistant path (i.e., path with minimum travel time) between an origin and destination. The concept of travel time presented here reflects the effective hydraulic conductivity within such a connected path. Application to a control situation demonstrated the efficiency of the simulated annealing method in identifying preferential pathways in a dominantly clayey environment with a thin sand layer connected across the site. The results obtained for one realization of the Ogden Valley aquifer reveals possible interconnection between the upper and lower aquifers, depending on the existing vertical and horizontal flow gradients, with very short travel time. The travel time estimated here is a relative measure of the real travel time, which can indicate the risk level of the contaminant potential.

The identification of such pathways directly improves groundwater monitoring programs by providing more information about potential movement of contaminants. The WHPA can be reevaluated based on investigation of the minimum travel time between existing sources of contaminant and the outside boundary of such areas. Also, the selection of new well locations may be based on maximizing the WHPA, whose outer boundary meets a prescribed minimum value of the travel time from an existing line source.

The algorithm developed in this study shows that simulated annealing is an attractive tool and promising technique in addressing the interconnectivity issue in heterogeneous porous media. The flexibility of this model allows incorporating other aquifer representations based on geologic and/or geostatistics models. Also, it allows the investigation of preferential pathways and true travel time if the velocity field can be computed at the three-dimensional lattice.

Like most simulated annealing applications, the algorithm developed in this study needs a sensitivity analysis of parameter specifications. Our experience with the control situation

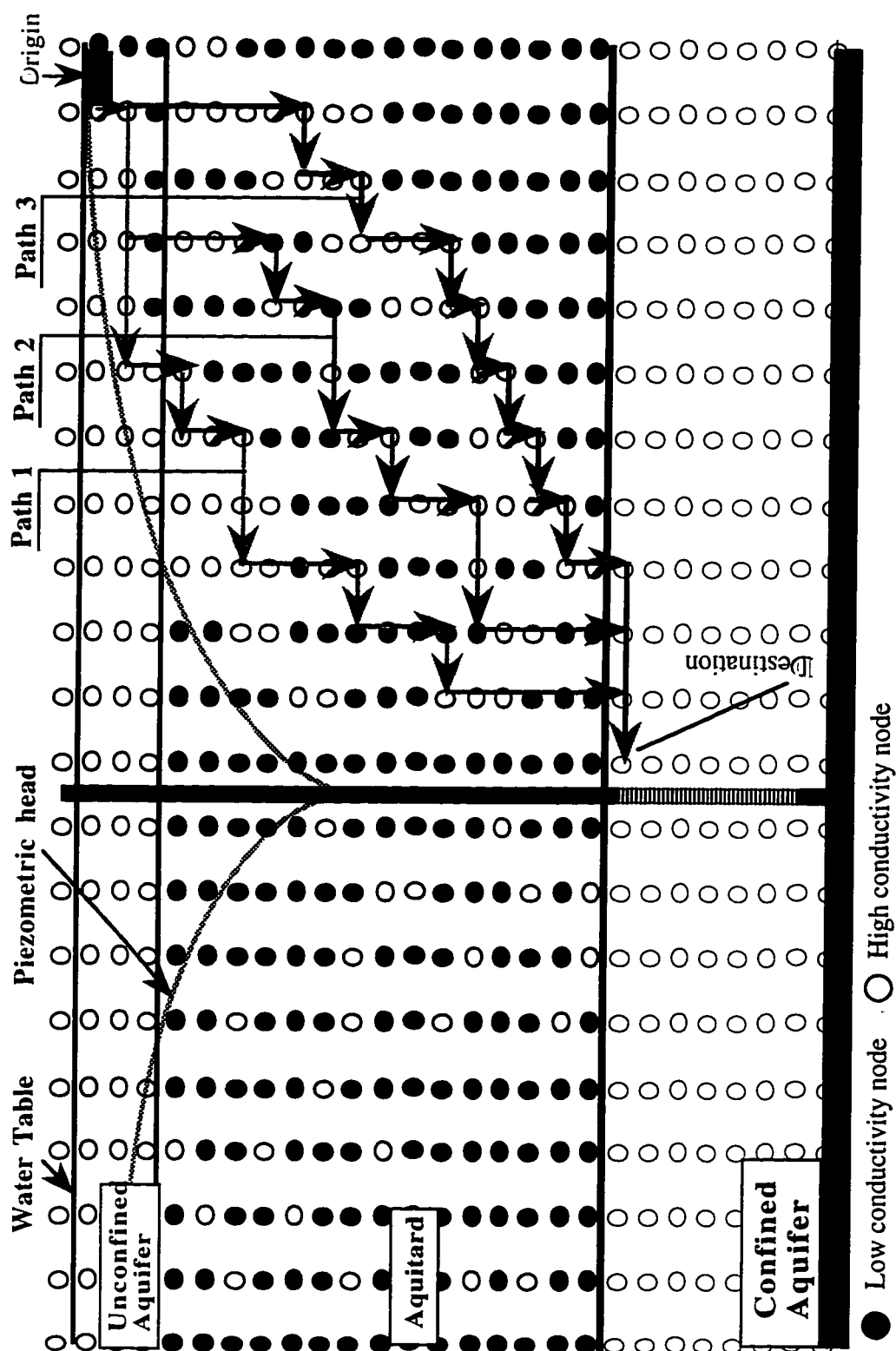
showed that the solution may get stuck in a local minimum if the initial value of control parameter is below a certain value. The multimodality observed in all pdf's indicates the system tendency to get stuck in some local minima with higher travel times. Such a local minimum can occur if the system gets stuck in one or two clay pixels. If the system energy is still high, it may get out of such a pixel; otherwise, it will remain in it, and may accumulate more clay pixels until it gets frozen. The range between the shortest and highest minimum travel times indicates that the system gets stuck in a few clay pixels. The physical explanation of a local minimum in this context is that the value of the hydraulic gradient, or the flow energy, may not be high enough to drive the flow particle away from such clayey zones. This is particularly true if the nearest high conductivity location cannot be reached by a perturbation with a certain length.

To make better parameter specifications in all simulated annealing applications, the gap between the theoretical and practical aspects should be narrowed. A better correlation between such aspects and the physics of the problem is needed. For example, the difficulty of a prior estimation the length of Markov chain leads to an unknown annealing schedule. Besides, the user cannot fully control the variation of the control parameter during the annealing progress. In our problem, a better understanding of the flow particle through dual porous media is useful to control the system energy according to the expected flow behavior during the annealing process

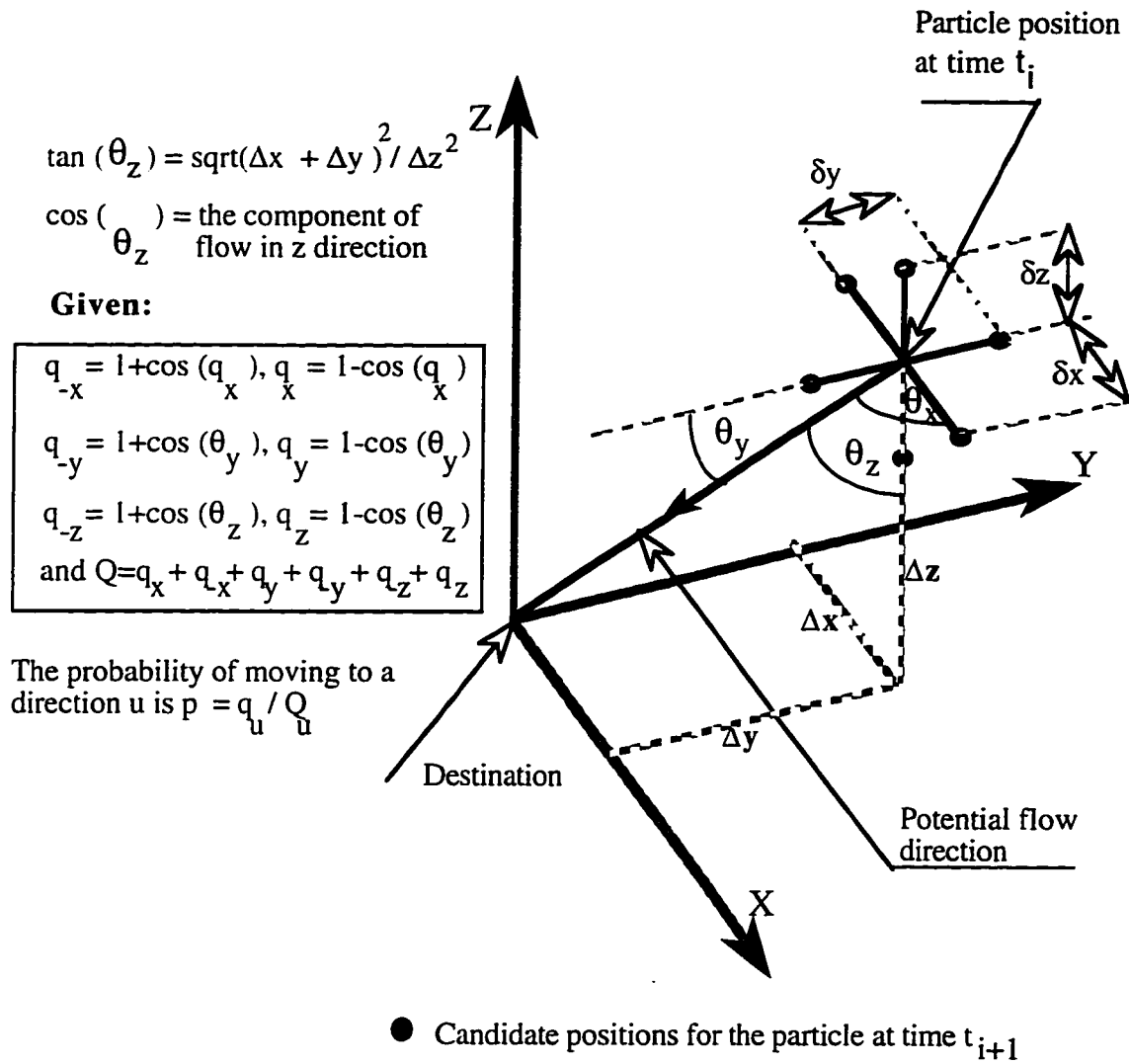
#### References

- Fogg, G. E., Groundwater flow and sand body interconnectedness in a thick, multiple-aquifer system, *Water Resour. Res.*, 22(5), 679-694, 1986.
- Hestir, K., and J. C. Long, Analytical expressions for the permeability of random two-dimensional Poisson fracture networks based on regular lattice percolation and equivalent media theories, *J. Geophys. Res.*, 95(B13), 21565-21581, 1990.
- Johnson, N. M., and S. J. Dreiss, Hydrostratigraphic interpretation using indicator geostatistics, *Water Resour. Res.*, 25(12), 2501-2510, 1989.

- Journel, A. G., and F. G. Albert, Focusing on spatial connectivity of extreme-valued attributes: Stochastic indicator models of reservoir heterogeneities, *Soc. Pet. Eng.*, 18324, 621-632, 1988.
- Kirkpatrick, S., Optimization by simulated annealing, Quantitative studies, *J. Stat. Phys.*, 34(5/6), 975-986, 1984.
- Rich, N., DLOG3D application at operable unit 2 Hill Air Force Base, M.S. thesis, 93 pp., Utah State University, Logan, 1995.
- Schafer-Perini, A., and J. Wilson, Efficient and accurate front tracking for two-dimensional groundwater flow models, *Water Resour. Res.*, 27(7), 1471:1485, 1991.
- Scheibe, T. D., and D. L. Freyberg, Understanding the impacts of spatial structure of natural porous media on the modelling of solute transport in groundwater, paper presented at Fifth Annual Canadian/American Conference on Hydrogeology, National Well Water Association, Calgary, Alberta, Canada, Sept. 18-20, 1990.
- Silliman S. E., and A. L. Wright, Stochastic analysis of paths of high hydraulic conductivity in porous media, *Water Resour. Res.*, 24(11), 1901-1910, 1988.
- Smith, L., and F. W. Schwartz, An analysis of the influence of fracture geometry on mass transport in fractured media, *Water Resour. Res.*, 20(9), 1241-1252, 1984.
- Webb, E. K., and M. P. Anderson, Simulation of preferential flow in three-dimensional, heterogeneous conductivity fields with realistic internal architecture, *Water Res. Res.*, 32 (3), 533-545, 1996.



**Figure 6-1.** Hypothetical flow paths from a contaminant source in an unconfined aquifer to a pumping well in an underlying confined aquifer.



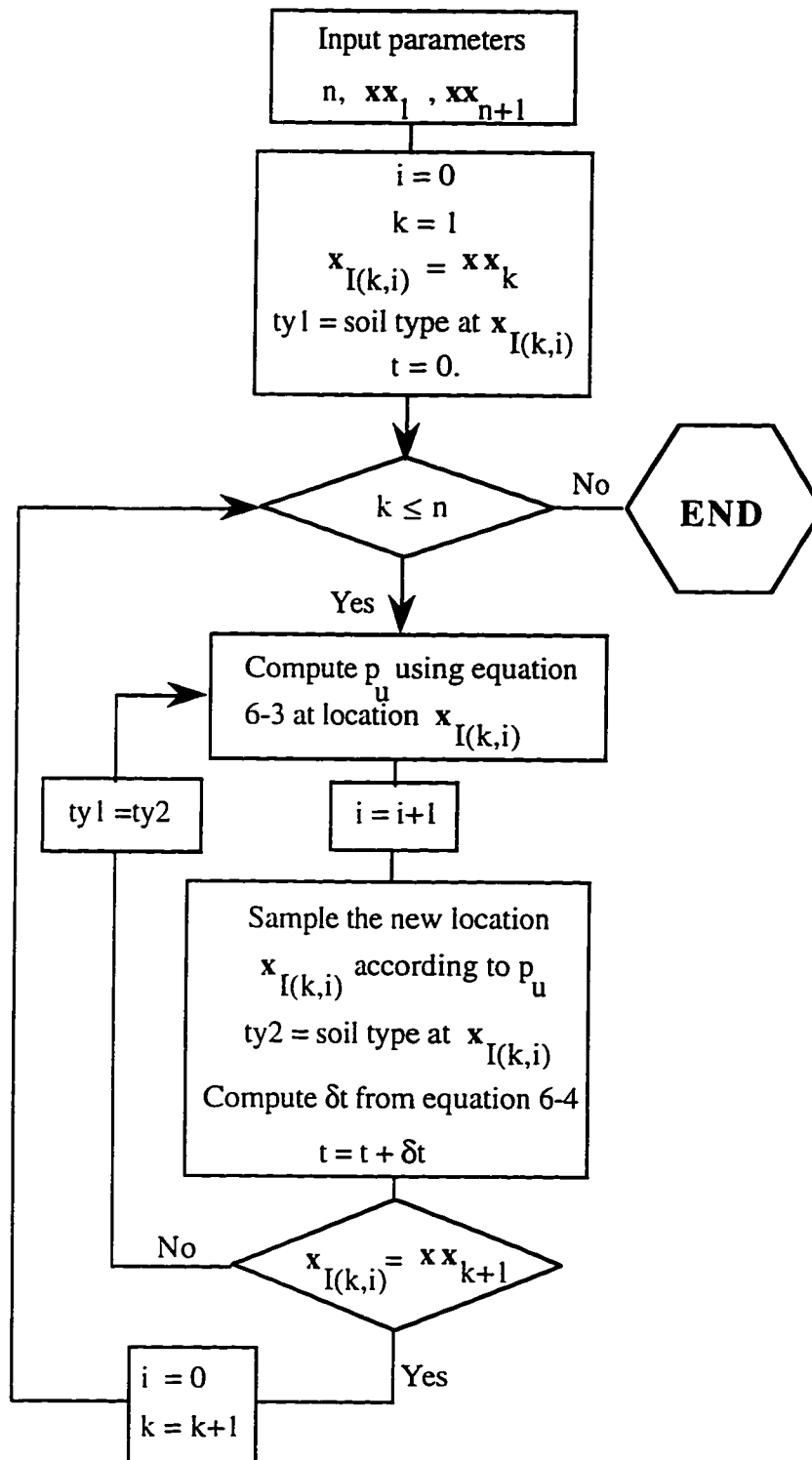
**Figure 6-2.** Assignment of probabilities for the generation of a candidate path.



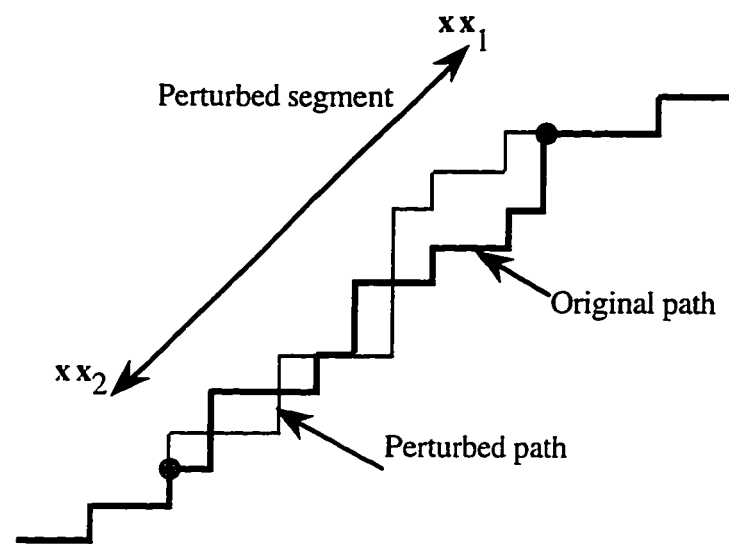
INPUT DATA
------------

- 1) Three-dimensional grid coordinates for aquifer ( $n_x \times n_y \times n_z$ )
- 2) Origin "O" and Destination "D" locations
- n = Number of path segments connected between
- 3) O and D through (n-1) randomly selected points
- 4)  $\mathbf{xx}$  = locations of the origin, destination and the n-1 points
- 5) Number of initially generated perturbations of the initial path to obtain an initial value for the control parameter C
- 6) Decrement coefficient for C
- 7) Decrement coefficient for perturbation length
- 8)
  - (  $K_x, K_y, K_z \phi$  ) for high conductivity soil
  - (  $K_x, K_y, K_z \phi$  ) for low conductivity soil

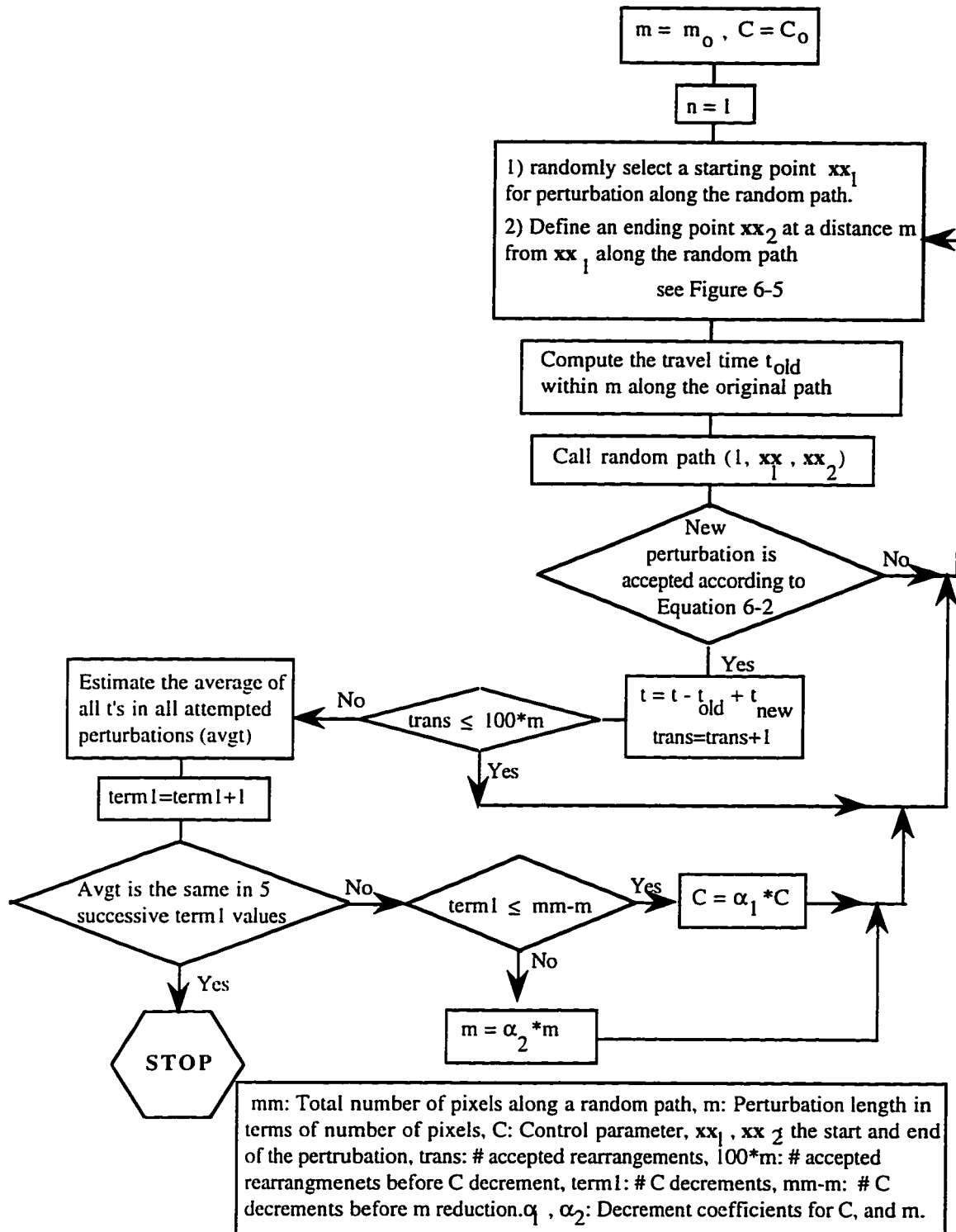
**Figure 6-3.** Input data for simulated annealing algorithm to identify a preferential pathways.



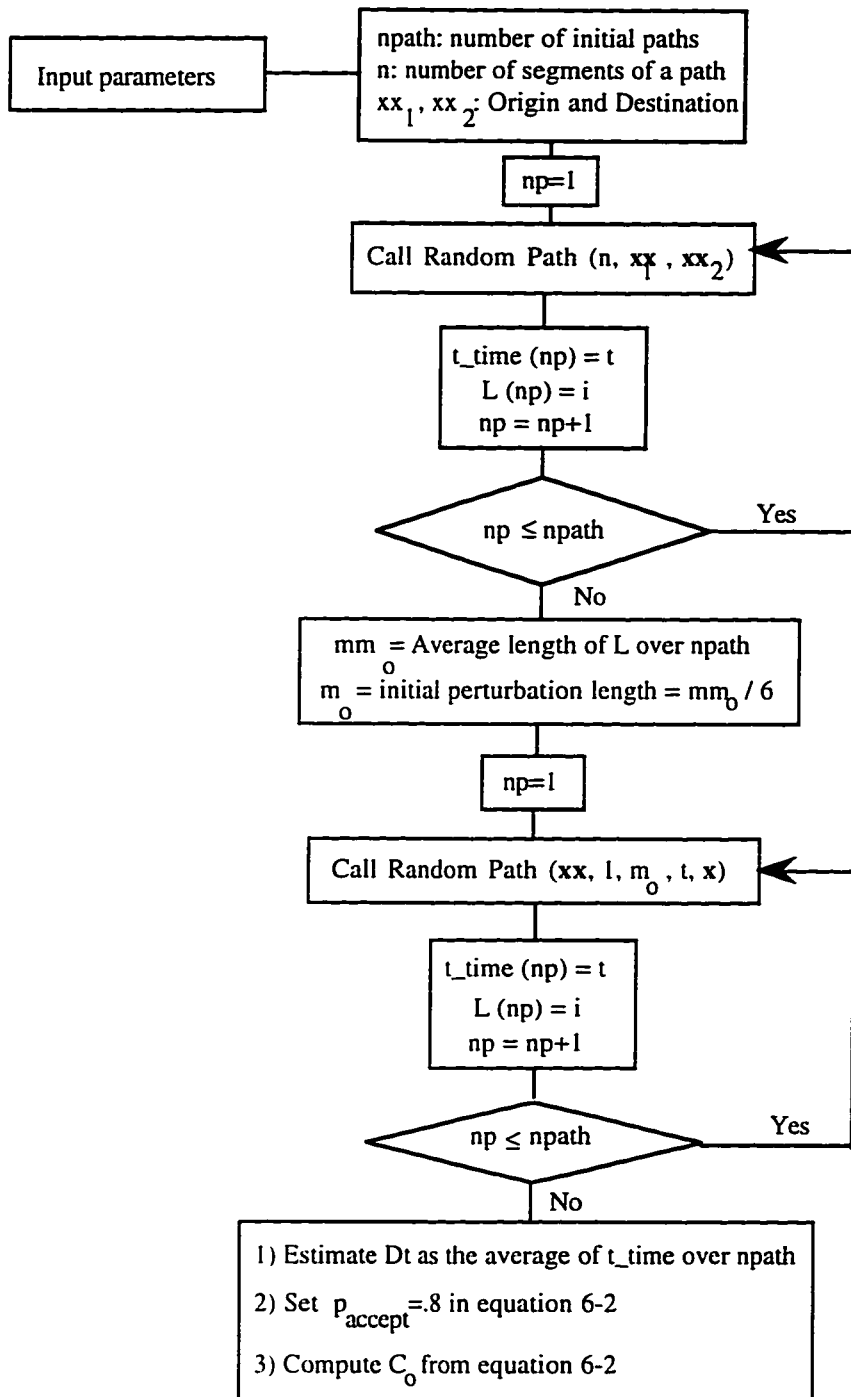
**Figure 6-4.** Procedures for constructing a connected random path between two points.



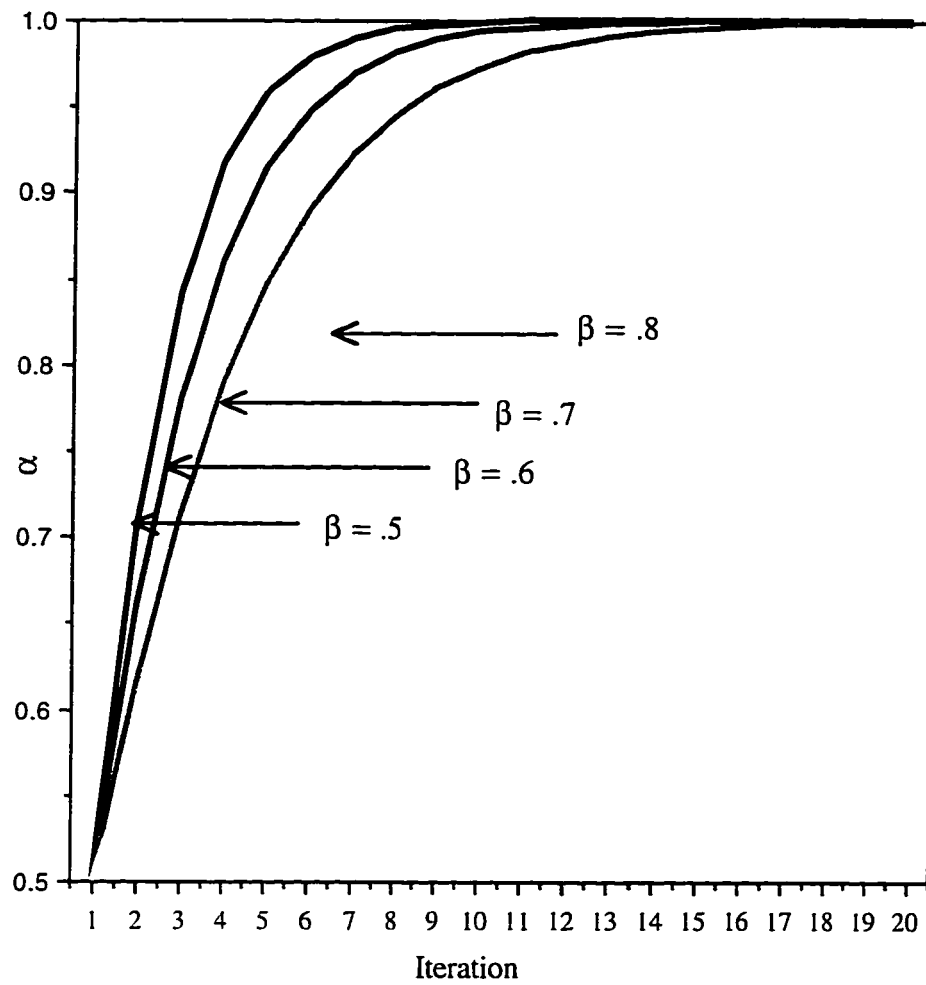
**Figure 6-5.** Local perturbation performed on the original path.



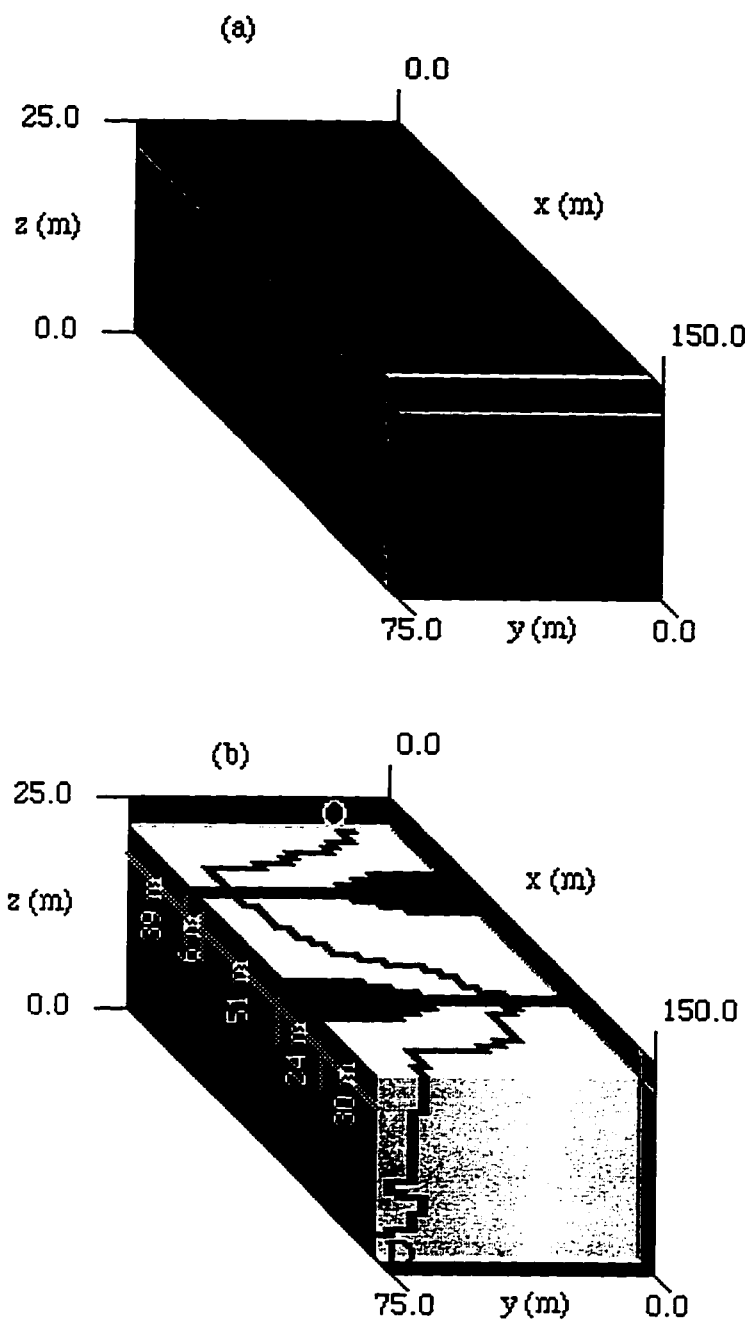
**Figure 6-6.** Procedures for simulated annealing in identifying preferential pathways and the associated travel times.



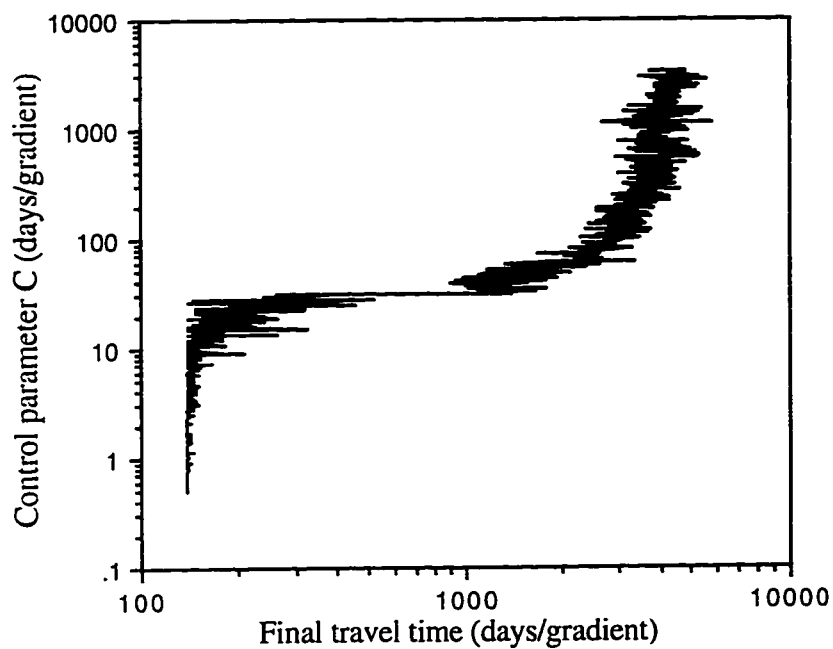
**Figure 6-7.** Procedures for estimating initial values of the control parameter and perturbation length.



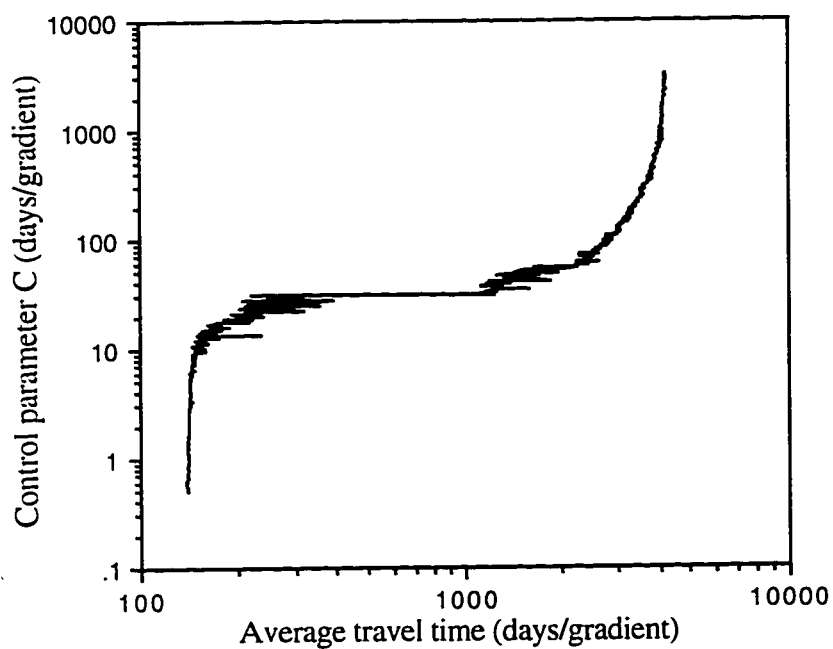
**Figure 6-8.** Decrement coefficient distribution with the number of iterations.



**Figure 6-9.** Two images of a hypothetical setting of sand/clay in porous media. The light color zones indicate sand, dark color zones indicate clay, and the gray line represents a preferential pathway (line with minimum resistivity).

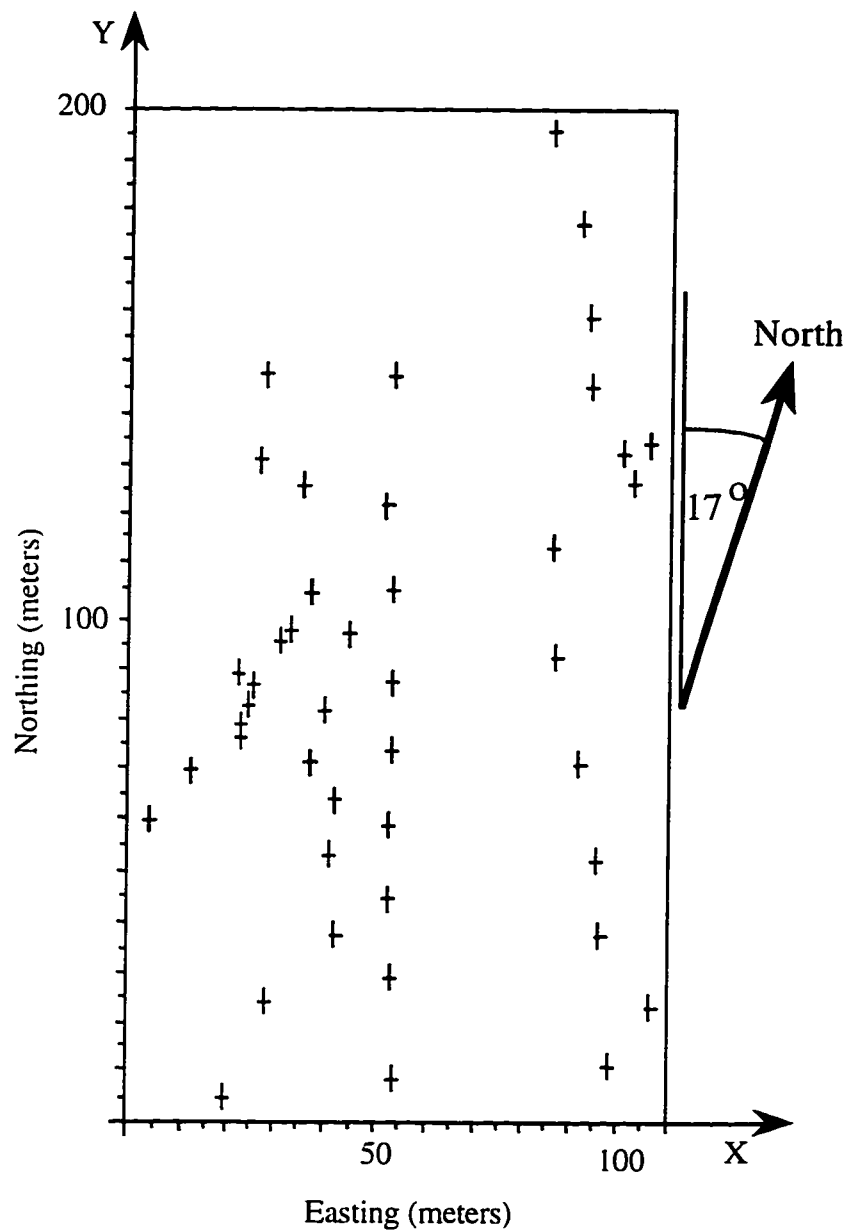


**Figure 6-10.** Control parameter, C, versus travel time at the end of each C step (for the setting in Figure 6-9).

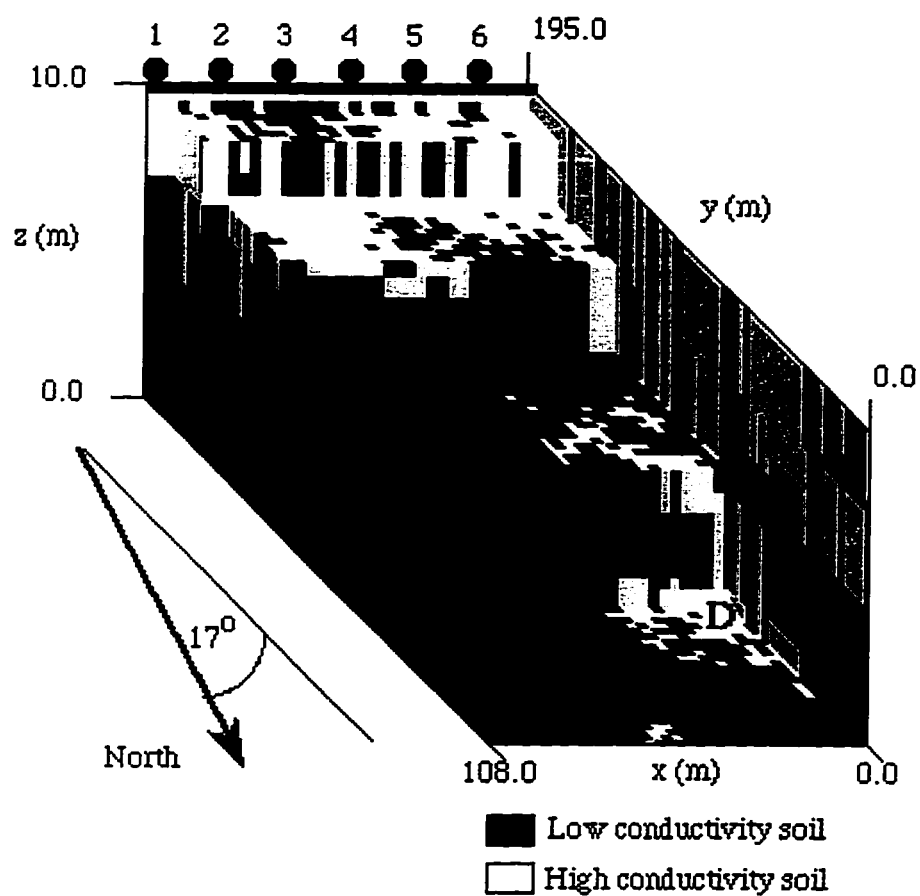


**Figure 6-11.** Control parameter, C, versus average travel time within each C step (for the setting in Figure 6-9).

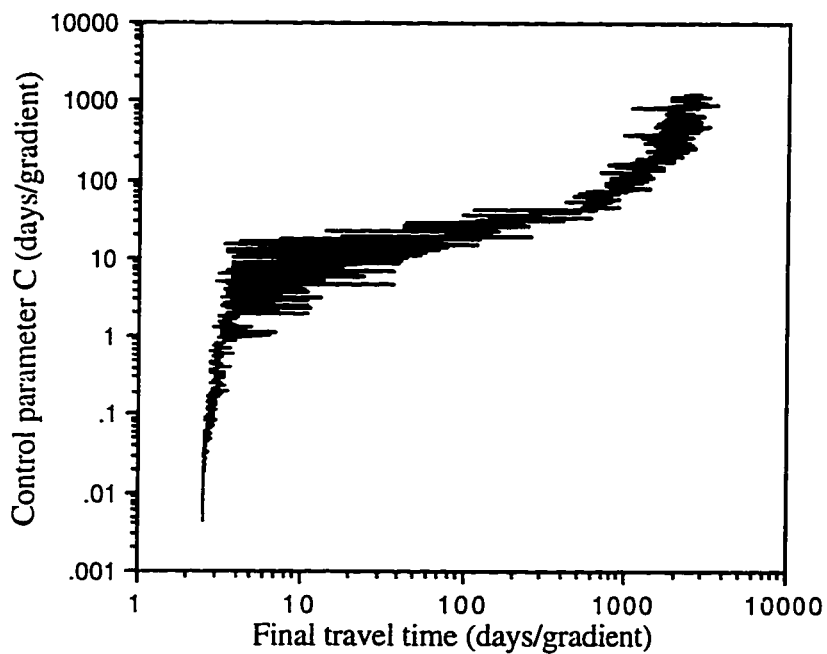




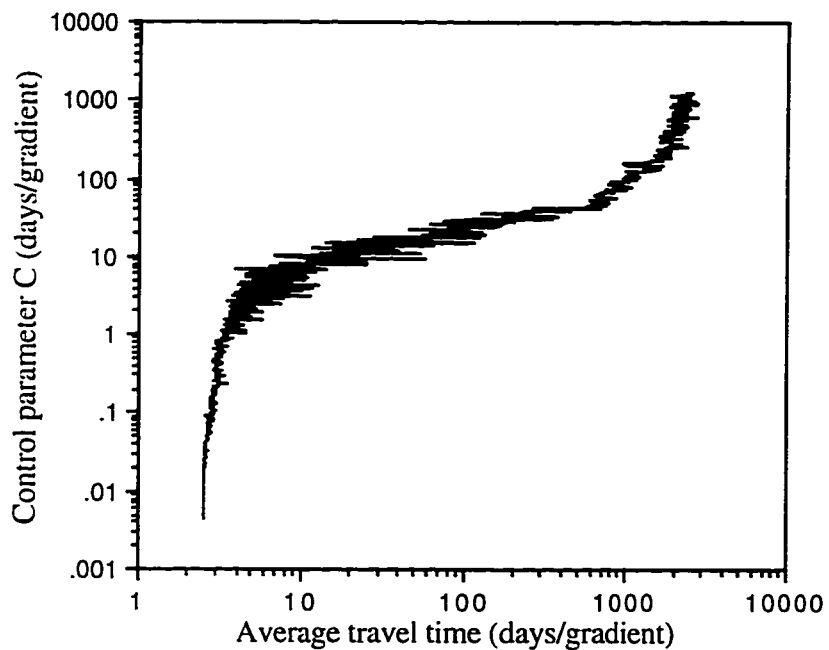
**Figure 6-12.** Plan view of the bore hole data layout and locations of estimate. Origin is located at 1870560 m (east), 292510 (north). Coordinate axes are rotated  $17^\circ$  counterclockwise.



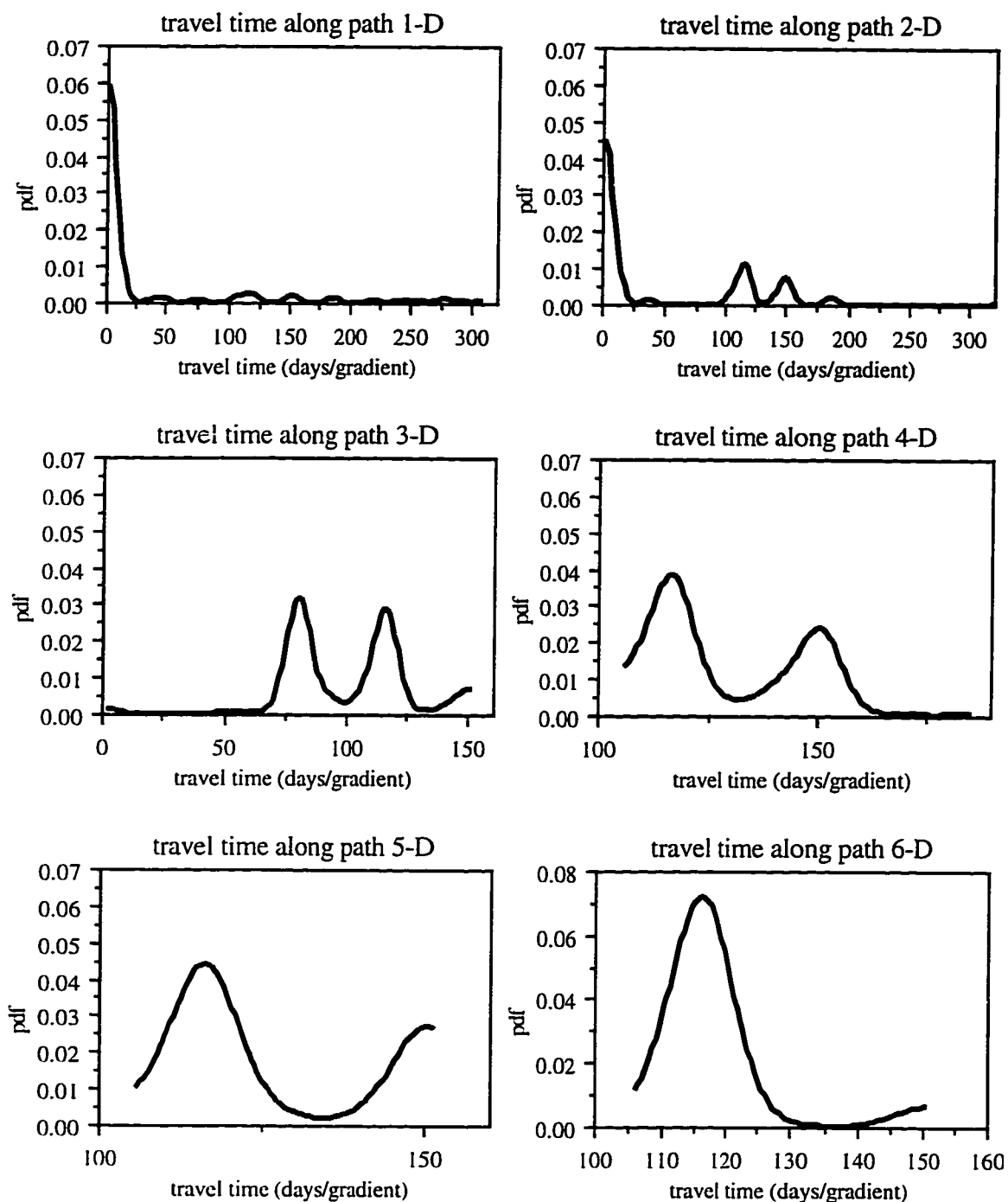
**Figure 6-13.** A realization of the KNN method applied to bore hole data from the Ogden Valley aquifer. This realization is used for the preferential pathway identification analyses between six points along the source line and point D.



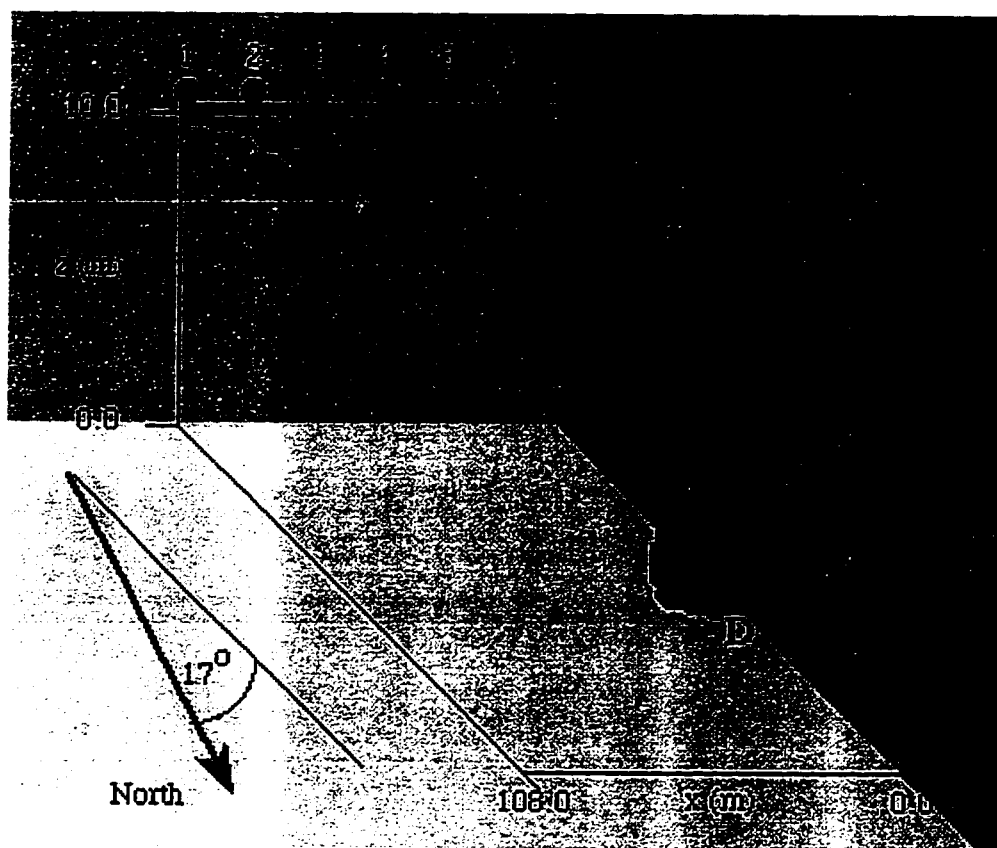
**Figure 6-14.** Control parameter,  $C$ , versus travel time at the end of each  $C$  step (for the Ogden Valley aquifer application).



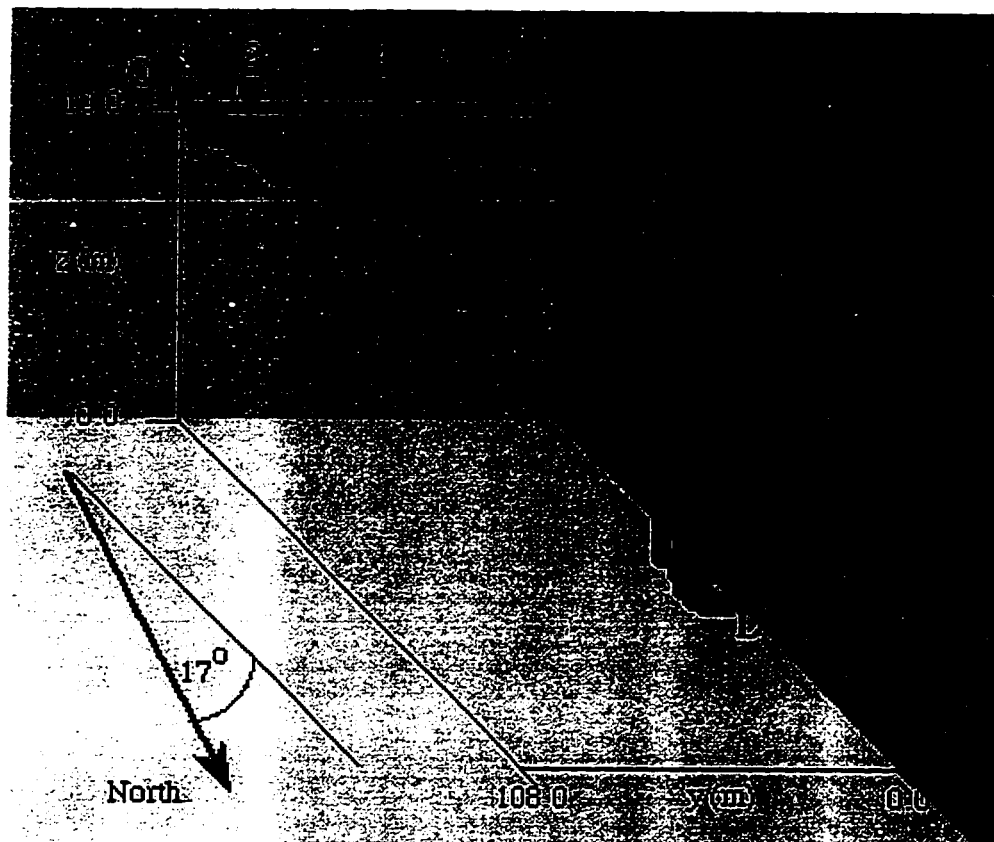
**Figure 6-15.** Control parameter,  $C$ , versus average travel time within each  $C$  step (for the Ogden Valley aquifer application).



**Figure 6-16.** Probability density functions of the travel time distribution between six points of origin along a line source and destination point D (see Figure 6-13).



**Figure 6-17.** Preferential pathways between six points of origin along a line source and destination point D.



**Figure 6-18.** Ten realization for optimal pathways between points O and point D.

## CHAPTER 7

### GENERAL SUMMARY

In this study, the main goal was to address the question: Given drill log data, how can a nonstationary subsurface environment be characterized?, and consequently, how can the contaminant potential be assessed? To answer the first question, nonparametric procedures were developed for subsurface characterization and simulation. These procedures are data-driven methods in which data have a direct and significant role in the estimation. The procedures avoid prior assumptions as to the kind of distribution and the dependence structure of the underlying probability models. To answer the second question, a simulated annealing technique was used as part of an optimization model for preferential pathways identification and travel time estimation.

In Chapter 2, a new method for characterizing subsurface geology using kernel methods with borehole data was presented. A product of this work is a probabilistic image that can provide insight into subsurface heterogeneities. The methodology is particularly useful where the environment is the outcome of nonstationary sedimentation process. No discretization of the spatial domain beyond the resolution of the available data is needed. The nearly continuous information in the vertical at each borehole is treated differently from and integrated with the sparse information in the horizontal. In these respects the methodology presented improves on Indicator Kriging, which is useful in the same context. Applications to a control situation and to a real data set demonstrated the efficacy of the kernel estimator in nonstationary situations, and the efficacy for identifying possible heterogeneities in the subsurface from scattered borehole data.

The model developed can be extended to provide realizations of the aquifer heterogeneity. The idea is to sequentially simulate all soil pixels of a three-dimensional

grid. At simulation step  $i$ , the model is applied to find the probability using all real data and the  $i-1$  simulated pixels. This probability is then used to sample soil pixel at step  $i$ . Such a procedure requires some modifications of the original model, including the method of selecting the model parameters (bandwidths). Also, an algorithm for preserving the vertical and lateral continuity is needed.

Chapter 3 presented a methodology for modeling stratigraphic sequences using a continuous parameter, homogeneous semi-Markov model. The simplicity and nonparametric flavor of the discrete parameter Markov chain models are retained, while problems and loss of information stemming from the sampling discretization in the discrete approach are circumvented. The applications presented demonstrated the utility of the method in preserving the statistical attributes of the stratigraphy as seen from the boreholes. The simulator presented should be useful in a number of contexts. One may be interested in testing hypotheses related to cyclicity in deposition at a site. The CHSM Monte Carlo simulations could be used to test the significance of any such claims or the adequacy of a physically or conceptually based model of deposition relative to the sampling of the environment using boreholes. This model functions properly in a statistically stationary stratigraphic environment. Simulated bore holes do not preserve the spatial location of lithologic sequence.

Chapter 4 extended the model presented in Chapter 3 to consider nonstationarity of the deposition process in the vertical. A nonparametric method was used to compute the unconditional probabilities and transition intensities for multiple types of soils from a well log. A semi-Markov framework is used to generate the pseudo-well logs. The parameters of this process are allowed to vary smoothly along the profile in the vertical. The applications presented showed the model utility in preserving the statistically nonstationary attributes as estimated from data. Two procedures for resampling the thickness of the beds



for each soil type were investigated. For the application presented, resampling from the observed bed thicknesses that lie within the window used for estimating local probabilities and transition intensities worked better than assuming the exponential distribution. The nonhomogeneous semi-Markov model represents a significant improvement compared to the homogeneous semi-Markov model. In the nonhomogeneous model, simulated bore holes are more visually consistent, with bore holes sampled from a nonstationary stratigraphic environment, than those generated by the homogeneous model.

In Chapter 5, a method for generating likely realizations of subsurface soils by bootstrapping  $k^{\text{th}}$  nearest drill logs was developed. A probability mass function (Kernel function), equation 1, was used to resample the nearby drill logs where decreasing probabilities are assigned to farther neighbors. Application to a control situation demonstrated the efficacy of the KNN resampler in reproducing the probability distribution of a given soil type. However, the probability images obtained are more rough than similar images obtained by kernel estimator in Chapter 2 due to the unsmoothed variation in the vertical. The results of Ogden Valley aquifer show that the KNN resampler preserves the unconditional probability of bed thickness and the transition intensity in the vertical from sand to clay and from clay to sand. Also, observation at drill logs along a certain line shows reasonable consistency with the horizontal trend in the probabilistic image. One of the problems encountered in the implementation of the KNN simulation model is the possibility of having drill logs that start at different elevations and are of different lengths. This is a problem that may be faced if there are significant terrain variations, and the sampling program is not designed specifically for the investigation. Changes in elevation may be accommodated by moving the top of the resampled drill log to match either 1) the surface elevation of the lattice point, or 2) the general dip and strike at the site as one moves from the original location to the lattice location. The KNN simulator is computationally

fast. Typical cpu times to generate a realization with 70200 lattice nodes and 2340 drill logs were of the order of 30 seconds on a DEC 3000 computer. The generation of a large number of realizations for analysis is thus feasible.

Chapter 6 presented a simulated annealing-based approach for preferential pathways identification and travel time estimation in three-dimensional heterogeneous porous media. Application to a control situation demonstrated the efficiency of the simulated annealing method in identifying preferential pathways in a dominantly clayey environment with a thin sand layer connected across the site. Application to the Ogden Valley aquifer showed sensitive areas of preferential paths with a very short travel time. The difficult part in this technique is a proper selection of the control parameters. A future work should focus on the theoretical aspect of simulated annealing to develop a robust strategy for control parameter specification. Also, the model developed assumes a high contrast of hydraulic conductivity and, hence, ignored the variation in the head field. In many environments, this may not be a valid assumption. A future work should incorporate the flow conditions in the simulated annealing process.

## APPENDICES

## APPENDIX A

## CURRICULUM VITAE

Alaa Ibrahim Ali  
(Oct., 1996)

Research Interest Geostatistics, Groundwater Hydrology , Geotechnical Engineering, Geo-environmental Engineering, Stochastic Subsurface Hydrology, and Stratigraphic Characterization.

### Education

Ph.D. in Civil and Environmental Engineering with an emphasis in Geostatistics and Groundwater, from Utah State University, Logan, Utah, 1996.

M.S. in Civil and Environmental Engineering with an emphasis in Geotechnical Engineering from Utah State University, Logan, Utah, 1992.

Pre-M.S. in Civil Engineering, with an emphasis in Hydrology, from Cairo University, Cairo, Egypt, 1989.

B.S. in Civil Engineering from Cairo University, Cairo, Egypt, 1987.

### Experience

Research Assistant, Utah Water Research Laboratory, Utah State University, Logan, Utah (1992-present), Developed series of nonparametric stochastic techniques for nonstationary aquifer system characterization and simulation. Models were validated against synthetic and real data. These models have been used for extensive investigations of several aquifers in several counties in Utah. Used simulated annealing technique to develop an optimization model for transport preferential pathways identification and travel time estimation. This model is used to assess the risk of contaminant migration from disposal sites to water supply well locations.

Research Assistant, Geotechnical Division, Utah State University, Logan, Utah (1990-1992). Conducted field scale-laboratory tests including instrumentation, data logging and analysis for earth embankment reinforcement. Developed a mathematical model for Pullout resistance for reinforced soil embankment walls. Conducted variety of soil mechanics tests such as: consolidation, direct shear, triaxial, unconfined compression, compaction, and in situ measurements by nuclear method.

### Selected Publications, Conference presentations

#### Publications, and Research Reports

Ali, A. I., and U. Lall, Identifying potential preferential paths for subsurface transport using simulated annealing, to be submitted to *Water Resources Research*, 1996.

Ali, A. I., and Lall U., A Kernel estimator for stochastic subsurface characterization from drill log data, submitted to *Ground Water*, 1996.

Ali, A. I., and U. Lall, A continuous parameter semi-Markov model for stratigraphic analyses from well log data, submitted to *Log Analyst*, 1996.

Ali, A. I., and U. Lall, A nonhomogeneous continuous parameter semi-Markov model for stratigraphic analyses from well log data, submitted to *Mathematical Geology*, 1996.

Lall, U., A. Ali, A K-nearest neighbor Simulator for pseudo-bore hole logs for subsurface characterization, to be submitted to *Water Resources Research*, 1996.

Ali, A. I., and Lall, U., Stratigraphic interpretation from drill log data, Kluwer Academic Publisher, Ontario, Canada, 1994.

Ali, A. I., Anderson L. R., Sampaco C. L., Womack K. C., and Christiansen V. T., Pullout capacity of flat-faced panel anchors for RSE walls, 28th Symposium on engineering geology and geotechnical engineering, Boise, Idaho, 1992.

Ali, A. I., and Lall, U, Stochastic characterization of aquifer heterogeneity from drill log data, Stochastic and statistical methods in hydrology and environmental engineering, international conference: University of Waterloo, Ontario, Canada, (1993).

Ali, A. I., and Lall, U., Stochastic simulation of aquifer heterogeneity from drill log data using kernel method, Nato-ASI, Recent Advances in Groundwater Pollution Control and Remediation, Antalya, Turkey, 1995.

Ali, A. I. and Lall, U., Interpretation of drill log data: DLOG3D-A probabilistic tool for analyzing subsurface soil variability No. report RR-93-HWR-UI/001. Utah Water Research Laboratory, Utah State University , Logan, UTAH, 1993.

Lall, U., and Ali, A. I., Nonparametric stratigraphic interpretation from drill log data (United States Geological Survey 104 Project completion Report No. United States Geological Survey, Utah Water Research Laboratory, Utah State University, Logan, UTAH, 1992.