**Bayesian Approaches for Subsurface Characterization Using Hydrogeological and Geophysical Data**

by

Jinsong Chen

B.E. (Tsinghua University, China) 1988
B.S. (Tsinghua University, China) 1988
M.E. (Tsinghua University, China) 1990
M.S. (Georgia Institute of Technology) 1997
M.A. (University of California, Berkeley) 2001

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering-Civil and Environmental Engineering

in the

GRADUATE DIVISION
of the
UNIVERSITY of CALIFORNIA, BERKELEY

Committee in charge:

Professor Yoram Rubin, Chair
Professor Rodney J. Sobey
Professor Charles Stone

Fall 2001

The dissertation of Jinsong Chen is approved:

Chair Date

Date

Date

University of California, Berkeley

Fall 2001

**Bayesian Approaches for Subsurface Characterization Using Hydrogeological and Geophysical Data**



by

Jinsong Chen

## Abstract

Bayesian Approaches for Subsurface Characterization Using Hydrogeological and Geophysical Data

by

Jinsong Chen

Doctor of Philosophy in Engineering-Civil and Environmental Engineering

University of California at Berkeley

Professor Yoram Rubin, Chair

Near-surface investigations often require detailed mapping of spatial variability of hydrogeological parameters. Conventional techniques for collecting densely sampled hydrogeological data are costly, time-consuming, and invasive; consequently, cost-effective and noninvasive geophysical methods can be used to provide additional information for the subsurface. This study explores the use of hydrogeological and geophysical data for site characterization within a Bayesian framework, and three case studies are presented with emphases on different aspects of subsurface investigations. The first study explores the use of ground penetrating radar (GPR) tomographic velocity, GPR tomographic attenuation, and seismic tomographic velocity using data collected from the Oyster site (VA) for hydraulic conductivity estimation

using the Bayesian method based on a normal linear regression model. Although the log-conductivity displays a small variation and the geophysical data vary over only a small range, results indicate that the geophysical data improve the estimates of the hydraulic conductivity and the improvement is the most significant where prior information is limited. Among the geophysical data, GPR and seismic velocity are more useful than GPR attenuation. The second study considers a scale disparity problem in which small-scale resistivity logs and large-scale electromagnetic (EM) surveys collected from the Lawrence Livermore National Laboratory site (CA) are used to map two-dimensional resistivity fields using a Bayesian method. Results reveal that the large-scale survey data enhance hydrogeological site characterization even when they have a relatively low resolution. The last study focuses on the issue of petrophysical relations between hydrogeological and geophysical data. A Bayesian model coupled with a fuzzy neural network (BFNN) is developed to alleviate the difficulty of using geophysical data in lithofacies estimation due to non-linearity of cross correlation between lithofacies and geophysical attributes. Results show that the BFNN model is the best method among indicator kriging, indicator cokriging, and the fuzzy neural network without considering spatial correlation.

---

Professor Yoram Rubin
Dissertation Committee Chair

# Contents

# List of Figures

# List of Tables

# Acknowledgements

I would like to thank Professor Yoram Rubin for his enthusiastic guidance as my advisor, and it has been a great pleasure to work with him. I also want to thank Professor Charles Stone, Professor Rodney J. Sobey, and Professor Alex Becker for their valuable and insightful comments on my work.

I wish to thank Susan Hubbard, John Peterson, and Ernie Majer from the Lawrence Berkeley National Laboratory for giving me excellent data to conduct this research. I also want to thank Souheil Ezzedine from the Lawrence Livemore National Laboratory for discussing his research with me when he worked here.

I want to gratefully acknowledge the Department of Energy and the National Science Foundation for providing financial support.

# Chapter 1

# Introduction

## 1.1 Motivation

The release and transport of contaminants from industrial, agricultural or waste disposal activities is a major environmental concern due to the potential for contamination of water supplies and sensitive areas. The remediation of the contaminants requires identification of natural heterogeneity of the geologic formation and detailed mapping of spatial variability of the hydraulic conductivity, for which densely sampled direct measurements of the hydrogeological parameters are needed, especially when the aquifers under investigation are complex. Conventional techniques for collecting hydrogeological data rely heavily on borehole drilling that is invasive, costly, and time-consuming. The use of the method is deemed necessary for collecting detailed information at some representative locations but is impractical for delineating spatial variability of the parameters. Consequently, modeling of heterogeneity of the hydrogeological parameters using hydrogeological measurements alone becomes difficult and subject to a large degree of uncertainty. With poor site characterization, the remediation schemes may be over-designed or even inapplicable and hence unnecessarily

increase the cost of the remediation.

Geophysical methods that are noninvasive, cost-effective, and have a broad spatial coverage along lateral directions can provide a large amount of additional information about the subsurface. These methods measure subsurface physical properties or their contrasts. Relations between hydrogeological and geophysical properties can allow those geophysical data to be used qualitatively in reconnaissance studies for identification of aquifer boundaries and stratigraphy or to be used quantitatively at the advanced stage of site characterization for estimation of hydrogeological parameters. Traditional methods for collecting hydrogeological data include core point measurements and volume-averaged pumping tests. The core measurements have high resolution and a small supporting volume whereas the pumping tests have low resolution and a large supporting volume, as shown in Figure 1.1. Geophysical methods, however, often provide data with scales between the core measurements and the pumping tests. As a result, geophysical data bridge the information gap between the two traditional measurements in terms of resolution and volume of aquifer sampled. With the use of two- and three-dimensional geophysical data, a better characterization becomes possible, and this may lead to a significant reduction of total cost for in-situ remediation of contaminants.

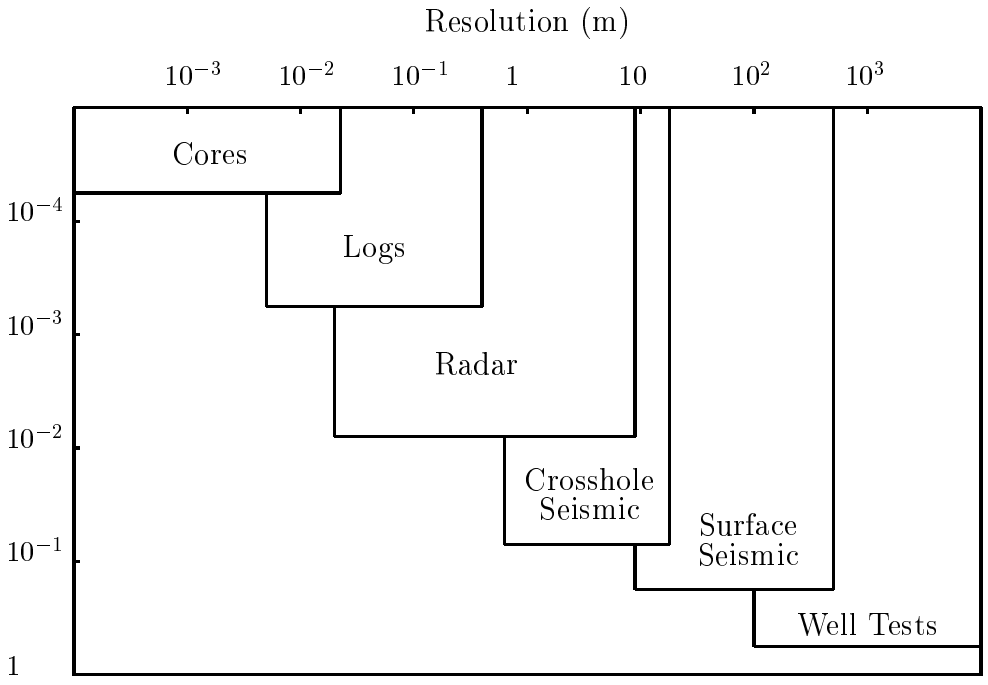


Figure 1.1: Comparison between resolution and volume of aquifer sampled for conventional core and well tests as well as for geophysical techniques. The vertical coordinate is the fraction of sampled aquifer volume [*Rubin et al.*, 1998].

## 1.2 Statement of the Problem

Several problems exist in the use of geophysical data for estimation of hydrogeological parameters. The first one comes from the fact that geophysical methods directly measure geophysical properties, such as electrical resistivity and seismic velocity, instead of hydrogeological parameters, such as hydraulic conductivity and porosity. Petrophysical relations are needed to incorporate the geophysical data into estimation of the hydrogeological parameters. Although most geophysical attributes have physical connections with hydrogeological properties, for unconsolidated sediments, the relations have not always been recognized or well-understood, even in the laboratory. They are often non-unique, site-specific, and affected by many compounding factors, which are very difficult to measure or control in field conditions. Developing data-driven petrophysical models for relating between geophysical data and hydrogeological parameters is one of the goals of this dissertation.

The second problem is due to the scale disparity between hydrogeological measurements and geophysical data and between different types of geophysical data. As shown in Figure 1.1, different types of geophysical measurements often have different resolution and supporting volumes, which are determined by the mechanics and acquisition parameters of the geophysical techniques such as source frequency, acquisition geometry, and sampling spacing. Because of ambiguity which arises from geophysical data interpretation, however, we often need to jointly use different types

of geophysical techniques at a given site. When using the multiple geophysical data for estimation of hydrogeological parameters, we have to take into account the scale differences. Chapter 3 of this dissertation is an effort at solving this problem. We combined the low-resolution electrical resistivity data obtained from crosshole electromagnetic (EM) surveys and the high-resolution electrical resistivity data obtained from borehole logging to generate two-dimensional resistivity images that can be used later to estimate hydrogeological parameters. The scale disparity also makes it difficult to infer petrophysical relations between geophysical attributes and hydrogeological parameters from training data sets.

The third problem results from measurement errors of hydrogeological data and uncertainty associated with acquisition and interpretation of geophysical data. Developing suitable assimilation models to account for the errors and the uncertainty is another goal of the dissertation. In addition, it has been well-recognized that hydrogeological parameters and geophysical attributes usually have good spatial structures. How to incorporate the spatial correlation into estimation of hydrogeological parameters is also an issue addressed in this dissertation.

## 1.3 Review of Previous Work

Geophysical techniques have been used in near-surface investigations for decades and have become increasingly important recently due to the development of high-

resolution crosshole tomography and the commercial availability of surveying techniques. Many efforts have been made to jointly use geophysical and hydrogeological data for site characterization. This section will briefly review some of the previous work along two different lines. Section 1.3.1 reviews some commonly used geophysical methods with the focus on the principal of each method and its applications in environmental site characterization. Section 1.3.2 reviews several assimilation methods used for combining geophysical and hydrogeological data with the focus on the advantages and limitations of using those methods.

### 1.3.1 Geophysical Characterization Methods

The commonly used geophysical methods for site characterization include surface electrical resistivity, electromagnetic (EM) induction, ground-penetrating radar (GPR), seismic methods, and borehole geophysical methods. Each of these methods can be used for different field situations and for different goals of site characterization.

Electrical methods have been used in groundwater investigations for many years. These methods measure electrical resistivity, the ability of electrical current to flow through materials, by inducing a time-varying current (DC) or very low frequency current into the ground between two current electrodes. The measured resistivity can be used qualitatively to map subsurface stratigraphy [*Zodhy et al.*, 1974] and to locate geological structures such as buried stream channels [*Burger*, 1992] because of the sensitivity of the current flow to the presence of geological layers. The resistivity

measurements can also be used quantitatively to infer information about hydrogeological parameters because of the physical connection between electrical resistivity and hydrogeological properties. Electrical conduction usually takes place in fluids in connected pore spaces, along grain boundaries, or within fractures, but not in the matrix of the materials. Electrical resistivity is affected by material texture, grain size, porosity, clay content, moisture content, and the resistivity of pore fluid. All of those factors are related to hydraulic conductivity [*Gassmann*, 1951; *Marion*, 1990; *Klimento and McCann*, 1990; *Knoll*, 1996;]. As a result, electrical resistivity can be used to estimate hydraulic conductivity or transmissivity in porous aquifers [*Kelly*,1977; *Urish*, 1981; *Mazac et al.*, 1985].

Recent applications of electrical resistivity methods are due to the significant progress in the development of crosshole electrical resistance tomography (ETR). This method involves using a crosshole geometry, where electrodes are placed in boreholes and at the surface; the differences in voltage between potential electrodes are then measured. Data acquisition of the method is automated and capable of recording 3500 measurements per hour with only one technician [*Rubin*, 1998]. This efficiency in data collection enables us to use the method to monitor flow and contaminant transport processes [*Alumbaugh et al.*, 2000] and movement of soil water content [*Daily et al.*, 1992; *Ramirez et al.*, 1993; *Daily and Ramirez*, 1995; *Zhou et al.*, 2001].

Electromagnetic (EM) methods have received much attention recently in groundwater and environmental site assessments due to the ability of the methods to detect

conductive objects under the ground. Controlled-source inductive electromagnetic (EM) methods use a transmitter to pass a time-varying current through a coil or dipole on the earth surface. This alternating current produces a time-varying magnetic field that interacts with the conductive subsurface to induce time-varying eddy currents, which give rise to a secondary EM field. Attributes of the secondary magnetic field, such as amplitude, orientation, and phase shift, can be measured by the receiver. By comparing these collected attributes with those of the primary field, we can detect subsurface conductors or distribution of electrical conductivity [*Rubin et al.*, 1998].

Since electrical conductivity is the inverse of electrical resistivity that can be related to hydraulic conductivity, the measured conductivity data can be used to estimate hydrogeological parameters. For example, EM methods have been used to estimate soil water content [*Kachanoski et al.*, 1988; *Sheets and Hendricks*, 1995] and to investigate the spatial variations of soil texture and pore fluid [*Kachanoski et al.*, 1988]. These methods have also been successfully used to detect both organic and inorganic groundwater contamination plums [*Buselli et al.*, 1990]. Because a conductive subsurface environment or target is required to set up the secondary field that can be measured by the receiver, the methods are suitable for detecting high-conductivity subsurface targets, such as salt water saturated sediments or clay layers, but unsuitable for detecting electrically resistive targets.

Ground-penetrating radar (GPR) is a relatively new geophysical tool that has

become increasingly popular due to its high resolution and the need to better understand near-surface conditions. GPR methods use electromagnetic energy at high frequencies (10 to 1000MHZ) to probe the subsurface, and the propagation of the radar signal depends on the electrical properties of the ground at the high frequency [*Davis and Anna*, 1989]. The methods measure the velocity and the attenuation of the radar waves, and these can be used to determine the dielectric constant or relative permittivity, which is the major electrical property of geologic materials at high frequencies. Generally, GPR methods have poor performances in electrically conductive environments, such as saturated systems or in systems dominated by the presence of expanding clays, and have better performances in unsaturated coarse-textured or moderately coarse-textured soil [*Hubbard et al.*, 1997].

GPR methods have found many applications in both saturated environments and unsaturated environments with substantial nonexpanding clay fractions. *Knoll et al.* [1991] used GPR methods to delineate near-surface conditions in a sand and gravel aquifer at Cape Cod site in Massachusetts. *Wyatt et al.* [1996] used the methods to detect shallow faults at the Savannah River site in South Carolina. *Hubbard et al.* [1997] used GPR data to estimate water saturation and permeability in unsaturated zones for sand-clay mixtures. *Greenhouse et al.* [1993] and *Brewster and Annan*, 1994] used the dielectric constants to detect contaminant transport in porous media. Additionally, GPR methods were also used to infer spatial variation in subsurface [*Rea and Knight*, 1998] and spatial correlation structure of hydrogeological parameters

[*Hubbard et al.*, 1999].

Seismic methods have been used to aid in environmental site characterization for many years. These methods use artificially generated high frequency (100 to 500Hz) pulses of acoustic energy to probe the subsurface [*Rubin et al*, 1998]. The commonly used seismic methods include seismic reflection, seismic refraction and crosshole transmission. The reflection method is better to be used for detecting structural and stratigraphic information about the subsurface. For unconsolidated and unsaturated materials, however, this is often difficult due to the lack of well-defined velocity contrasts in the ground. Seismic refraction methods sometimes are chosen in site characterization to replace reflection methods for determining the locations of the water table and the top of bedrock and locating significant faults because they are inexpensive. However, they yield much lower resolution than seismic reflection and crosshole methods [*Lankston*, 1990].

Crosshole seismic methods have the highest resolution compared to other methods, and this permits a very detailed estimate of seismic P-wave velocity structure [*Rector*, 1995]. These high-resolution seismic velocity data can be incorporated into estimation of hydrogeological parameters. For example, *Rubin et al.* [1992] and *Copty et al.* [1993] used seismic velocity together with hydraulic pressure data to map hydraulic permeability. *Hyndman et al.* [1994] coupled seismic velocity data with tracer experiment data to estimate lithofacies and hydraulic conductivity. Although not yet established as a field method, seismic imaging of organic contaminants in the labora-

tory has provided a fundamental step toward the application of seismic tomographic imaging of interwell contamination [*Geller and Myer*, 1995].

Borehole geophysical techniques as basic methods have been applied almost to all the contaminant transport sites in the United States [*Keys*, 1997]. These methods can obtain much more information from a well than can be obtained from drilling, sampling, and testing [*Keys*, 1989]. Geophysical logs provide continuous analog or digital records that can be interpreted in terms of physical properties of soil texture, the contained fluid, and even construction of the well. The various borehole logs can be used to estimate lithofacies or hydrogeological parameters along boreholes [*Doveton*, 1986; *Rogers et al.*, 1992; *Benaouda et al.*, 1999] or used in conjunction with surface geophysical data to provide information about aquifer and hydrogeological parameters to a large extent [*Lortzer and Berkhout*, 1992; *Copty and Rubin*, 1995].

### 1.3.2 Data Assimilation Methods

The development of methods for combining hydrogeological and geophysical data receives less attention than that of geophysical techniques in near-surface investigations. Most of studies using geophysical data focus on the improvement of geophysical data acquisition and interpretation methods, not assimilation models. Due to the need for better understanding of contaminant transport processes in porous media and the availability of multiple sources of information for site characterization, however, an effective data assimilation method is required for the purpose of environmental site

characterization. The following is a brief review of the methods used for incorporating geophysical data into estimation of hydrogeological parameters.

Petrophysical or empirical models, if they are available and applicable, are often the first choice for connecting geophysical data to hydrogeological parameters. For consolidated sediments, many petrophysical or empirical relations have been found in the laboratory, which may be used to estimate hydrogeological parameters in field conditions [*Mavko et al.*, 1998]. For unconsolidated sediments, however, obtaining reliable relations between geophysical measurements and hydrogeological parameters is very difficult [*Marion et al.*, 1992; *Knoll*, 1996; *Bachrach and Nur*, 1998; *Bachrach et al.*, 2000]. The major reason is that different types of geological materials in unconsolidated sediments have much smaller contrasts in geophysical attributes than the ones in consolidated sediments. Consequently, very few applications have been found so far to directly use the petrophysical or empirical models obtained from the laboratory for estimation of hydrogeological parameters in field conditions for unconsolidated sediments.

Regression models, mostly linear regression models, were popular in the early applications of geophysical data for estimation of hydrogeological parameters. Such methods have been used by *Kelly* [1977] and *Mazac et al.* [1985] for resistivity data, used by *Han* [1986] and *Vernik and Nur* [1992] for seismic data, and used by *Topp et al.* [1980] for ground-penetrating radar (GPR) data. These models divide available data into two parts, one for training and the other for testing. The testing results

can be considered as the verification of the trained regression model. Advantages of using these models are that they are simple and can be verified directly using field data. Measurement errors and scale disparity between geophysical data and hydrogeological parameters are also implicitly considered in the fitted relations. However, the fitted models are site-specific and should be used with caution. In addition, these methods are considered as a deterministic approach, and the estimated values are often considered as the trend, drift or mean of hydrogeological parameters [*Kitanidis*, 1998].

Inverse models have been used to estimate hydrogeological parameters from piezometric head measurements for many years [*Yeh*, 1986; *Ginn and Cushman*, 1990; *Sun*, 1994; *McLaughlin and Townley*, 1996]. The models were also used by *Lortzer and Berkhout* [1992] to combine seismic data and lithologic information for lithology estimation and used by *Hyndman et al.* [1994] to combine seismic and tracer experiment data for hydraulic conductivity estimation. Such methods provide a better way to account for the special hydrogeological characteristics of individual sites and have the ability to incorporate a wide range of field information [*McLaughlin and Townley*, 1996]. Creating forward models when using the models for certain types of geophysical data, however, may be very difficult or even impossible sometimes, such as gamma-ray measurements. Additionally, ambiguity in geophysical data interpretation may also cause problems in constructing forward models. The potential of using inverse models for assimilation of hydrogeological and geological data still need

further studies.

Geostatistical models for site characterization have been used for decades to estimate hydrogeological parameters using piezometric head and hydrogeological measurements at boreholes [*Dagan*, 1985; *Rubin and Dagan*, 1987]. Geophysical attributes and hydrogeological parameters in these models are considered as spatial random functions, which form two- or three-dimensional random fields; the measurements of the geophysical attributes and the hydrogeological parameters at some locations are considered as samples of the random functions from the random fields. Using those measurements or samples, we can first infer spatial structures of those random fields, defined by variogram models and their associated parameters such as means, variances and integral lengths. Then we can interpolate or extrapolate those geophysical and hydrogeological data, using kriging or cokriging methods, to the locations where directly measurements are not available. Similar to regression models, these models are also simple yet efficient in cases where well-defined spatial structures exist and can be derived from sampling data. However, cross correlations between the hydrogeological and geophysical data and between various types of geophysical data may not always be identifiable from given measurements; they may be highly nonlinear.

Bayesian methods provide a general framework for data assimilation [*Box and Tiao*, 1973; *Bernardo and Smith*, 1994]. These methods have been shown by *McLaughlin and Townley* [1996] to be consistent with numerous inverse models if prior information is considered. The use of these models for combining geophysical and hydro-

geological data can be found in *Lortzer and Berkhout* [1992], *Copty et al.* [1993], and *Copty and Rubin* [1995]. The key focus of using the methods is the inference of the likelihood functions. Many applications of Bayesian models rely on the assumptions that the multiple spatial random variables have multivariate normal distributions, but this is not a general case. Under field conditions, we often need to infer site-specific likelihood functions directly from the in-situ hydrogeological and geophysical data.

## 1.4 Scope of the Dissertation

This dissertation is divided into five chapters, including the introduction and the summary given later in the dissertation. Each of Chapter 2 through 4 focuses on one important aspect of subsurface characterization and includes a separate introduction, methodology, case study, and discussion or summary sections. These studies are characterized by three factors: 1) the hydrogeological and geophysical data used in each chapter are either real-life field data or synthetic data closely mimicking field conditions, 2) the petrophysical relations between the hydrogeological parameters and the geophysical attributes are site-specific and derived directly from the training data sets using data-driven models, 3) the assimilation of the hydrogeological and geophysical data is performed using Bayesian models.

Chapter 2 explores the use of ground penetrating radar (GPR) tomographic velocity, GPR tomographic attenuation, and seismic tomographic velocity for hydraulic

conductivity estimation. The hydrogeological and geophysical data were collected from the Narrow Channel Focus Area at the South Oyster site in Virginia and include flowmeter and slug test data at each well and GPR and seismic tomographic data along some transects. The integration of those data was carried out by the Bayesian method based on a normal linear regression model. Since the GPR and seismic tomographic data have very high resolution and small sampling volumes at the site, the scale disparity between the hydraulic conductivity and the geophysical tomographic data is relatively small. This allows us to derive the site-specific relations between the hydraulic conductivity and the geophysical attributes from both hydraulic conductivity and geophysical data available at each well using a data-driven model. Although the log-conductivity displays a small variation ($\sigma^2 = 0.30$) and the geophysical data vary over only a small range, results indicate that the geophysical data improve the estimates of the hydraulic conductivity. The improvement is most significant where prior information is limited. Among geophysical data, GPR and seismic velocity are more useful than GPR attenuation.

Chapter 3 considers a scale disparity problem. A Bayesian model is developed to combine small-scale resistivity logs with large-scale electromagnetic (EM) surveys data for mapping two-dimensional resistivity fields, which can be used later to infer hydrogeological parameters. The method is oriented towards the Lawrence Livemore National Laboratory project, where lithofacies and gamma-ray logs have good spatial structures but resistivity does not. Firstly, the small-scale or high-resolution

resistivity data from boreholes are interpolated to crosshole areas using the site-specific relations among gamma-ray, lithofacies and resistivity obtained from field data. Secondly, the estimated resistivity is considered as the prior and updated by the collocated large-scale resistivity data obtained from crosshole EM surveys. The relation between small- and large-scale resistivity is a reflection of the scale disparity between those data, and it becomes weak when the scale disparity is large. Finally, the updated resistivity estimate is compared to the corresponding prior estimate to evaluate the effectiveness and limitations of the Bayesian model. Results reveal that the proposed method enhances hydrogeological site characterization even when the resistivity surveys have a relatively low resolution.

Chapter 4 focuses on the issue of petrophysical relations. A Bayesian model coupled with a fuzzy neural network (BFNN) is developed to alleviate the difficulty of using geophysical data in lithofacies estimation due to nonlinearity of cross correlation between lithofacies and geophysical attributes. The Bayesian model allows for the incorporation of spatial correlation of lithofacies as well as the nonlinear cross correlation into lithofacies estimation. The prior estimate is inferred from lithofacies measurements at boreholes using indicator kriging based on the spatial correlation, whereas the posterior estimate is updated from the prior using the geophysical data based on the nonlinear cross correlation. The key to using the model is the inference of the likelihood function, which is obtained from training data sets using a fuzzy neural network. The fuzzy neural network takes advantages of both fuzzy logic and

neural networks. Fuzzy logic provides an approach to incorporate human knowledge into lithofacies estimation, and this knowledge may be used to justify the rules learned directly from data. Neural networks provide a powerful tool to fit nonlinear functions from given input and output data with no or few assumptions about the form of the functions. The efficiency of the method in lithofacies estimation is demonstrated by two synthetic case studies generated from measurements at the Lawrence Livemore National Laboratory site. Results show that the BFNN model is the best method among indicator kriging, indicator cokriging and the fuzzy neural network without considering spatial correlation.

# Chapter 2

# Estimating the Hydraulic Conductivity at the South Oyster Site from Geophysical Tomographic Data Using Bayesian Techniques Based on a Normal Linear Regression Model

## 2.1 Introduction

Heterogeneity of hydraulic conductivity in porous media is a major control of groundwater flow and contaminant transport [*Dagan*, 1982; *Gelhar and Axness*, 1983]. Modeling of this heterogeneity is difficult and subject to a large degree of uncertainty due to the lack of densely sampled in-situ hydrological measurements.

Conventional borehole techniques such as flowmeter and slug tests for collecting hydrological data are costly, time-consuming and invasive; therefore, a large effort is undertaken to explore the potential of using geophysical data to compensate for the

scarcity of in-situ hydrological measurements [*Rubin et al.*, 1992; *Copty et al.*, 1993 and 1995; *Hubbard et al.*, 1997; *Rubin et al.*, 1998; *Ezzedine et al.*, 1999; *Hubbard and Rubin*, 2000]. Geophysical data used for hydrogeological characterization often include electrical resistivity [*Kelly*, 1977; *Ahmed et al.*, 1988], seismic velocity [*Rubin et al.*, 1992; *Copty et al.*, 1993 and 1995; *Hyndman et al.*, 1994] and ground penetrating radar (GPR) velocity [*Hubbard et al.*, 1997 and 1999 ]. Methods for integration of hydrological and geophysical data include regression models [*Kelly*, 1977] , cokriging models [*Ahmed et al.*, 1988], inversion models [*Rubin et al.*, 1992] and Bayesian models [*Copty et al.*, 1993; *Ezzedine et al.*, 1999] . Despite the difference in the methods and the geophysical data, it has been widely recognized that the most difficult part of the integration is tying hydrological measurements to geophysical data because of the scale and resolution disparity between hydrological and geophysical measurements [*Ezzedine et al.*, 1999] , and because of their non-unique relationships due to the uncertainty associated with field data acquisition and interpretation [*Urish*, 1981]. This chapter proposed an approach to dealing with this issue based on the normal linear regression model. It extends the previous work reported in *Copty et al.* [1993] and *Ezzedine et al.* [1999], and its main novelty is in formulating the petrophysical models in a probabilistic fashion, using likelihood functions.

This chapter explores the potential use of GPR tomographic velocity, GPR tomographic attenuation and seismic tomographic velocity as well as hydrological measurements for estimating hydraulic conductivity. It is focused on the usefulness of

geophysical measurements for hydraulic conductivity estimation and on the integration of hydrological and geophysical data.

The approach given in this chapter is to explore the correlations between the geophysical attributes and the hydraulic conductivity, following the ideas explored in *Rubin et al.* [1992], *Copty et al.* [1993], *Hyndman et al.* [1994] and *Ezzedine et al.* [1999]. Another approach to this problem is to analyze transient effects, for example through time-lapse tomography [*Shapiro et al.*, 1999]. This approach was not pursued here because it requires injecting fluids, which can potentially influence the geophysical signals and thus may have detrimental effects on the overall goals of the field experiments.

This chapter is organized as follows. Section 2.2 introduces the South Oyster Site, available data at the site and some preliminary data analyses. Section 2.3 describes the Bayesian method and the normal linear regression model. Section 2.4 explores the use of geophysical data within a Bayesian framework for estimating hydraulic conductivity. Discussion and conclusions are given in sections 2.5.

## 2.2 Site and Data Descriptions

### 2.2.1 South Oyster Site

The South Oyster Site is located near the town of Oyster on Virginia's Eastern Shore Peninsula between the Chesapeake Bay and the Atlantic Ocean. A field-scale experiment has been undertaken by a multi-disciplinary research team within an uncontaminated aquifer at the Oyster Site to evaluate the importance of chemical and physical heterogeneity in controlling bacteria that are injected into the ground for bioremediation purposes [*DeFlaun et al.*, 2000] . The sediments at the South Oyster Site consist of unconsolidated to weakly indurated, well-sorted, medium- to fine- grained Late Pleistocene sands and pebbly sands. The upper 9 meters of the South Oyster Site consists of the Wachapreague Formation, which was deposited in a shallow, open marine to back-barrier environment, north of the tide-dominated mouth of the Chesapeake Bay [*Mixon*, 1985] . The water table at the South Oyster Site is located approximately 3 meters below ground surface.

Within the South Oyster Site two study focus areas exist: the South Oyster Focus Area and the Narrow Channel Focus Area (Figure 2.1). Locations of the focus areas were chosen based primarily on groundwater chemistry: the South Oyster Focus Area is situated within a suboxic portion of the aquifer, while conditions at the Narrow Channel Focus Area are predominantly aerobic. Forced gradient chemical and bacterial tracer test experiments were performed within what is called the 'Aerobic Flow

Cell' of the Narrow Channel Focus Area in 1999 [*Johnson et al.*, 2001]; similar experiments will be performed within the 'Suboxic Flow Cell' of the South Oyster Focus Area in 2000. At both locations, extensive geological, geophysical and hydrological data were and are being collected to characterize the subsurface prior to the tracer test experiments. This study explores the use of geophysical tomographic data, collected within the saturated portion of the Aerobic Flow Cell (approximately between depths of 0-6.0 m below mean sea level [MSL]), for providing detailed hydraulic conductivity estimates there.

The Aerobic Flow Cell layout within the Narrow Channel Focus Area is shown in Figure 2.2. Hydraulic conductivity measurements are available at the wellbore locations, indicated by circles, and geophysical tomographic profiles are available between several well pairs, as indicated by the solid lines. Descriptions of these available data are discussed in next section. The chemical and bacterial tracer injection well is NCB2. Groundwater flow direction and geologic dip are aligned approximately parallel to the transect NCB2-NCM3, and geologic strike is aligned approximately parallel with the transect NCT3-NCT1. Twenty-four multi-level samplers were installed between the wells NCB2 and NCM3, and NCT3 and NCT2 to detect the bacterial passage of chemical and bacterial tracers over time during the tracer test experiments as described by *Johnson et al.* [2001]. The log-conductivity estimates within the Aerobic Flow Cell, obtained using geophysical tomographic data as described in this study, will be used to help constrain the stochastic numerical flow

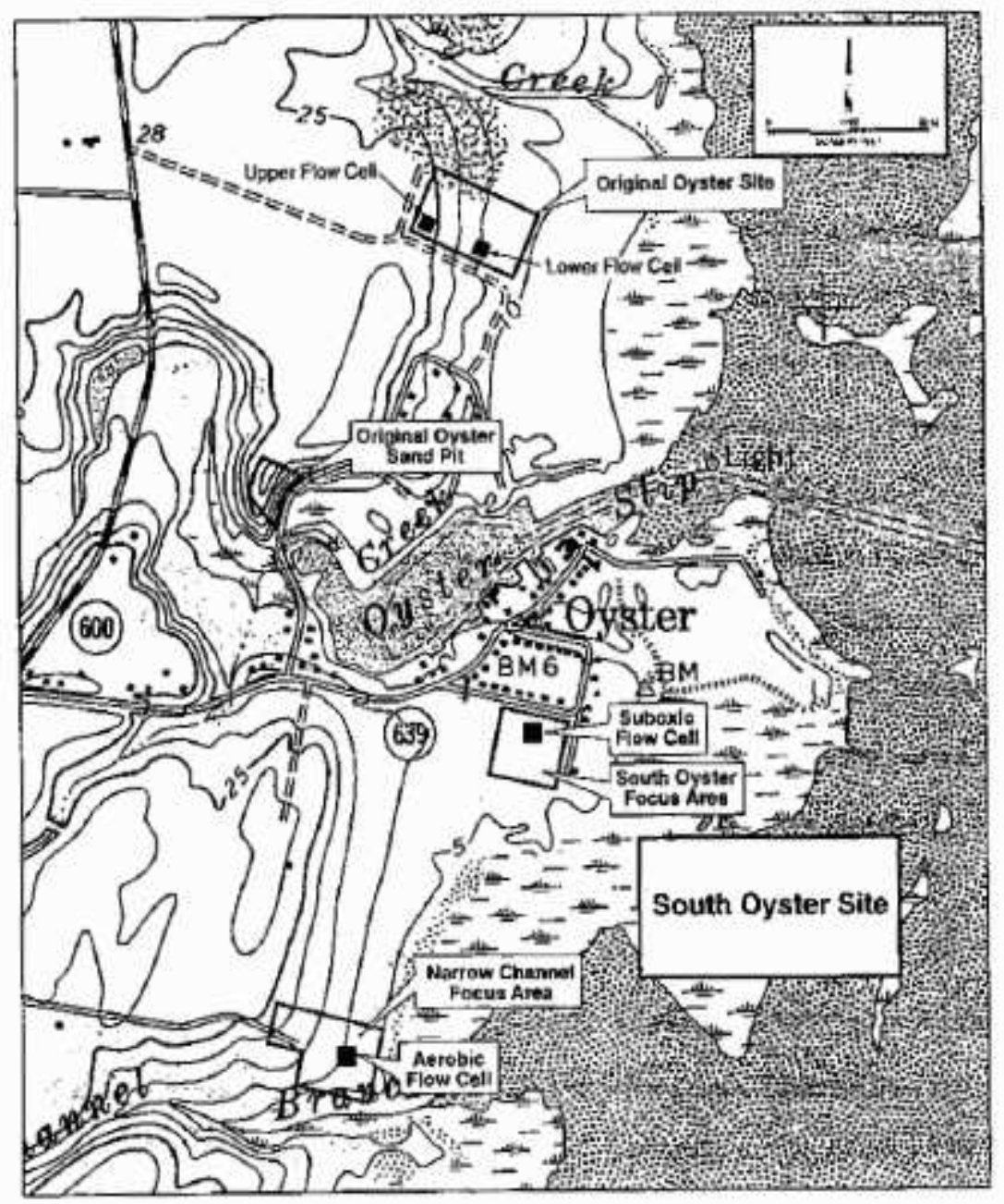


Figure 2.1: Location of the South Oyster Site and the Aerobic Flow Cell (*Golder Associates, 1998*).

studies being performed to understand the transport experiment results [*Scheibe et al.*, 1999].

### 2.2.2 Field Sampling

#### Hydraulic Conductivity Measurements from Flowmeter Data

Hydraulic conductivity values were calculated from flowmeter and slug test data within the Aerobic Flow Cell. Electromagnetic borehole flowmeter data were collected from all wells whose locations are shown in Figure 2.2. Each well was approximately 9.4 m deep and had two 3.05 m long screens positioned in the lower 6.1 m, or from approximately 0.5 m to 5.8 m below MSL [*Waldrop and Hamby*, 1998]. The flowmeter data provided relative hydraulic conductivity measurements at discrete intervals of 0.15 m for each well. Slug test data, where available, were used to provide average hydraulic conductivity values over the screened well intervals. Where slug test data were not available, the geometric mean of the slug tests over the entire flow cell was used. These average hydraulic conductivity values were then used to convert the relative hydraulic conductivity measurements, obtained from flowmeter data, into absolute hydraulic conductivity values for that well [*Molz and Young*, 1993].

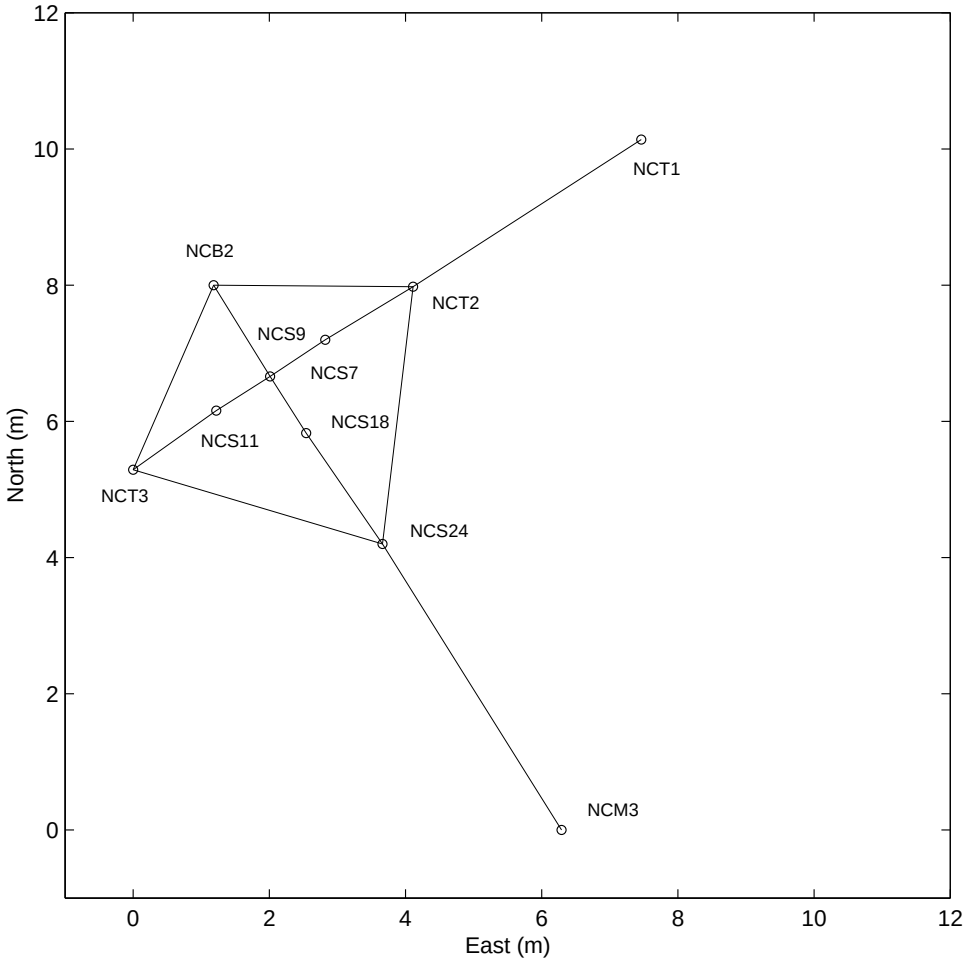


Figure 2.2: Aerobic Flow Cell in the Narrow Channel Focus Area (the circles denote flowmeter measurement wells and the solid lines denote geophysical tomographic profiles).

**Geophysical Measurements from Tomographic GPR and Seismic Data**

Both GPR and seismic tomographic data were collected within the Aerobic Flow Cell along the transects indicated in Figure 2.2. For tomographic acquisition geometry, GPR transmitting antenna (or seismic source) and GPR receiving antennas (or seismic geophones) are located in separate wellbores, and direct energy from a transmitting antenna in one wellbore is recorded by a receiving antenna located in the other wellbore. The transmitter position is changed and the recording repeated until both the transmitter and the receiver have occupied all possible positions within the two wellbores. The direct electromagnetic or seismic P-wave wave travel time between all transmitter/receiver positions, as well as the amplitude of the direct arrival, is obtained from the recorded data. The interwell area is then discretized into a grid composed of cells or pixels, and inversion algorithms are used to transform the recorded travel time and amplitude information into estimates of velocity and attenuation, respectively, at each pixel. The discretization that is chosen for the inversion is typically based on consideration of several factors including: the wavelength of the propagating signal, expected material properties and their contrasts, acquisition geometry including wellbore spacing and geophone spacing (which control propagation distance, raypath density and illumination angles) and inversion damping parameters. The reader interested in different types of geophysical inversion approaches and applications is referred to *Parker* [1994], *Williamson and Worthington* [1993], and *Rector* [1995].

The seismic tomographic profiles were collected along the same transects where GPR tomographic data were collected (Figure 2.2. These data were collected using a Geometrics Strataview seismic system with a piezoelectric source. The central frequency of the pulse was 4000 Hz, with a bandwidth from approximately 1000 to 7000 Hz, rendering an average seismic wavelength of approximately 0.4 m. The source and geophone spacing in the wellbores was 0.125 m, which resulted in dense raypath coverage of over 13,200 traces in the interwell area. Based on the seismic wavelength, small wellbore spacing of approximately 3-5 m, and dense raypath coverage over a variety of illumination angles, a discretization of 0.25 m x 0.25 m was chosen for the discretization inversion. The travel times were picked for all source-receiver pairs. The travel time data were then inverted using a straight-ray algebraic reconstruction technique [*Peterson et al.*, 1985] to produce seismic velocity estimates for each 0.25 m x 0.25 m cellblock along all transects. The small velocity range observed in the data suggests that distortion caused from ray bending should be minimal, and that raypath density should be fairly evenly distributed in the interwell area.

Seismic amplitudes can yield information about the attenuative properties of subsurface sediments or rocks. In a manner similar to travel time inversion, amplitude information can be extracted from the tomographic data and inverted for attenuation in the interwell area. Many theories exist to incorporate the great variety of mechanisms that can influence seismic wave attenuation. For porous, granular, sedimentary rocks, the generally accepted mechanisms may be grouped into three broad

categories: scattering attenuation, fluid-flow attenuation, and fluid-matrix attenuation [*Nihei*, 1992]. Because of the variety of influences on the seismic amplitudes, it is often difficult to extract meaningful characterization information from seismic amplitude data. Additionally, seismic amplitudes are extremely sensitive to the presence of trapped gas and the state of consolidation. The presence of a shallow water table, small amounts of organic material (potential sources of trapped gas), and varying states of consolidation of the Oyster sediments are suspected to have affected the seismic source radiation pattern and coupling, as well as the receiver properties. If not adjusted, these radiation pattern and coupling variations often yield inversion artifacts [*Vasco et al.*, 1996; *Keers et al.*, 2000]. Because inversion artifacts were observed in the Oyster seismic attenuation tomograms, the confidence in these data was not high enough to use rigorously in this hydrological property estimation procedure. Use of the Oyster seismic amplitude data is left for future studies, after further investigation and pre-processing of the seismic radiation pattern and coupling effects have been performed.

The tomographic GPR data were collected using a PulseEKKO 100 system with 200 MHz central frequency wellbore antennas. These data were collected using a transmitter/receiver spacing in the wellbores of 0.125 m, which again resulted in over 13,200 traces per tomogram. The effective range of the radar propagation frequencies was 40-140 MHz, rendering effective GPR wavelengths of approximately 0.5 m. Although the wavelengths of the radar data are on average greater than those of the

seismic data, the high signal to noise ratio of the radar data relative to the Oyster seismic data permitted inversion using the same discretization as was used for the seismic data of 0.25 m x 0.25 m. The similar discretization used for both the seismic and radar tomographic data inversions facilitated the hydrogeological parameter estimate computations. For the high radar frequencies employed and in the sandy environment at Oyster, the radar propagation velocities are primarily governed by variations in the dielectric constant, and the amplitudes are primarily affected by variations in dielectric constant and electrical conductivity [*Davis and Annan*, 1989] of the interwell sediments. Unlike the seismic amplitude data, the Oyster radar amplitude radiation patterns and source-receiver coupling appeared to be consistent, and thus inversion was performed on both the picked travel times and amplitudes using straight-ray algebraic reconstruction techniques [*Peterson et al.*, 1985; *Peterson*, 2000] to yield electromagnetic wave velocity and attenuation estimates for each cellblock along all tomographic profiles.

### 2.2.3 Data Analysis

Since the goal of this study is to explore and test the use of geophysical tomographic measurements for hydraulic conductivity estimation, only those geophysical data at the wellbores are used, where hydraulic conductivity data are available. Geophysical data at these wells were approximated by the ones at the cellblocks located one column away from the wells rather than directly near the wells on their surround-

ing transects. This is because geophysical data located directly near the wells may be affected by the presence of disturbed zones around the wells [*Peterson*, 2000]. For some wells such as NCB2, there are three transects passing them; the averaged values of the geophysical data extracted from each transect are used. Consequently, geophysical data are obtained at each well location with an interval of 0.25 m along vertical directions. Since log-conductivity was sampled with an interval of 0.15 m rather than 0.25 m, the log-conductivity data at the vertical locations where geophysical data were also sampled for each well are interpolated from the direct hydraulic conductivity measurements. Finally, a data set in which each log-conductivity value has corresponding co-located geophysical data is created, which will be used later in this chapter.

**Hydraulic Conductivity**

The histogram of natural log-conductivity at the Aerobic Flow Cell is shown in Figure 2.3(a). It is asymmetric and negatively skewed, and suggests the existence of two sub-populations of hydraulic conductivity—high- and low-conductivity zones [*Copty et al.*, 1995; *Welhan and Reed*, 1997].

The spatial structure of the log-conductivity was identified through covariance analyses and the results are shown in Figure 2.4. The vertical covariance given in Figure 2.4(a) can be fitted with an exponential covariance model with a range of 0.6 m and a sill of 0.30 for small lags (¡0.6 m), but hole-type structure appears at

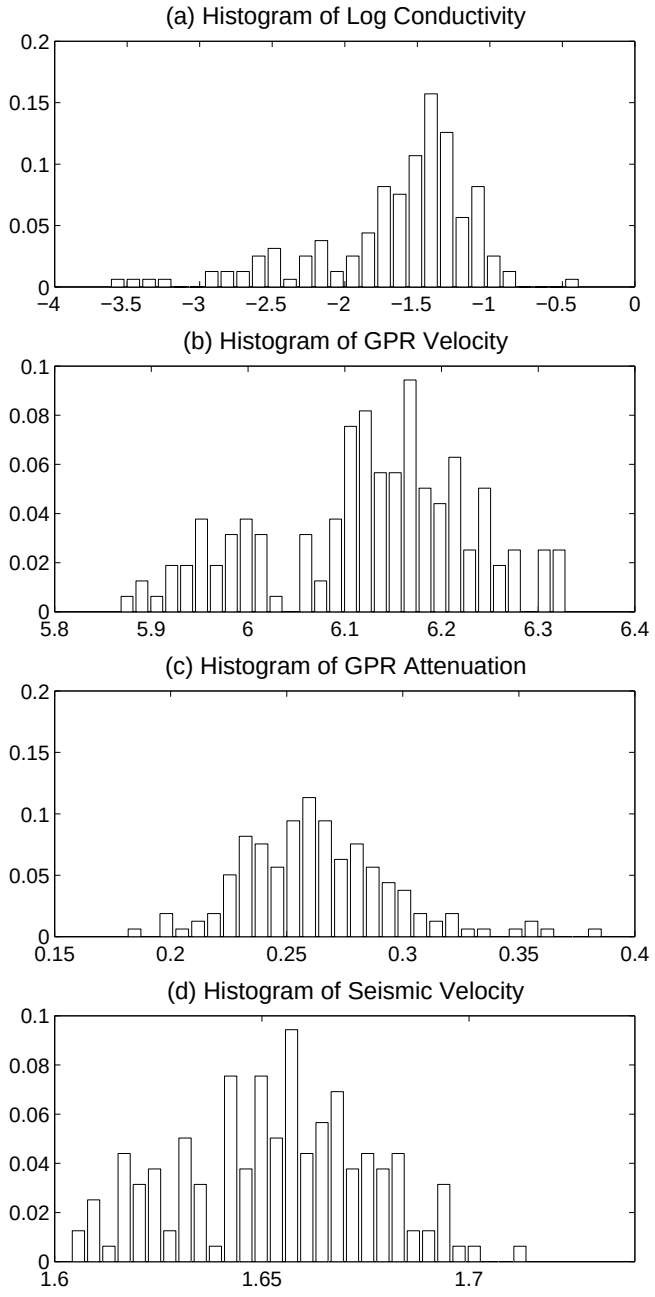


Figure 2.3: (a) Histogram of log-conductivity (hydraulic conductivity in $m/h$), (b) Histogram of GPR velocity ($cm/ns$) (c) Histogram of GPR attenuation ($1/m$) (d) Histogram of seismic velocity ($km/s$).

large lags (¿0.6 m). The hole effect may be the outcome of repetitive sequences or periodic variations [*Isaaks and Srivastava*, 1989]. As pointed out by *Journel and Huijbregts* [1978], the hole effect may also be due to an artificial pseudo-periodicity of available data, and can be ignored in practice if not very remarkable. The lateral covariances along the directions perpendicular and parallel to the geologic strike are shown in Figure 2.4(b) and Figure 2.4(c), respectively. Both covariances are fitted with exponential models with a range of 5 m.

### Geophysical Data

Histograms of GPR velocity, GPR attenuation and seismic velocity are shown in Figure 2.3(b,c,d), respectively. Similar to log-conductivity (Figure 2.3(a)), negative skewness is observed in the histograms of GPR and seismic velocity. This suggests correlations between log-conductivity and GPR and seismic velocity. The GPR and seismic velocity change over small ranges and exhibit small variations (coefficient of variation CV=1.7% for GPR velocity and CV=1.4% for seismic velocity) compared to the GPR attenuation (CV=12.5%).

### Correlations Between Log-Conductivity and Geophysical Data

Physical connections between log-conductivity and GPR velocity, GPR attenuation and seismic velocity exist, but not straightforward. For instance, hydraulic conductivity correlates to porosity, as evidenced by the Kozeny-Carman equation

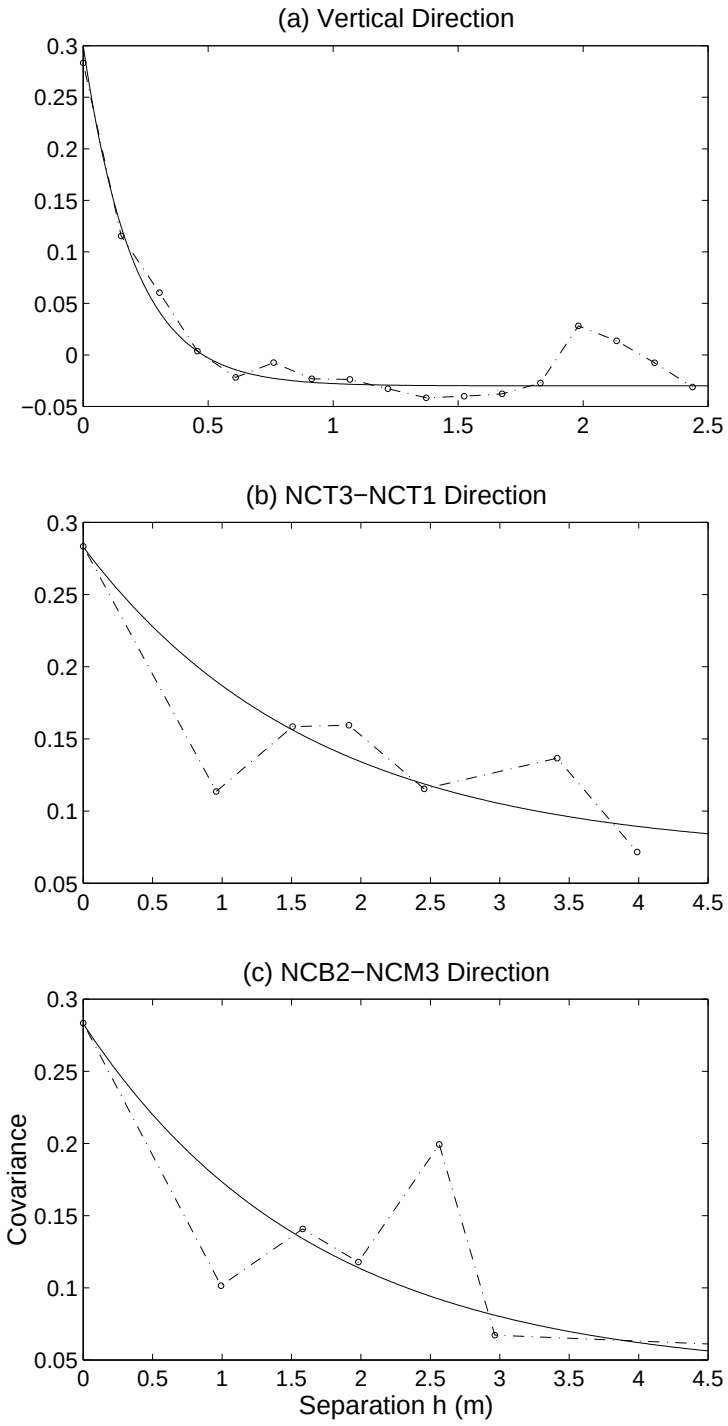


Figure 2.4: (a) Covariance along vertical direction (b) Covariance along the direction perpendicular to geologic strike (NCB2-NCM3) (c) Covariance along the direction parallel to geologic strike (NCT3-NCT1).

[*Carman*, 1956], and GPR velocity, GPR attenuation and seismic velocity also relate to porosity [*Knoll*, 1996; *Marion*, 1990; *Mavko et al.*, 1998].

Figure 2.5 depicts scatter-plots of log-conductivity versus GPR velocity, GPR attenuation and seismic velocity based on the data available at the Aerobic Flow Cell. The log-conductivity correlates with GPR and seismic velocity, and it generally increases as GPR and seismic velocity increase. The GPR attenuation and log-conductivity appear to be uncorrelated, and the GPR attenuation associated with low log-conductivity ($log(k) < -2$) seems to be less variable.

## 2.3 Methodology

Estimating log-conductivity suffers from much uncertainty due to the lack of densely sampled in-situ hydrological measurements, and due to the absence of unique relations between log-conductivity and geophysical data. To address this uncertainty, a stochastic framework is adopted in which log-conductivity, GPR velocity, GPR attenuation and seismic velocity are considered as spatial random functions.

### 2.3.1 Bayesian Formula

A Bayesian methodology is developed in this section for estimating log-conductivity from hydrological and geophysical data. Let the random variable $Y$ denote log-conductivity and $V_g$, $\alpha$ and $V_s$ denote GPR velocity, GPR attenuation and seismic

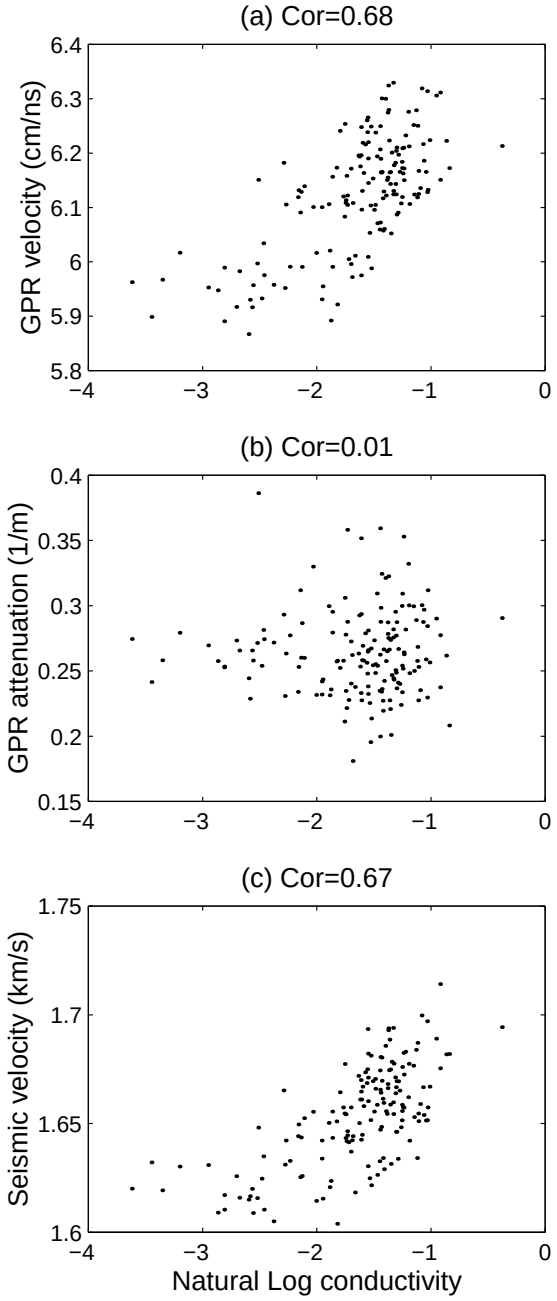


Figure 2.5: Scatter-plots of log-conductivity, GPR velocity, GPR attenuation and seismic velocity.

velocity, respectively. All data are mean-removed and normalized by their corresponding standard deviations. The log-conductivity estimate at a given location $\mathbf{x}$, in terms of probability density function (pdf), is obtained using the Bayes theorem as follows [*Box and Tiao*, 1973; *Kitanidis*, 1986]:

$$f'_Y(y(\mathbf{x})) = CL(y(\mathbf{x})|v_g(\mathbf{x}), \alpha(\mathbf{x}), v_s(\mathbf{x}))f_Y(y(\mathbf{x})), \tag{2.1}$$

where $y(\mathbf{x})$ is an unknown value of $Y$ being estimated at $\mathbf{x}$, $v_g(\mathbf{x})$, $\alpha(\mathbf{x})$ and $v_s(\mathbf{x})$ are the measured values of $V_g$, $\alpha$ and $V_s$ at the same location, $C$ is a normalizing coefficient, $L(y(\mathbf{x})|v_g(\mathbf{x}), \alpha(\mathbf{x}), v_s(\mathbf{x}))$ is the likelihood function given $v_g(\mathbf{x})$, $\alpha(\mathbf{x})$ and $v_s(\mathbf{x})$, and $f'_Y(y(\mathbf{x}))$ and $f_Y(y(\mathbf{x}))$ are the posterior and prior pdfs of $Y$ at $\mathbf{x}$, respectively. Note that only co-located geophysical data have been used to update the prior pdf since they are most informative compared to the measurements at adjacent locations [*Copty et al.*, 1993].

The Bayesian method has been used for many years in the water resources field. One of the earliest applications in groundwater hydrology was provided by *Kitanidis* [1986] for analyzing parameter uncertainty in estimation of spatial functions. In that work, the mean and covariance matrix of the posterior distribution were derived analytically by choosing a prior distribution that is conjugate to the likelihood function in the sense that the posterior has the same form as the prior. Following the same line, *Copty et al.* [1993] applied the method to subsurface characterization of hydrological properties using geophysical data, and the analytical forms of the posterior mean and variance were also obtained under certain assumptions. This study develops a new

approach, which allows for large flexibility in the form of the likelihood function and posterior pdf, to get numerical rather than analytical posterior mean and variance.

### The Prior pdf

The prior pdf $f_Y(y(\mathbf{x}))$ was estimated based on the hydraulic conductivity data using kriging [*Journel*, 1989]. A similar approach was also used by *Copty et al.* [1993] and *Ezzedine et al.* [1999] . The prior distribution is normal if $Y$ is multivariate normal [*Deutsch and Journel*, 1998] .

### The Likelihood Function

The likelihood function $L(y(\mathbf{x})|v_g(\mathbf{x}), \alpha(\mathbf{x}), v_s(\mathbf{x}))$ plays a central role in the Bayesian method and was inferred from the hydrological and co-located geophysical data. It is expressed as follows [*Bernardo and Smith*, 1994]:

$$\begin{aligned} L(y(\mathbf{x})|v_g(\mathbf{x}), \alpha(\mathbf{x}), v_s(\mathbf{x})) &= f_{V_g}(v_g(\mathbf{x})|y(\mathbf{x})) \cdot f_\alpha(\alpha(\mathbf{x})|y(\mathbf{x}), v_g(\mathbf{x})) \\ &\quad \cdot f_{V_s}(v_s(\mathbf{x})|y(\mathbf{x}), v_g(\mathbf{x}), \alpha(\mathbf{x})), \end{aligned} \tag{2.2}$$

where $f(\cdot|\cdot)$ denotes a conditional pdf. If $V_g$, $\alpha$ and $V_s$ are independent such that $f_\alpha$ $(\alpha(\mathbf{x})\,|y(\mathbf{x}), v_g(\mathbf{x})) = f_\alpha\,(\alpha(\mathbf{x})\,|y(\mathbf{x}))$ and $f_{V_s}(v_s(\mathbf{x})|y(\mathbf{x}), v_g(\mathbf{x}), \alpha(\mathbf{x})) = f_{V_s}(v_s(\mathbf{x})|y(\mathbf{x}))$, the inference of the likelihood function becomes simple since each conditional pdf involves only two variables. This is however not the case in the present study where

there are four dependent variables ($Y$, $V_g$, $\alpha$ and $V_s$) and the prior pdf needs to be updated based on all the co-located geophysical data.

### 2.3.2 Normal Linear Regression Model

The normal linear regression model [*Stone*, 1995] provides a systematic approach to the inference of the conditional pdfs shown in equation 2.2. A similar approach has also been suggested by *Kitanidis* [1991] to model a linear drift of a spatially dependent variable, such as log-conductivity. This section will demonstrate the inference of $f_{V_s}(v_s(\mathbf{x})|y(\mathbf{x}),\ v_g(\mathbf{x}), \alpha(\mathbf{x}))$, and the method can be applied to the other functions appearing in equation 2.2.

In the normal linear regression model, the seismic velocity $V_s$ at $\mathbf{x}$ is assumed to be normally distributed with mean $\mu$ and variance $\sigma^2$. The mean $\mu$ is assumed to be a member of the linear function space $G$ whose basis functions consist of $m$ distinct monomials $g_1(\mathbf{x})$, $g_2(\mathbf{x})$, $\cdots$, $g_m(\mathbf{x})$, formed from combinations of powers and products of $y(\mathbf{x})$, $v_g(\mathbf{x})$ and $\alpha(\mathbf{x})$, such as 1, $y(\mathbf{x})$, $v_g(\mathbf{x})$, $\alpha(\mathbf{x})$, $y^2(\mathbf{x})$, $v_g^2(\mathbf{x})$, $\alpha^2(\mathbf{x})$, $y(\mathbf{x})v_g(\mathbf{x})$, $y(\mathbf{x})\alpha(\mathbf{x})$ and $v_g(\mathbf{x})\alpha(\mathbf{x})$. It is modeled as follows:

$$\mu(\mathbf{x}) = \sum_{i=1}^{m} \beta_i g_i(\mathbf{x}), \tag{2.3}$$

where $\beta_i$ is a coefficient of basis function $g_i(\mathbf{x})$, $i = 1, 2, \cdots, m$. The variance $\sigma^2$ is taken to be a constant, independent of $y(\mathbf{x})$, $v_g(\mathbf{x})$ and $\alpha(\mathbf{x})$. The final set of basis functions is determined by following a model selection procedure, given later in this

section.

**Estimation of $\mu$ and $\sigma^2$**

The mean function $\mu(\mathbf{x})$ and the variance $\sigma^2$ are estimated from the data ($v_s(\mathbf{x}_j)$, $y(\mathbf{x}_j)$, $v_g(\mathbf{x}_j)$, $\alpha(\mathbf{x}_j)$), $j = 1, 2, \cdots, n$, where $y(\mathbf{x}_j)$ is the log-conductivity at location $\mathbf{x}_j$, and $v_s(\mathbf{x}_j)$, $v_g(\mathbf{x}_j)$ and $\alpha(\mathbf{x}_j)$ are the dimensionless co-located seismic velocity, GPR velocity and GPR attenuation, respectively.

Estimating $\beta_1$, $\beta_2$, $\cdots$, $\beta_m$ is achieved by minimizing the residual sum of squares

$$RSS = \sum_{j=1}^{n} (v_s(\mathbf{x}_j) - \mu(\mathbf{x}_j))^2. \tag{2.4}$$

Let $\boldsymbol{\beta}$= $(\beta_1, \beta_2, \cdots, \beta_p)^T$ and $\mathbf{Z} = (v_s(\mathbf{x}_1), v_s(\mathbf{x}_2), \cdots, v_s(\mathbf{x}_n))^T$, where the exponent $T$ denotes the transpose operator. The estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, which minimizes equation 2.4, is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{Z}, \tag{2.5}$$

where $\mathbf{D}$ is a design matrix, given by

$$\begin{pmatrix} g_1(\mathbf{x}_1) & g_2(\mathbf{x}_1) & \cdots & g_m(\mathbf{x}_1) \\ g_1(\mathbf{x}_2) & g_2(\mathbf{x}_2) & \cdots & g_m(\mathbf{x}_2) \\ \cdots & \cdots & \cdots & \cdots \\ g_1(\mathbf{x}_n) & g_2(\mathbf{x}_n) & \cdots & g_m(\mathbf{x}_n) \end{pmatrix}. \tag{2.6}$$

Once $\beta_1$, $\beta_2$, $\cdots$, $\beta_p$ are estimated, the mean and the variance of $f_{V_s}(v_s(\mathbf{x})|y(\mathbf{x}),$

$v_g(\mathbf{x}), \alpha(\mathbf{x}))$ are defined as follows:

$$\hat{\mu}(\mathbf{x}) = \sum_{i=1}^{m} \hat{\beta}_i g_i(\mathbf{x}), \tag{2.7}$$

$$\hat{\sigma}^2 = \frac{1}{n-m} \sum_{j=1}^{n} (v_s(\mathbf{x}_j) - \hat{\mu}(\mathbf{x}_j))^2. \tag{2.8}$$

It is clear that the estimate $\hat{\sigma}^2$ of the variance in equation 2.8 is stationary and independent of locations. The estimate $\hat{\mu}(\mathbf{x})$ of the mean function in equation 2.7, however, depends on $y(\mathbf{x})$, $v_g(\mathbf{x})$ and $\alpha(\mathbf{x})$ and thus on the location $\mathbf{x}$.

**Selection of Basis Functions**

Selecting and eliminating basis functions is the key to the normal linear regression model. The initial set of the basis functions consists of all possible distinct monomials of $y(\mathbf{x})$, $v_g(\mathbf{x})$ and $\alpha(\mathbf{x})$ with a degree of 4 at most. The final set is obtained from the initial set by deleting some of the initial basis functions based on certain criteria described in the following.

Removing or retaining a basis function $g_i(\mathbf{x})$ $(1 \leq i \leq m)$ from the initial set is based on testing of the null hypothesis $H_o : \beta_i = 0$, at the common testing level 0.05. The statistic $u = \hat{\beta}_i / SE(\hat{\beta}_i)$ is first computed, where $SE(\hat{\beta}_i)$ is the standard error of $\hat{\beta}_i$, which is the i-th element in the diagonal of the matrix $\hat{\sigma}^2(\mathbf{D}^T\mathbf{D})^{-1}$ ($\hat{\sigma}^2$ is obtained from equation 2.8 and $\mathbf{D}$ from equation 2.6). This statistic follows the t-distribution with $n-m$ degrees of freedom, based on the properties of the normal linear regression model [*Stone*, 1995]. Consequently, the p-value, defined by $2(1 - t_{n-m}(|u|))$ where

$t_{n-m}$ is the probability function of the t-distribution with $n-m$ degrees of freedom, can be calculated and compared to the testing level 0.05. If the p-value is larger than 0.05, the null hypothesis is accepted and $g_i(\mathbf{x})$ is removed from the basis function set; otherwise, the null hypothesis is rejected and $g_i(\mathbf{x})$ is retained.

The procedure of deleting basis functions is an iterative process, which is executed as follows: (1) fitting a model $\mu(\mathbf{x})$ as shown in equation 2.3 to the sampled data and obtaining $\hat{\beta}_i$ (equation 2.5) and $SE(\hat{\beta}_i)$ $(i = 1, 2, \cdots, m)$; (2) computing their corresponding p-values; (3) comparing all the p-values with 0.05; (4) removing the basis function with the largest p-value from the initial set. The procedure is repeated until no members of the basis function set can be removed. The same method is also applied to determine the other pdfs, which compose the likelihood function (equation 2.2).

## 2.4 Hydraulic Conductivity Estimation

### 2.4.1 Outline of the Approach

To apply and test the previous approach, the hydraulic conductivity and geophysical data available at the Aerobic Flow Cell are split into a training set and a testing set. To avoid the bias in selecting the testing set, each well shown in Figure 2.2 is in turn considered as a testing well; thus there are ten different combinations of training and testing sets. For each of those combinations, the steps outlined in Figure 2.6 are

followed, which can be grouped into three categories:

1. Prior estimation: The spatial correlation structure of log-conductivity is first inferred from the hydraulic conductivity data of the training set and the mean log-conductivity and its variance at each testing location are then estimated using kriging. The prior estimate is a random variable having the normal distribution with the mean and the variance.

2. Posterior estimation: The conditional pdfs, which form the likelihood function as shown in equation 2.2, are derived from both hydraulic conductivity and geophysical data of the training set using the normal linear regression model. The posterior pdf is obtained by following equation 2.1.

3. Evaluation: Various statistics of the prior and posterior pdfs are compared with the actual measurements to evaluate the proposed model. More details on that are given in the subsequent sections.

### 2.4.2 Estimating the Hydraulic Conductivity Using GPR Velocity

Figure 2.7 compares the log-conductivity measurements at well NCS7 with the means of the prior pdfs and the posterior pdfs updated using its co-located GPR velocity only. The improvement of the posterior mean estimates is not significant,

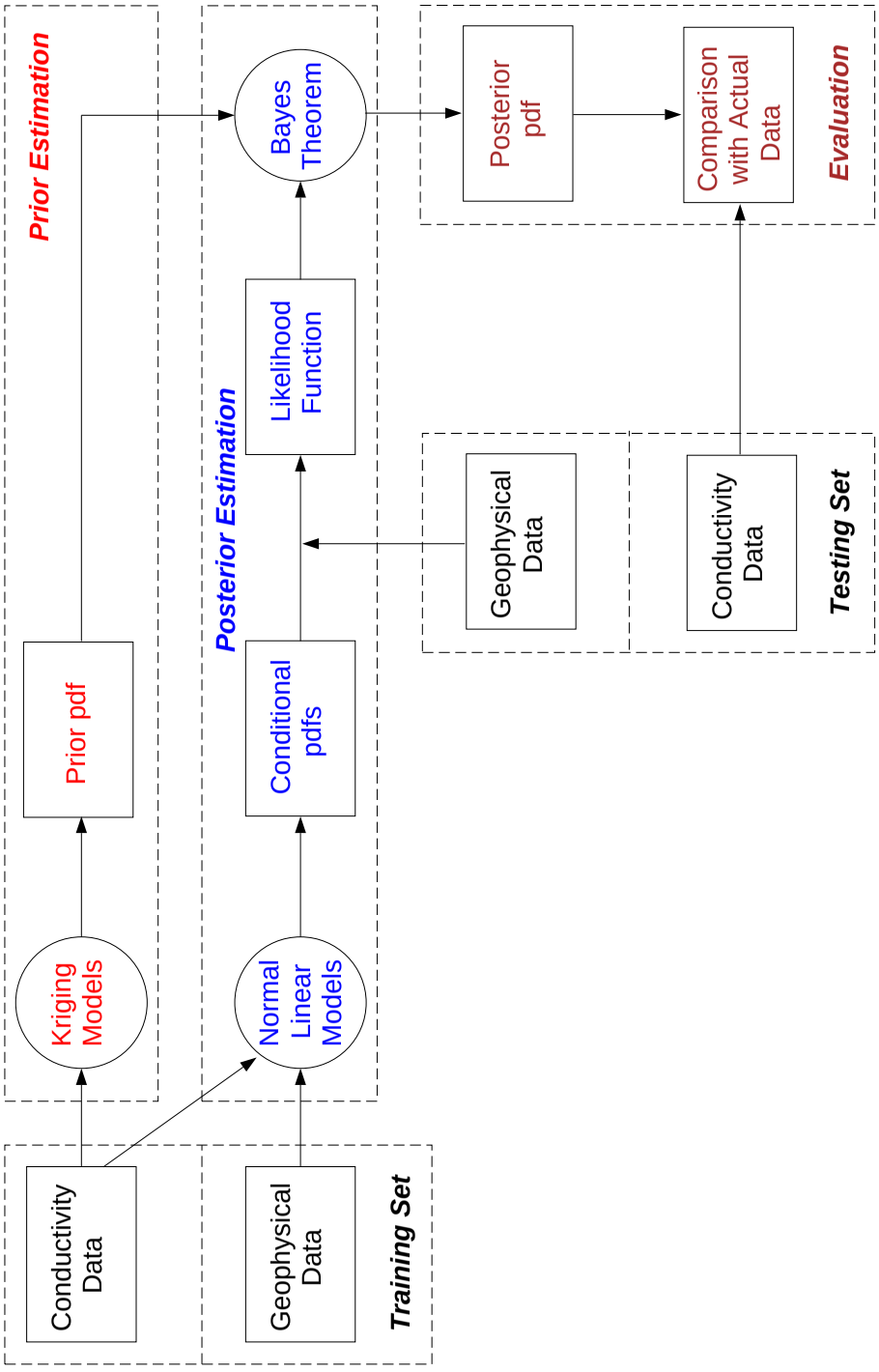


Figure 2.6: Schematic map of the Bayesian method.

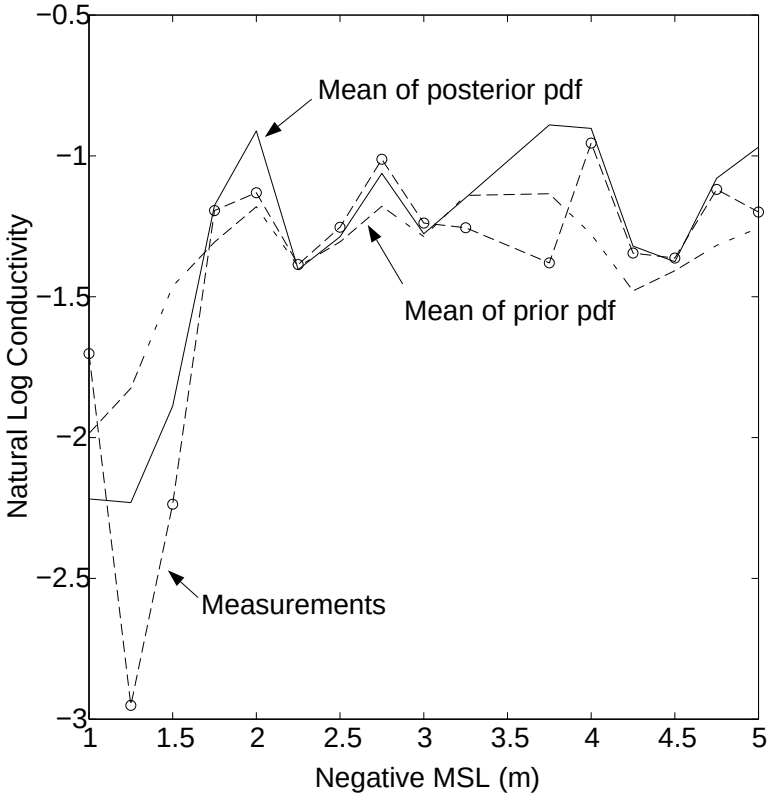


Figure 2.7: Comparisons of measurements, prior and posterior means at testing well NCS7.

and both prior and posterior means follow quite closely the trend outlined by the actual measurements. This is because well NCS7 is close to wells NCS9 and NCT2 (Figure 2.2) which belong to its training set, and thus the prior mean estimates based on the hydraulic conductivity data alone are fairly accurate. Negligible improvement was also observed at other testing wells located in close proximity to training wells, such as wells NCS9, NCS11 and NCS18.

As the distances between wells of the training sets and the testing wells increase,

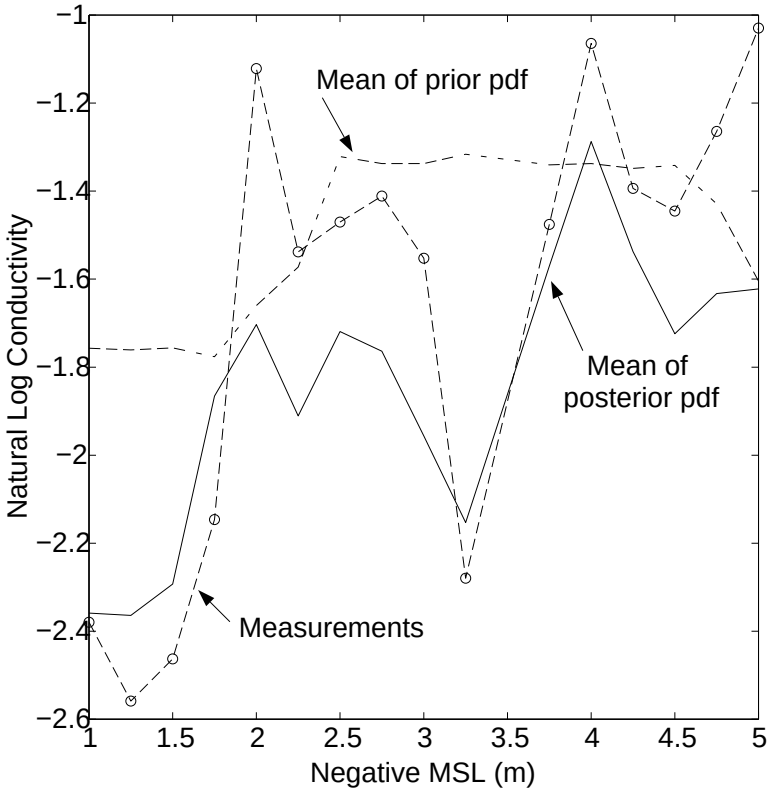


Figure 2.8: Comparisons of measurements, prior and posterior means at testing well NCM3.

the contribution of the co-located GPR velocity data becomes apparent. This is demonstrated in Figure 2.8 which compares actual measurements with the means of prior and posterior pdfs at well NCM3, located no less than 4.95 m away from other training wells. The posterior estimates follow more closely the measurements compared to the prior estimates. Similar results were also observed at the testing wells NCT1, NCT2, NCT3, NCS24 and NCB2.

Table 2.1 shows the vertical averages of standard deviations of prior pdfs and

Table 2.1: Comparisons of prior and posterior standard deviations

| Testing Well | Prior standard deviations $\sigma_{prior}$ | Posterior standard deviations $\sigma_{post}$ | Reduction(%) $(\sigma_{prior} - \sigma_{post})/\sigma_{prior}$ |
|---|---|---|---|
| NCS7 | 0.47 | 0.38 | 19 |
| NCS9 | 0.42 | 0.32 | 24 |
| NCS11 | 0.47 | 0.37 | 21 |
| NCS18 | 0.49 | 0.37 | 24 |
| NCM3 | 0.60 | 0.42 | 30 |
| NCB2 | 0.53 | 0.38 | 28 |
| NCS24 | 0.55 | 0.40 | 27 |
| NCT1 | 0.59 | 0.42 | 29 |
| NCT2 | 0.52 | 0.39 | 25 |
| NCT3 | 0.53 | 0.40 | 25 |

posterior pdfs updated using GPR velocity for each well shown in Figure 2.2. The prior variances are estimated from kriging, and the posterior variances are computed from the posterior pdfs using numerical integration:

$$\sigma^2_{post}(\mathbf{x}) = \int (y(\mathbf{x}) - \langle Y(\mathbf{x})\rangle)^2 f'_Y(y(\mathbf{x}))dy, \tag{2.9}$$

where $\langle Y(\mathbf{x})\rangle$ is the mean of the posterior pdf $f'_Y(y(\mathbf{x}))$ at $\mathbf{x}$. For all the testing wells, the standard deviations are significantly reduced by the use of GPR velocity, and the reductions at wells NCM3 and NCT1 are most evident because their prior estimates are less informative due to the relatively large distances of these wells from the wells used for calculating the priors.

Figure 2.9 shows the 95% confidence intervals at well NCM3 for prior and posterior estimates. The errors predicted by the prior model are on the conservative side, and at

each testing location, the updating provides more narrow bounds, which are consistent with the actual errors.

### 2.4.3 Estimating the Hydraulic Conductivity Using GPR Velocity, GPR Attenuation, and Seismic Velocity

This section explores the efficiency of using GPR velocity, GPR attenuation and seismic velocity in the hydraulic conductivity estimation. Since the addition of GPR attenuation and seismic velocity after using GPR velocity did not lead to further changes in the estimates of the mean log-conductivity in the current study, the discussion focuses on prior and posterior variances of all the ten testing wells.

Figure 2.10 compares the averaged actual errors over the ten testing wells with the averaged standard deviations of prior and posterior pdfs along depth and over the ten testing wells for different combinations of geophysical data. The actual errors are the spatial averages of absolute differences between actual and estimated values along each testing well. It is evident that the standard deviations of the model using the hydraulic conductivity data only are much larger than the actual errors, but with the addition of various types of geophysical data, the standard deviations consistently decrease until they are of the same order as the actual errors.

To evaluate the efficiency of geophysical data in reducing uncertainty, different combinations of geophysical data are used in the hydraulic conductivity estimation,

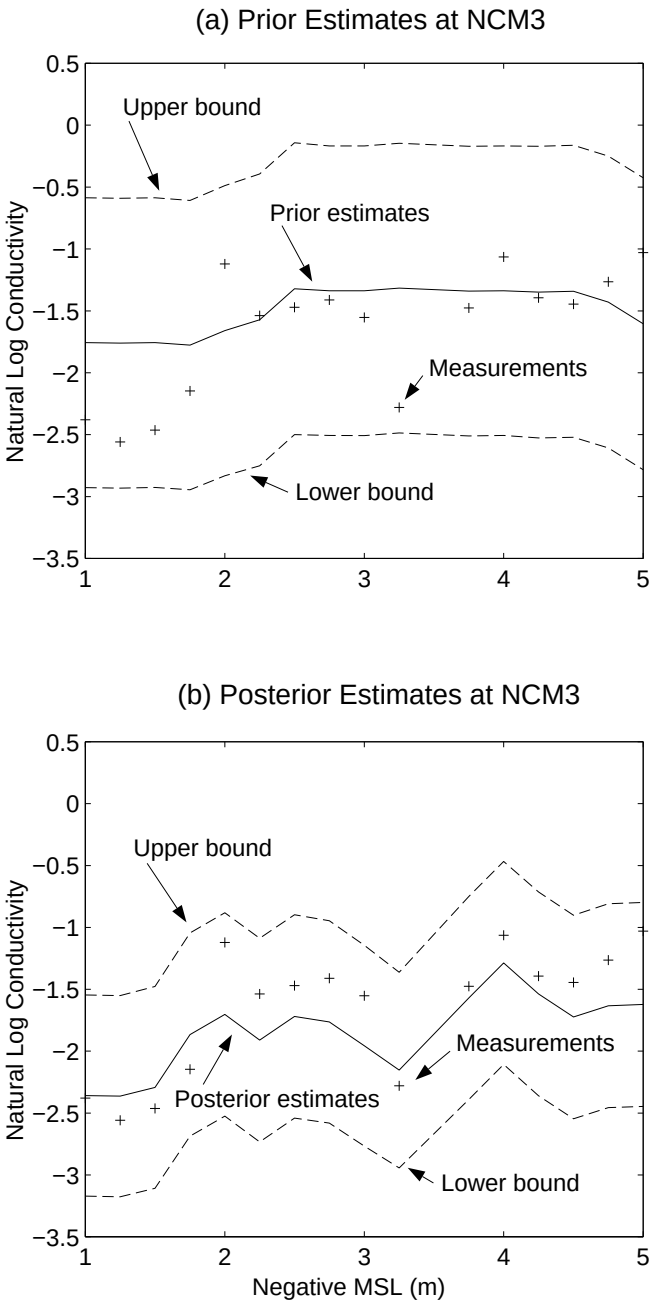


Figure 2.9: 95% confidence intervals for testing well NCM3.

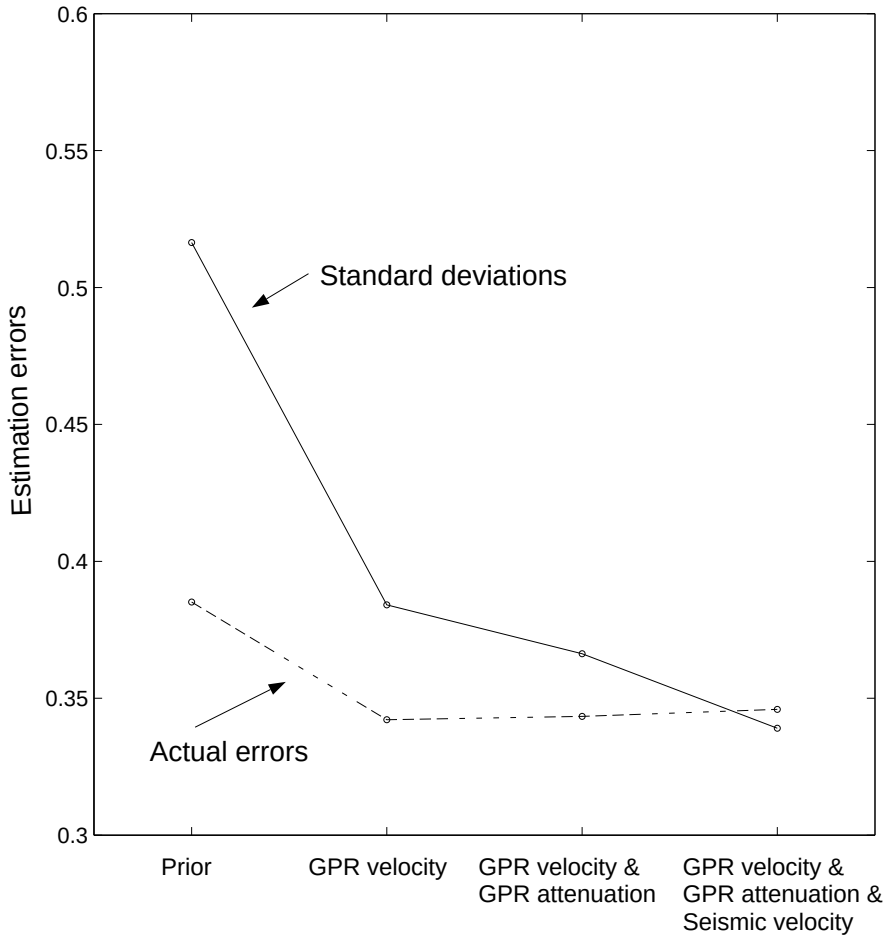


Figure 2.10: Comparison of the standard deviations and actual errors, which are the space averages along the wells of the absolute differences between actual and es timated values.

and calculate their reductions in the standard deviations, given by

$$\frac{\sigma_{prior} - \sigma_{post}}{\sigma_{prior}} \times 100\%, \tag{2.10}$$

where $\sigma_{prior}$ and $\sigma_{post}$ denote the vertical averages of the standard deviations of the prior model and of the posterior model at a testing well, respectively. Table 2.2 summarizes the averaged reductions over the ten testing wells for different models. GPR and seismic velocity were found to be more useful than GPR attenuation as only one type of geophysical data is used. The zero reduction in the standard deviations by GPR attenuation reflects either that the current method does not effectively account for the non-linear relationship between the log-conductivity and the GPR attenuation, or that the GPR attenuation is non-informative at this site. As two or more types of geophysical data are used, the average reductions in the standard deviations increase but only incrementally. This is the outcome of the various degrees of correlations that exist between the various types of data which leads to redundancy in data and hence to only minor improvements.

## 2.5 Discussion and Conclusions

This chapter paper explored the use of geophysical tomographic data for hydraulic conductivity estimation using a Bayesian framework. The prior estimates were inferred from the hydraulic conductivity data measured at wellbores, and the posterior estimates were obtained by updating the prior using co-located geophysical data.

Table 2.2: Reductions of the standard deviations using various types of geophysical data

| Number of Data Types | Data Types | Average Reduction (%) | Standard Errors (%) |
|---|---|---|---|
| 1 | $v_g$ | 25.3 | 3.5 |
| | $\alpha$ | 0.0 | N/A |
| | $v_s$ | 29.2 | 3.5 |
| 2 | $v_g, \alpha$ | 28.7 | 3.6 |
| | $v_g, v_s$ | 31.9 | 2.7 |
| | $\alpha, v_s$ | 31.0 | 4.1 |
| 3 | $v_g, \alpha, v_s$ | 34.0 | 4.1 |

*$v_g$ GPR velocity, $v_s$ seismic velocity and $\alpha$ GPR attenuatio n*

Both the prior and the posterior estimates were compared with the actual measurements to evaluate the usefulness of geophysical data for hydraulic conductivity estimation. The key findings of this study are summarized below.

Geophysical tomographic data hold the potential to improve estimation of hydraulic conductivity even when log-conductivity displays small variations and geophysical data vary over narrow ranges. This is possibly true also for domains of large variability, since large variability implies large contrasts in the geophysical measurements and consequently, better correlations between the hydraulic conductivity and the geophysical measurements. Thus, the advantages of using geophysical data in the hydraulic conductivity estimation may become more evident, as shown in *Rubin et al.* [1992], *Copty et al.* [1993], *Hyndman et al.* [1994], *Hubbard et al.* [1997], *Ezzedine et al.* [1999] and *Hubbard and Rubin* [2000].

The Bayesian approach coupled with the normal linear regression model is effective

in combining geophysical data into hydraulic conductivity estimation. Petrophysical models relating hydraulic conductivity to geophysical measurements are often non-linear and hard to define [*Hyndman et al.*, 1994; *Ezzedine et al.*, 1999], yet well-defined models are critical for successfully employing geophysical data in the estimation. In this study, petrophysical models are summarized in the form of likelihood functions, and each of the likelihood function is expressed as the product of several conditional pdfs (equation 2), which were defined using the normal linear model. This method alleviates the difficulty commonly encountered in the inference of petrophysical models for multivariate dependent variables by following a systematic model selection procedure. Another advantage of the method is that the normal linear model provides flexibility in fitting the non-linear relations between hydraulic conductivity and geophysical measurements since the likelihood functions can be of an arbitrary shape even if each of the conditional pdfs (equation 2) are assumed to be normal. However, this method is limited in situations where each of the conditional pdfs (equation 2) is multimodal and asymmetrical. In this case, other techniques, such as described in *Ezzedine et al.* [1999] and *Hubbard and Rubin* [2000]., are needed.

Using site-specific petrophysical models, as done in this study, rather than empirical relations obtained in the laboratory is a rational and efficient way to employ geophysical data in estimation of hydraulic conductivity. Petrophysical relations between hydraulic conductivity and geophysical measurements in unconsolidated sediments are more difficult to obtain compared to those in consolidated sediments or

rocks [*Marion*, 1990; *Knoll*, 1996; *Bachrach and Nur*, 1998; *Bachrach et al.*, 2000] . No empirical models are currently available to relate hydraulic conductivity to geophysical measurements in unconsolidated sediments, due to the disparity between the frequencies used in the laboratory and those employed in the field.

# Chapter 3

# Bayesian Method for Hydrogeological Characterization using Borehole and Geophysical Data: Theory and Application to the Lawrence Livemore National Laboratory Site

## 3.1 Introduction

Combining ground-surface or crosswell geophysical surveys with well logs for enhancing the quality of subsurface characterization has been the goal of recent studies. The primary motivation has been the recognition that geophysical surveys offer unique opportunities for improving crosswell interpolation, and are particularly promising in situations of data scarcity. *Hyndman et al.* [1994] developed an inversion algorithm that employs both seismic crosswell travel times and solute tracer concentration to estimate the interwell geology and therefore the hydraulic parameters. *Sheets and*

*Hendricks* [1995] used regression techniques to build a site-specific petrophysical relationship between the soil water content estimated from borehole neutron probes and the bulk electrical conductivity of the soil estimated from electromagnetic (EM) induction surveys, and this model was then used for mapping the soil water content. *Daily et al.* [1992] conducted an infiltration experiment to build a site-specific regression model between the resistivity and the moisture and showed the potential capability of Electrical Resistivity Tomography (ERT) to monitor capillary barriers performance and flow in the vadose zone. *Doyen* [1988] used cokriging to estimate porosity from surface seismic data and well logs. *Cassiani et al.* [1998] included seismic tomography data and sonic data using a geostatistical approach to improve the hydraulic conductivity estimation. *Lucet and Mavko* [1991] combined crosswell seismic tomography, logs and petrophysical relationships between porosity, velocity and clay content to estimate porosity and lithofacies. *Rubin et al.* [1992] and *Copty and Rubin* [1995] used a Bayesian approach and maximum likelihood principles to combine seismic velocity with sparsely measured hydraulic conductivity and pressure for the purpose of mapping the spatial distribution of the hydraulic conductivity. *Hubbard et al.* [1997] used a similar approach to incorporate the spatial distribution of dielectric constant obtained from ground penetrating radar (GPR) to estimate soil saturation and permeability in the vadose zone. More recently *Hubbard et al.* [1999] combined acoustic tomography with borehole data to estimate the spatial covariances of the log-conductivity.

A few observations based on these studies are as follows: (i) No universal methods or petrophysical models are available for converting geophysical attributes to hydrogeological properties; (ii) The most challenging problem is tying well-logging measurements to the geophysical surveys. This issue involves problems of scale disparity between different measurements and inconsistencies in the methods used for data acquisition and interpretation. The last point is manifested by noticing that resistivity at the Lawrence Livemore National Laboratory (LLNL) site was measured using several tools with different support volumes and some of them lead to dramatically different results.

This chapter investigates the use of geophysical surveys for mapping lithofacies and thus soil properties in the subsurface using a Bayesian approach. The study focuses primarily on the issues associated with the assimilation of weakly or non-linearly correlated data with different spatial resolutions in a geologically complex environment.

Section 3.2 introduces the LLNL superfund site, followed by geostatistical analyses of the data. A petrophysical relationship between lithofacies and geophysical attributes is also presented. Section 3.3 outlines in detail the proposed approach for data assimilation, and section 3.4 introduces the synthetic electromagnetic survey. Section 3.5 discusses Bayesian updating of pre-simulated lithofacies and resistivity random fields and evaluates the effectiveness of this method, and some findings of this study are summarized in section 3.6.

## 3.2 Site Description, Sources of Data, and Geostatistical Analysis

### 3.2.1 Lawrence Livermore Superfund Site

Volatile organic compounds (VOC) were used at the LLNL superfund site (Figure 3.1) as solvents when the site was an active Naval Air Force Base in the 1940's. Fuel petroleum hydrocarbons associated with a gasoline spill have also contaminated the underlying aquifer. The VOCs are classified as mainly Trichloroethylene (TCE), Tetrachloroethylene(PCE) and Chloroform. Tritium and Chromium are also present but in smaller concentrations [*Noyes*, 1991]. The site is located in an unconsolidated alluvial basin. The hydrogeology of the area is very complex, but a considerable amount of geological, geophysical, hydraulic and geochemical data are available. The data provide a unique opportunity to study the relationship between hydraulic conductivity and sediment texture. This study focuses on the area near treatment facility D (TFD) shown in Figure 3.1, and the detailed locations of the boreholes used in the analysis are depicted in Figure 3.2a.

The contaminants are distributed within a thick, complex sequence of unconsolidated alluvial sediments [*Blake et al.*, 1995]. A hydrostratigraphic analysis has been conducted to divide this sequence of layers into hydrostratigraphic units (HSUs). These latter are defined as sedimentary sequences whose permeable layers show evi-

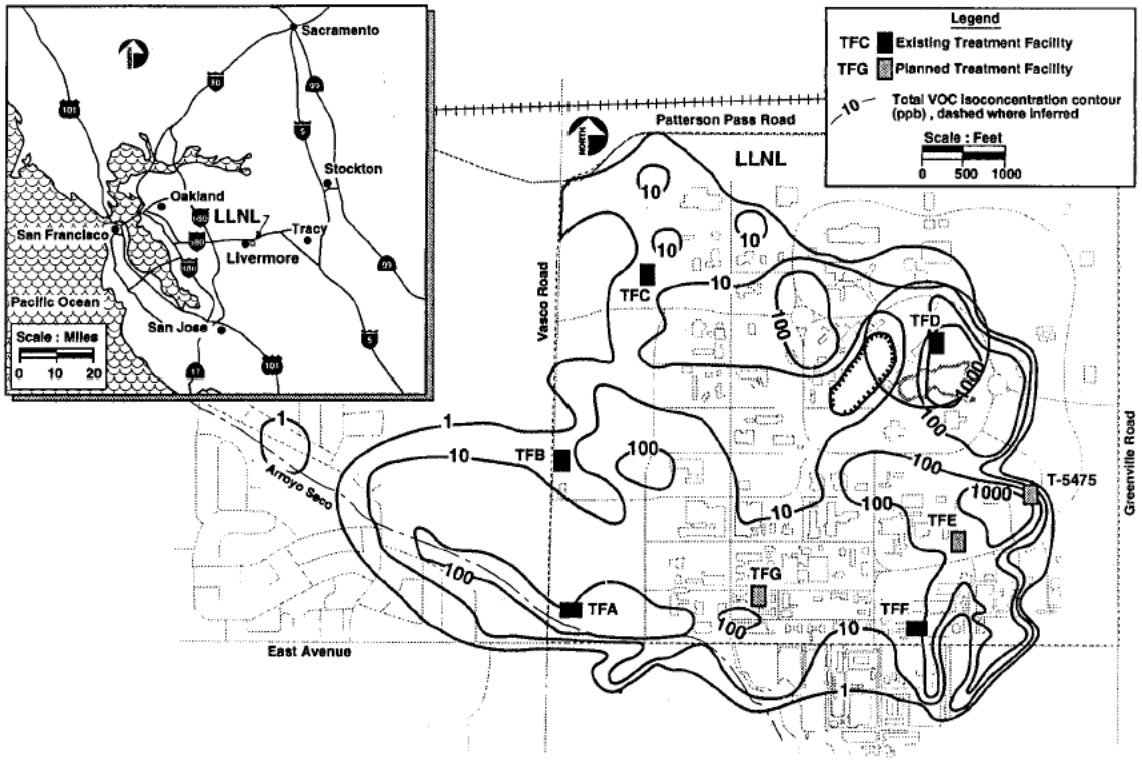


Figure 3.1: Site map of LLNL showing treatment facility (TF) areas and total volatile organic compounds (VOCs) contoured without respect to depth [*Blake et al.*, 1995].

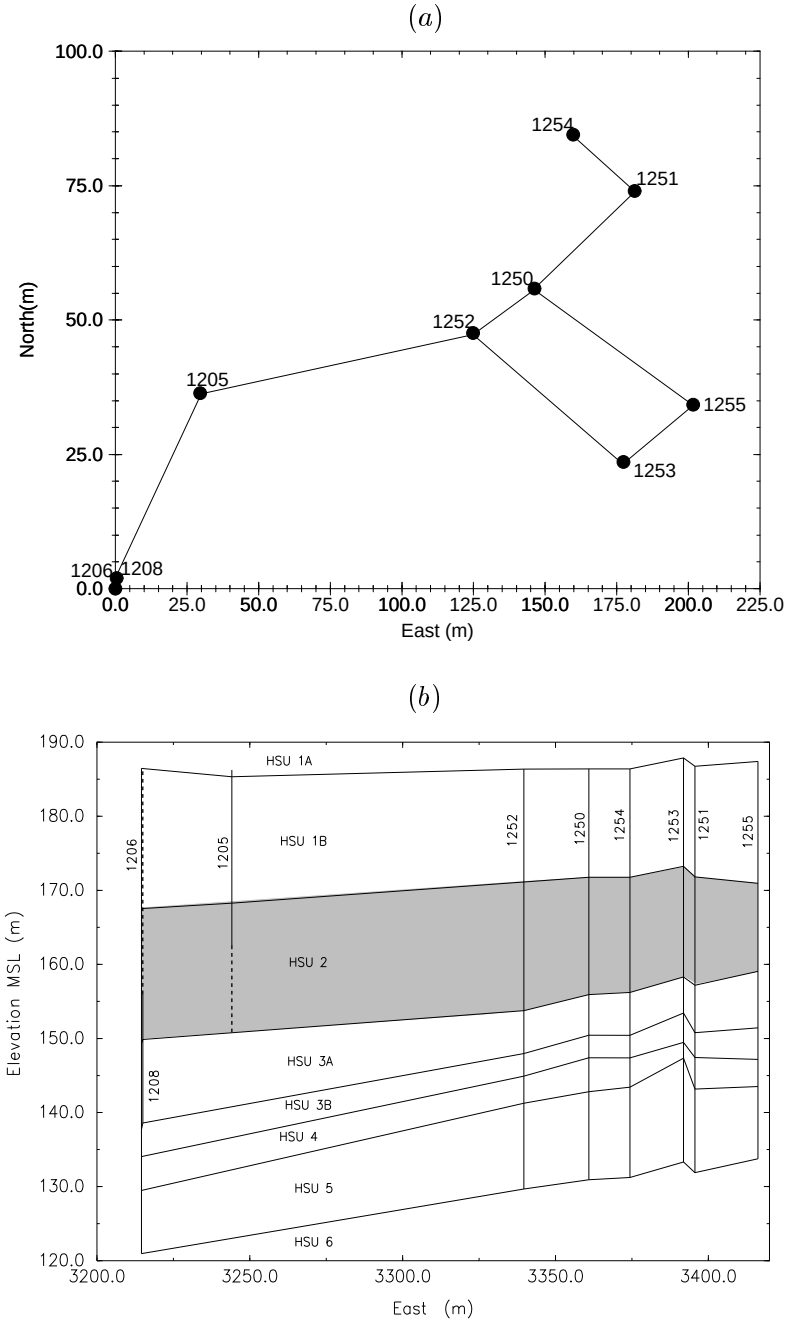


Figure 3.2: (a) Location of the wells available for the present study in TFD. Nine wells are depicted and labeled as 1205, 1206, 1208, and 1250 through 1255. (b) Vertical cross section over all hydrostratigraphic units (HSUs) through all wells depicted on (a). Vertical dash lines represent missing data, and HSUs are referred by their name HSU 1 through 6.

dence of hydraulic connectivity, using several complementary sources of information [*Blake et al.*, 1995], including chemical (concentration in ground water and soil), geological (lithological core description), geophysical (wire-line borehole electrical logs), and hydrogeological (hydraulic well tests, hydraulic communication between layers). Thicker aquitards were also defined as HSUs, while minor aquitards define HSU boundaries across which little or no vertical hydraulic leakage takes place (Figure 3.2b).

### 3.2.2 Lithological and Geophysical Raw Data

The study focuses on the cross section between wells 1206, 1208, 1205, 1252, 1250, and 1251-1254 (Figure 3.2a) within HSU2 as shown in Figure 3.2b. Types of data collected along the wells include geophysical and lithofacies well logs, which have different spatial resolutions, varying from 3 cm to 15 cm, along the vertical direction. Spatial statistics of the geophysical attributes and lithofacies are inferred from all the data available at those wells.

The collected geophysical log data include induction resistivity, short and long normal resistivity, spontaneous potential, single point resistance, guard resistivity, caliper, and gamma-ray, among others. A general description of those log types is given by *Keys* [1997]. Lithofacies were classified as 16 different types, including gravel, clay, sand, silt, and their mixtures, such as, gravely clay and clayey sand. To simplify the lithofacies mapping and because the main concern is to map the high and low

hydraulic conductivity zones, only two main classes will be used: 1) silt, including silt, clay, and their mixtures, and 2) sand, including sand, gravel, and their mixtures.

### 3.2.3 Geostatistical well log analysis

Since HSU2 is not horizontal or not defined by constant thickness (Figure 3.2b), the vertical coordinates were normalized by the average thickness of the HSU, which is approximately 17 m (Figure 3.3). Indicator semivariograms were used to describe the spatial variability of lithofacies based on a binary representation for sand and silt, and semivariograms were used to describe the spatial variability of gamma-ray (G) and resistivity (R). Those variables were investigated with and without log transformation. Semivariograms of other geophysical attributes within each lithofacies were also investigated, but only the statistics of these three variables are presented because they form the basis of this method.

#### Lithofacies Indicator Semivariograms

Based on the lithofacies classification and adopting an indicator coding of 0 for sand and 1 for silt, a geostatistical analysis was performed. Figures 3.4a and 3.4b show the vertical and horizontal indicator semivariograms. Because the vertical semivariogram was computed after normalizing the depth by the thickness of HSU2, the maximum lag is equal to the average thickness. The volume fractions of silt ($p$) and sand ($1-p$) are 0.48 and 0.52, respectively. The sills of the semivariograms are 0.25,

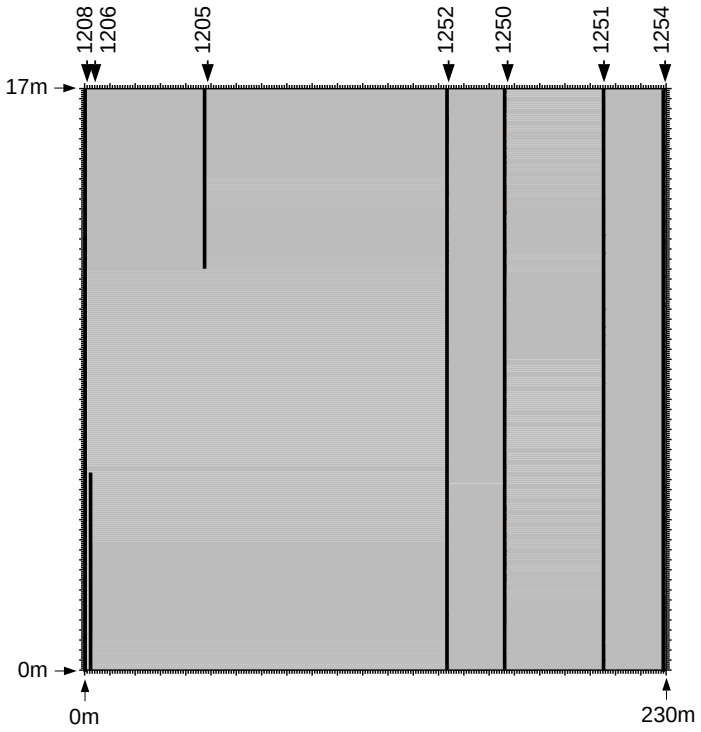


Figure 3.3: Vertical cross section of the present study over HSU2. Distances are reported from well 1206 and along the cross section, and available data along the wells are depicted using continuous vertical lines.

which is approximately equal to the theoretical value of the variance $p(1-p)$. The fitted theoretical models are exponential with a range of 1.5 m in the vertical direction and 30 m in the horizontal direction.

**Resistivity Measurement Analysis**

The analogies between the flow of electrical current and fluid through porous media have made electric logging of formation resistivity a commonly employed technique in geophysical prospecting [*Keys*, 1997]. Since a crosswell electromagnetic resistivity survey is considered at the LLNL site, resistivity well logs were used as the primary link for correlating the tomographic survey with other soil properties.

Semivariograms of induction resistivity, guard resistivity and short and long resistivity were investigated, but well-defined, long-range patterns of spatial correlation of the resistivity were not found. Despite the apparent lack of the spatial correlation, the combination of induction resistivity with gamma-ray and lithofacies logs offers an opportunity for indirect projection of gamma-ray and lithofacies pairs into resistivity.

Induction resistivity among all the resistivity logs was chosen as the primary candidate for correlating with the resistivity survey based on several reasons. (1) The measurement procedure does not require conductive fluid in the borehole or direct physical contact with the formation. (2) Induction tools minimize the contribution of the borehole, invaded zone and surrounding formations on the measurement. (3) Induction logs are automatically corrected for skin effect during recording. (4) Al-

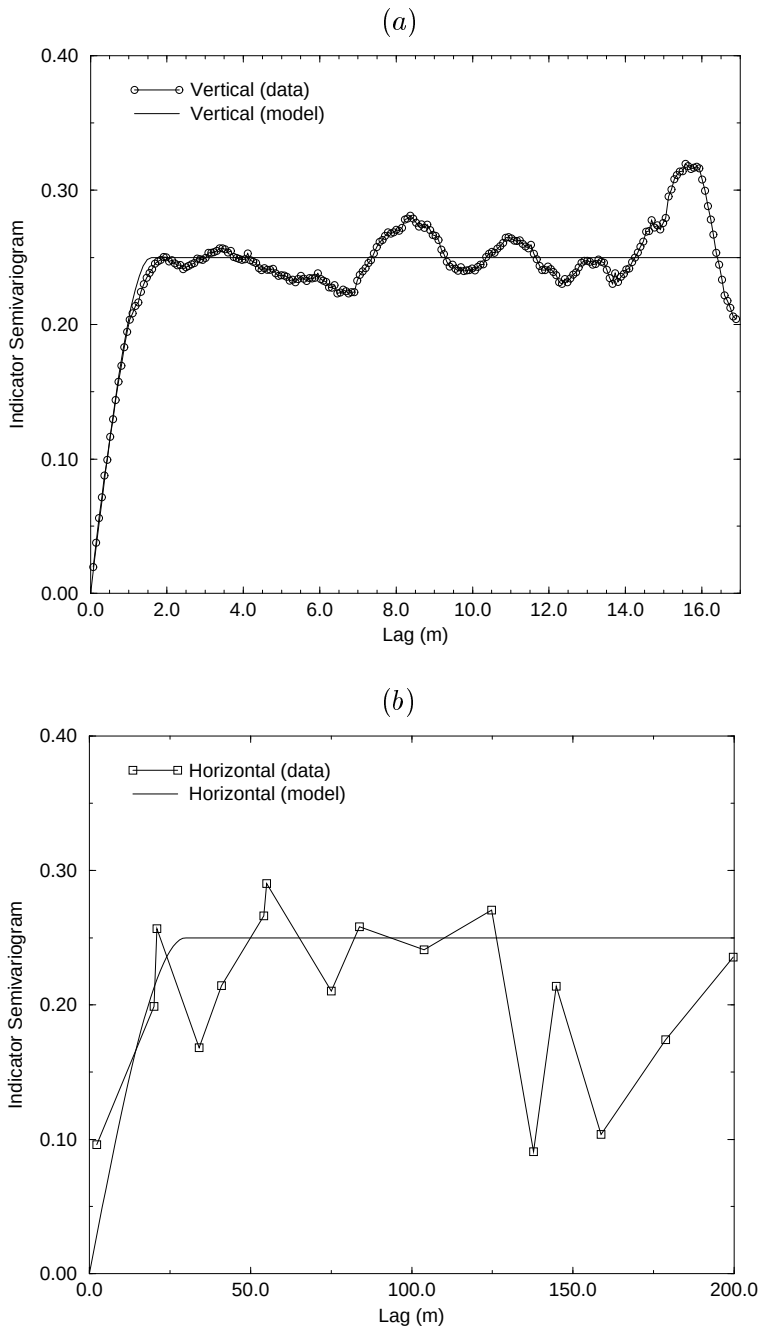


Figure 3.4: Indicator experimental and theoretical semivariograms: (a) vertical direction and (b) horizontal direction. Both theoretical semivariograms are found to be exponential.

though induction tools were designed for nonconductive borehole environments, they were found to yield excellent measurements in water-based mud, provided that the water is not too salty, the formation is not too resistive, and the borehole diameter is not too large [*Keys*, 1997]. The induction log measurements at the LLNL are of excellent quality, and this reinforced the previous choice.

### Gamma-Ray Analysis

Gamma-ray logs measure naturally occurring gamma emissions around the borehole. The sources of the radioactive decay series in nature are primarily Potassium 40, Uranium 238 and 235, and Thorium 232 [*Serra*, 1984]. Potassium 40 is by far the most abundant radioactive isotope found in sediments. As the content of Potassium 40 increases, the response of the gamma-ray probe increases. Gamma-ray response decreases from shale and clay, to silt-stone, to sandy silt-stone, to sandstone and gravel.

Conversion of gamma-ray measurements to shaliness helps to remove inconsistencies in the data introduced by using different tools and calibration techniques [*Doveton*, 1986; *Hill*, 1986]. Shaliness for unconsolidated rock is given by [*Serra*, 1984]:

$$S = 0.083[2^{3.7I_G} - 1], \tag{3.1}$$

where $I_G$ is the gamma-ray index defined as follows [*Serra*, 1986]:

$$I_G = \frac{G - G_{Min}}{G_{Max} - G_{Min}}. \tag{3.2}$$

Figures 3.5a and 3.5b depict the vertical and horizontal semivariograms of the shaliness and their corresponding fitted models. The best fit was found to be the Gaussian model with a nugget of 0.011 m and ranges of 2.5 m in the vertical direction and 25 m in the horizontal direction. Integral scales are set to 1.46 m ($\simeq$ 1.5 m) in the vertical direction and 14.43 m ($\simeq$ 14.5 m) in the horizontal direction.

**Shaliness vs. Resistivity Relationship**

Figure 3.6a displays a cross-plot of the resistivity and the shaliness. Two main clusters are shown, corresponding to the different lithofacies. It suggests that resistivity/shaliness pairs are useful for lithofacies identification. The overlap between the sand and silt clusters indicates that a unique identification of lithofacies based on resistivity and shaliness is not possible for all pairs. The main reason for the overlap between the two clusters is data reduction: the lithofacies classification originally consisted of 16 members, but now only two. Despite the ambiguous interpretation of several pair combinations, it appears that this cross-plot is a good analytical tool. This is one of the fundamental results of this analyses so far because it suggests a systematic approach for tying the resistivity survey with well logging information, which is fully developed in Section 3.3.

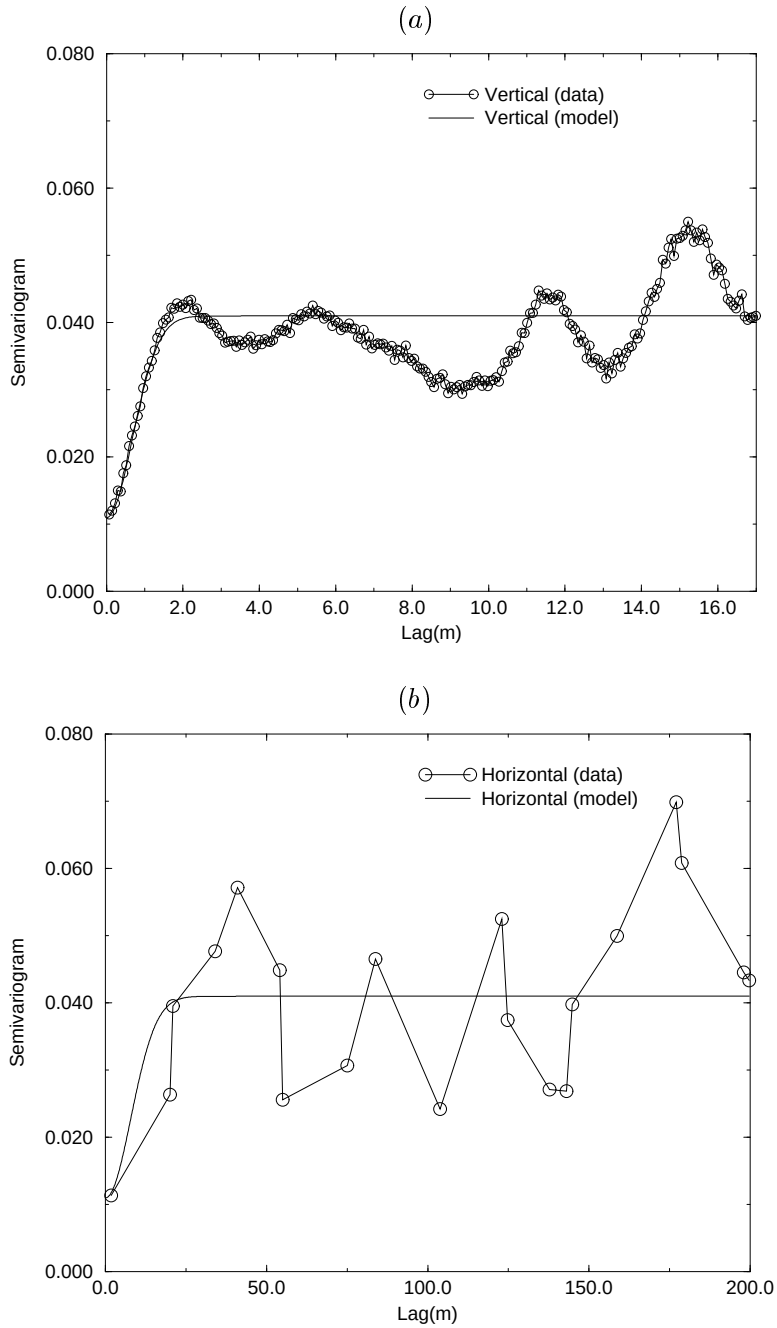


Figure 3.5: Shaliness experimental and theoretical semivariograms: (a) vertical direction and (b) horizontal direction. Both theoretical semivariograms are found to be Gaussian.

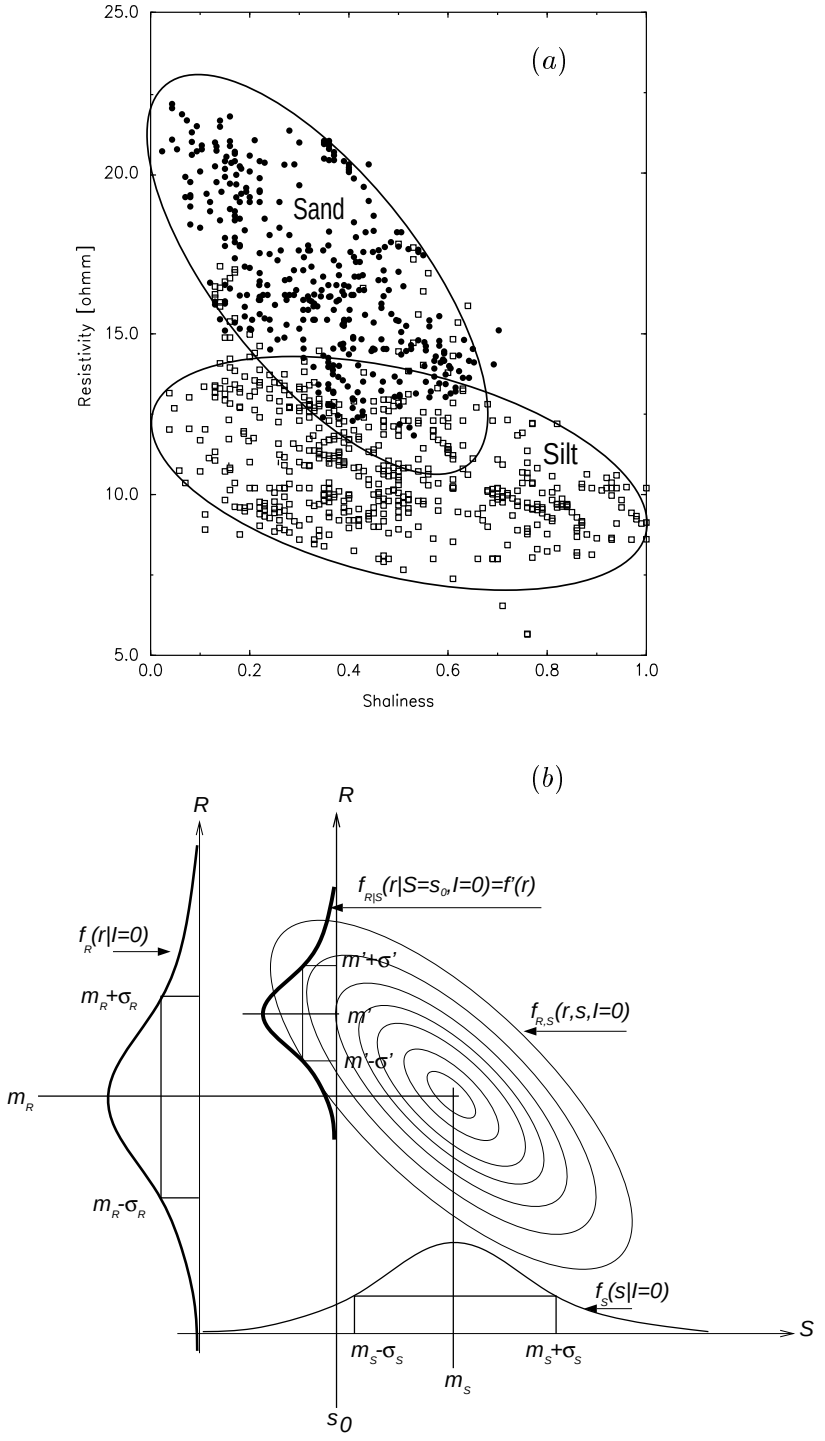


Figure 3.6: (a) Petrophysical relationship between shaliness and resistivity plotted from available data at the wells crossing HSU2. (b) Generic scheme for constructing resistivity pdfs to conditional lithofacies and shaliness.

Figure 3.6a was obtained using all well log data within the HSU2. The use of shaliness instead of gamma-ray activity considerably improved the clustering analysis. This analysis has been applied to other HSUs as well, and a behavior similar to Figure 3.6a was observed in all the cases.

## 3.3 Bayesian Data Assimilation

In an ideal situation, the geophysically measured attributes may correlate well with the hydrogeological ones, such as permeability, and the conversion of the geophysical survey to a hydrogeological distribution map is straightforward. In a more realistic situation, such as the one described here, the conversion of the geophysical attributes to the hydrogeological ones is convoluted and non-unique. Couple of difficulties exist in the implementation of the geophysical survey. Firstly, the survey resistivity is of relatively low resolution, but high-resolution permeability images need to be developed. This causes a scale disparity problem. Additionally, although the database available at the site was developed over many years, some types of data are still not available. For example, the cores were not tested for permeability.

The purpose of this section is to develop a conceptual data-driven approach for lithofacies mapping based on the well log data. This approach is general in its basic principles, but meanwhile is site-specific since the employed petrophysical models are not universal. The general approach is stochastic, which is justified given the large

uncertainty associated with crosswell interpolation, with the petrophysical models, and with the interpretation of the geophysical surveys. The rationale for this approach is based on the following observations:

1. Resistivity and shaliness can be used for lithofacies identification through the cross-plot (Figure 3.6a). Once a type of lithofacies is determined, further mapping of hydrogeological properties can be pursued.

2. Facies identification based on the shaliness-resistivity cross-plot is non-unique due to some overlap between the sand and silt clusters.

3. Borehole resistivity measurements display a short correlation range, and it is impractical to develop spatial images of the resistivity using crosswell geostatistical interpolation.

4. Shaliness displays a well-defined spatial structure. It can be used for projecting resistivity measurements indirectly through a combination of geostatistical interpolation/simulation techniques, in conjunction with the nonlinear correlation with the resistivity, as expressed through the cross-plot (Figure 3.6a).

Based on these observations, an approach that consists of sequentially generating a series of collocated attributes is proposed. At the basis of the hierarchy, images of the lithofacies are generated, conditional to well logs and possibly also to the survey resistivity. Each lithofacies image serves then as the basis for generating a series of shaliness images, again conditional to well data. The shaliness images

are then used to correlate the survey resistivity with the hydrogeological attributes obtained experimentally. Since all the generated images are conditioned to the well data and have the same underlying spatial structure, they are physically plausible. The variations between the images provide a measure of the spatial variability and uncertainty associated with the estimation. The focus here is on resistivity estimation, but it can be converted to porosity and hydraulic conductivity through well-known models such as Archie's, Waxman-Smits', Kozeny-Carmen's [*Mavko et al.*, 1998], or through site-specific calibration curves [ *Daily et al.*, 1992].

### 3.3.1 Outline of the approach

Figure 3.7 is a flow chart showing the proposed approach, which includes the following four steps.

**Step 1: Generating lithofacies images using sequential indicator simulation (SIS)**

The lithofacies is defined through an indicator variable $I$ according to [*Rubin*, 1995]:

$$I(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is in silt} \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

Note that boldface letters denote vectors, i.e., $\mathbf{x}$ is a location coordinate vector. Lowercase $i$ is a realization of the spatial random function (SRF) $I$, which is assumed

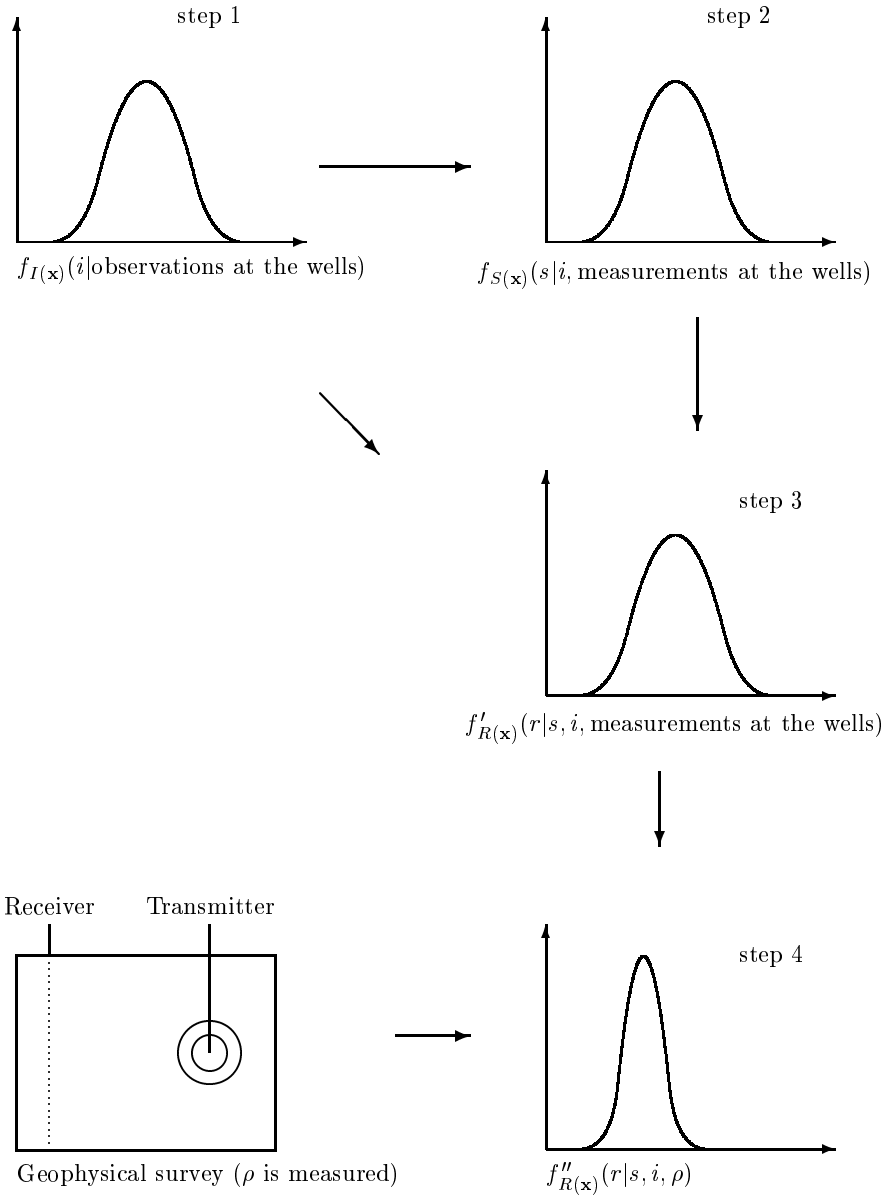


Figure 3.7: Flow chart of the approach

stationary and characterized through its unconditional expected value:

$$E\{I\} = p, \tag{3.4}$$

where $E$ denotes the expected value operator and $p$ is the volume fraction of silt. The spatial structure of the variable is shown in Figure 3.4a and 3.4b and defined through the semivariogram:

$$\gamma_I(\mathbf{x}, \mathbf{x}') = \frac{1}{2} E\{(I(\mathbf{x}) - p)(I(\mathbf{x}') - p)\}, \tag{3.5}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are two location vectors.

The unconditional statistics define the crudest level of probabilistic characterization. A more advanced characterization is possible through the conditional moments of $I$. These statistics are the cornerstone of the SIS algorithm [*Deutsch and Journel*, 1998; *Rubin and Bellin*, 1998]. The SIS algorithm consists of computing the expected value of $I$ conditional to the borehole data:

$$p^c = E^c\{I\} = E\{I|\{\text{measurements}\}\} \tag{3.6}$$

with a superscript $c$ denoting conditioning.

A realization of $I$ can be drawn once $p^c$ is defined because $I$ is binary and $p^c$ is statistically exhaustive. The process of computing $p^c$ and drawing realizations is done sequentially over a grid. Spatial continuity is maintained by conditioning $I$ on both well data and the previously generated values. Specifically,

$$p^c = p + \sum_{n=1}^{N} \lambda_n (I(\mathbf{x}_n) - p), \tag{3.7}$$

where the weights $\lambda_n$ are obtained by solving the following system of linear equations:

$$\sum_{n=1}^{N} \lambda_n \gamma_I(\mathbf{x}_m, \mathbf{x}_n) = \gamma_I(\mathbf{x}_m, \mathbf{x}), \quad m = 1, \cdots, N. \tag{3.8}$$

The important point to note is that $N$, the number of lithofacies measurements, includes all the observations as well as all the values previously generated at other locations.

**Step 2: Generating shaliness images**

This step is similar in principle to the previous one. The differences are in the fact that (i) the shaliness $S$ is not a binary variable and (ii) the spatial structure of the shaliness may be different between the sand and silt lithofacies; i.e. $\gamma_{S|i}$, the semivariogram of the shaliness $S$, depends on the lithofacies $i = 0$ or 1:

$$\gamma_{S|i}(\mathbf{x}, \mathbf{x}') = \frac{1}{2} E\{(S(\mathbf{x}|i) - m_{S|i})(S(\mathbf{x}'|i) - m_{S|i})\}. \tag{3.9}$$

The univariate and spatial statistics of the shaliness were discussed in Section 3.2. Here the conditional mean and variance of $S$ are computed, which uniquely determines the Gaussian distribution of the shaliness at location $\mathbf{x}$, using the kriging equation. With the use of a Gaussian random generator, a local value for $S$ is drawn from the distribution. Defining the shaliness $S$ through its mean $m_{S|i}$ and its semivariogram $\gamma_{S|i}$ for a given lithofacies $i$, the conditional mean $m^c_{S|i}$ and variance $\sigma^{2c}_{S|i}$ of the shaliness are given by:

$$m^c_{S|i} = \sum_{l=1}^{N} \alpha_l S(\mathbf{x}_l|i), \tag{3.10}$$

$$\sigma^{2c}_{S|i} = \sigma^2_{S|i} - \sum_{l=1}^{N} \alpha_l Cov_{S|i}(\mathbf{x}_l, \mathbf{x}), \tag{3.11}$$

where the covariance is identified by $Cov_{S|i}(\mathbf{x}_l, \mathbf{x}) = \sigma^2_{S|i} - \gamma_{S|i}(\mathbf{x}_l, \mathbf{x})$, and the weights $\alpha_l$ are obtained by solving the following system of linear equations:

$$\sum_{l=1}^{N} \alpha_l Cov_{S|i}(\mathbf{x}_m, \mathbf{x}_l) = Cov_{S|i}(\mathbf{x}_m, \mathbf{x}),\ m = 1, \cdots, N. \tag{3.12}$$

All points $l, m = 1, \cdots, N$ are located within the lithofacies $i$.

**Step 3: Computing the resistivity prior pdfs**

Once lithofacies (sand or silt) at location $\mathbf{x}$ is identified and the corresponding shaliness is assigned, a prior pdf for the resistivity $f_{R(\mathbf{x})}(r|I = i, S = s)$ can be defined through Figure 3.6a. $R$ and $S$ denote the space random functions (SRFs) of the resistivity and the shaliness, respectively, and $r$ and $s$ denote their corresponding realizations. Figure 3.6b illustrates the joint pdf of $R$ and $S$ given $I = 0$ (sand) and the marginal pdfs $f_R(r|I = 0)$ and $f_S(s|I = 0)$. Conditioning further on $S = s_0$ leads to $f_{R|S}(r|S = s_0, I = 0)$, which is the Bayesian prior. Scarcity of data led to condition on ranges of $S$ values rather than on a single value. Figure 3.8 shows examples of $f_{R(\mathbf{x})}(r|I = i, 0.3 \leq s < 0.4)$, $f_{R(\mathbf{x})}(r|I = i, 0.4 \leq s < 0.5)$, and $f_{R(\mathbf{x})}(r|I = i, 0.5 \leq s < 0.6)$ for $i = 0, 1$.

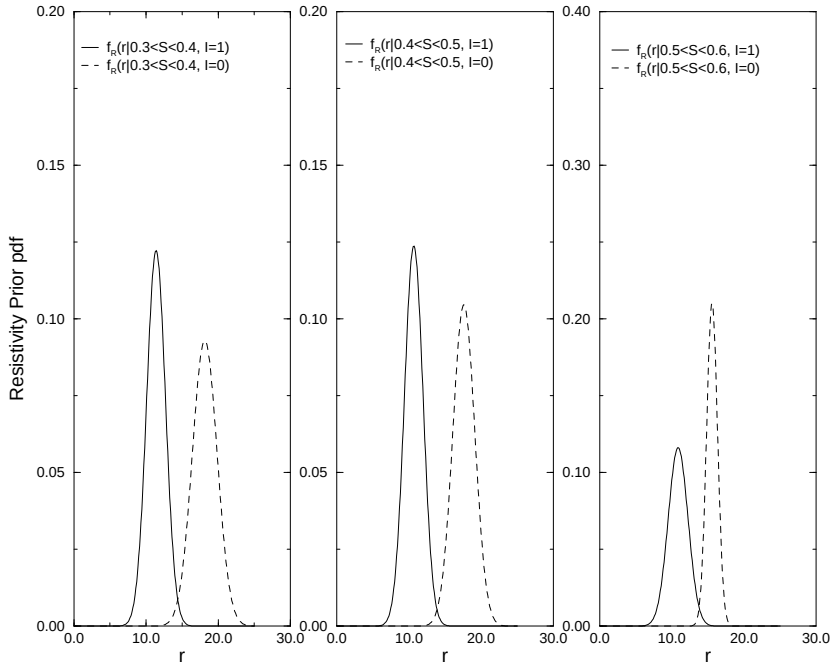


Figure 3.8: Example of prior pdf's $f'(r)$ for $I = 0, 1$ and shaliness $0.3 \leq s < 0.6$.

**Step 4: Updating $f_{R(\mathbf{x})}(r|I=i,S=s)$ based on crosswell resistivity survey $\rho(\mathbf{x})$**

Defining $f_{R(\mathbf{x})}(r|I=i,S=s) = f'_{R(\mathbf{x})}(r)$ for brevity and given collocated survey resistivity $\rho(\mathbf{x})$, the posterior pdf $f''_{R(\mathbf{x})}(r|\rho)$ can be defined through Bayes' rule [*Box and Tiao*, 1973]:

$$f''_{R(\mathbf{x})}(r|\rho) = C_R L(r|\rho) f'_{R(\mathbf{x})}(r) \tag{3.13}$$

where $L(r|\rho)$ is the likelihood function and $C_R$ is a normalized factor defined as [*Box and Tiao*, 1973]

$$C_R = \left( \int_{-\infty}^{\infty} L(r|\rho) f'_{R(\mathbf{x})}(r) dr \right)^{-1} \tag{3.14}$$

In general, $\rho$ is defined over a support volume larger than the support volume of $r$. Note that in the case of a high resolution geophysical survey, $\rho(\mathbf{x}) \rightarrow r(\mathbf{x})$, and Bayesian updating is unnecessary. In this case, $\rho$ can simply be converted to the hydrogeological properties of interest if a petrophysical model is available. That conversion will be as reliable and accurate as the petrophysical model used for conversion. This is however not a general case, and the alternative is to update $f'_{R(\mathbf{x})}(r)$ given $\rho$. Typically, $\rho$ is defined by a block of scale $\sim 3$ m or greater, and the estimate of $R$ is expected to have a scale of $\sim 1$ m. The inference of the likelihood function, $L(r|\rho)$, is critical for the success of the updating and discussed in Section 3.5. Once $f''_R(r|\rho)$ is defined, a realization of $R$ at $\mathbf{x}$ can be drawn. The entire process is repeated for all $\mathbf{x}$ until a complete image of the resistivity field is completed.

An attractive property of Bayesian updating is that the posterior $f''_{R(\mathbf{x})}(r)$ is at least as informative as $f'_{R(\mathbf{x})}(r)$. In the case of a totally non-informative likelihood function, equation (3.13) yields $f'_{R(\mathbf{x})}(r) = f''_{R(\mathbf{x})}(r)$. It is emphasized that the method does not always guarantee better estimates for a couple of reasons. Firstly, the Bayesian approach provides a pdf, not a single valued estimate. Secondly, the improvement achieved in the posterior pdf is dictated by the quality of external factors such as the accuracy of the geophysical survey and the petrophysical model.

A modification of step 1 is appropriate and useful if the lithofacies images can also be conditioned on the resistivity. As is apparent from Figure 3.6a, the lithofacies images can be improved through the resistivity survey: sand tends to be characterized by high resistivity and silt by low resistivity, although there is some overlap at midrange values. This approach calls for Bayesian updating of $p^c$ as well, through the relationship

$$p^{c\prime} = C_I L(I|\rho) p^c, \tag{3.15}$$

where $L(I|\rho)$ is the likelihood function, of a similar nature to (3.13), only relating $\rho$ to $I$ rather than $R$. $C_I$ is a normalized factor similar to $C_R$ (3.14).

### 3.3.2 Synthetic "True" Database

The concept outlined in Section 3.3.1 is demonstrated here using a synthetic example, generated to simulate closely the conditions of HSU2. Figure 3.9a shows a realization of HSU2 lithofacies conditional on the lithofacies observed at the wells.

The field is 230 m in the horizontal direction and 17 m in the vertical direction. Realizations of the shaliness and resistivity fields, generated based on the previously described method, conditional on borehole data, are given in Figure 3.9b and 3.9c. The spatial statistics used are those described in Section 3.2.

## 3.4 Electromagnetic Surveying

Field EM surveying is a complex mapping of the detailed, high-resolution $R(\mathbf{x})$ distribution into a low-resolution $\rho(\mathbf{x})$ field. In reality, the geophysical response is distorted by both data acquisition and the inversion process.

An electromagnetic survey was conducted at LLNL through polyvinyl chloride (PVC) cased wells. Two surface to borehole profiles were measured with a surface transmitter loop (frequency 11.3 kHz) and a vertical magnetic coil receiver placed in well 1250. The profiles were in the region between wells 1250-1251 and 1250-1252 shown in Figure 3.2. Seven cross-well EM data sets were collected. A vertical magnetic coil transmitter (frequency of 9.6 kHz) was placed in well 1250 and 1251. From well 1250, five data sets were collected with a vertical magnetic receiver placed successively in wells 1251 through 1255. The last two data sets were collected between wells 1251-1253 and wells 1251-1254. All data sets from the crosswell EM survey have been processed, but results are not yet available. To explore the Bayesian updating approach, synthetic surveys of the resistivity are simulated.

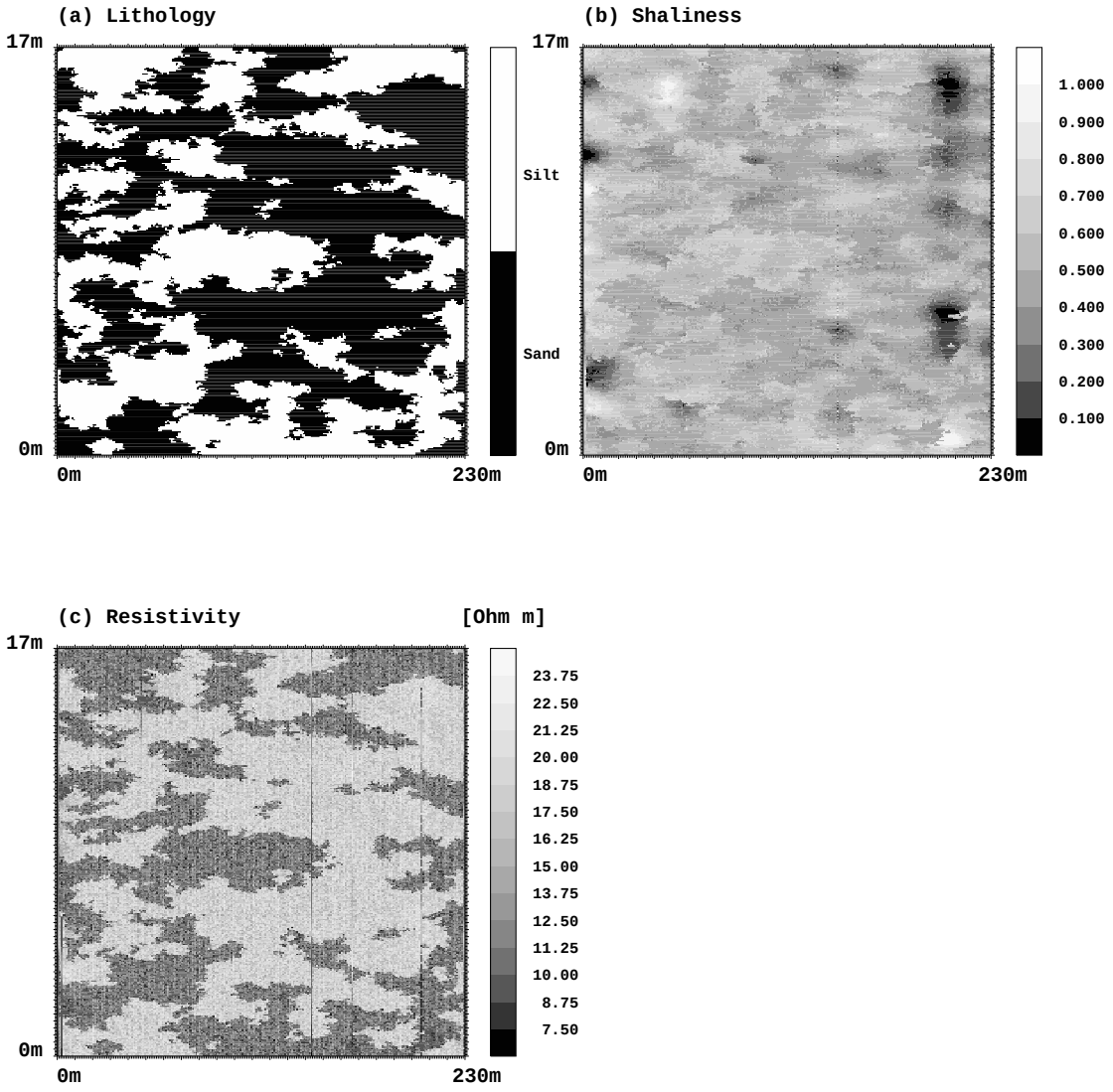


Figure 3.9: "True" geological sitting. (a) Sequential indicator simulation of lithofacies conditional on borehole data.(b) Sequential Gaussian simulation of shaliness conditional on borehole shaliness measurements. Darker shades represent sand (low clay content), and brighter shades represent silt (high clay content). (c) True resistivity random field built by projecting the "true" shaliness field using the petrophysical relationship given in Figure 3.6a.

The theoretical foundation of the EM survey is based on Maxwell's wave propagation equations as given in Appendix A. These equations couple the electric field to the magnetic one. Under reasonable approximation of low variability of the resistivity between the sand and the silt bodies (Figure 3.6a), the wave propagation problem can be reduced to an electric current diffusion problem. Identical problems have been considered in fluid flow in porous media [*Dagan*, 1989] and flow of electric currents [*Abramovich and Indelman*, 1995]. Borrowing from their results and considering the two-dimensional survey, the electrical conductivity $\kappa_b$ of a block which covers $n_x$ by $n_z$ small-scale blocks (Figure 3.10), where $n_x$ is the number of blocks in $x$ direction and $n_z$ in the $z$ direction, is given by the geometric mean:

$$\kappa_b = \left( \prod_{l,m=1}^{n_x n_z} \kappa_{l,m} \right)^{\frac{1}{n_x n_z}}, \tag{3.16}$$

where $\kappa_{l,m}$ are the small-scale blocks. This formula is applicable for blocks that are large relative to the characteristic length scale of resistivity heterogeneity. In the present case, since the characteristic length of the spatial variability is small, geometric averaging appears to be an appropriate homogenization procedure.

From simple algebra, $\rho = \left( \prod_{l,m=1}^{n_x n_z} R_{l,m} \right)^{\frac{1}{n_x n_z}}$ (see Appendix A). Hence for the purpose of this study, the resistivity of a block detected in a survey equals the geometric averaging of the small-scale resistivity. $n_x = n_z = 3, 6, 9$ will be considered. Figures 3.11a and 3.11b show results of the synthetically surveying resistivity field shown in Figure 3.9c using different resolutions. As the resolution decreases, small-scale details become obscure and fuzzy, and the range of resistivity values detected narrows.

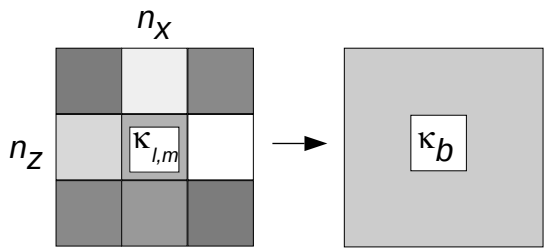


Figure 3.10: Upscaling small-scale block conductivities $\kappa_{l,m}$, $l = 1, \cdots, n_x$; $m = 1, \cdots, n_z$ into survey scale block conductivity $\kappa_b$.

## 3.5 Synthetic Case Study

The cross section shown in Figure 3.3 is investigated in this case study, assuming that Figures 3.9a to 3.9c, which were generated conditional to the borehole data, are the "true" images of that cross section. A geophysical survey of the same cross section is simulated using equation (3.16). The goal of the case study is to test the capability of the method described in Section 3.3 to reconstruct the base case's images while benefiting from the resistivity survey.

Typical images obtained using prior pdfs are only shown in Figures 3.12a to 3.12c. It is noted that these images are in good agreement with the corresponding images (Figures 3.9a to 3.9c) only in the well-sampled areas, on the right-hand side of the images. Figures 3.12a and 3.12c will be updated following the methods outlined in section 3.3.1. Updated images will be compared to the assumed "true" images depicted in Figures 3.9a and 3.9c.

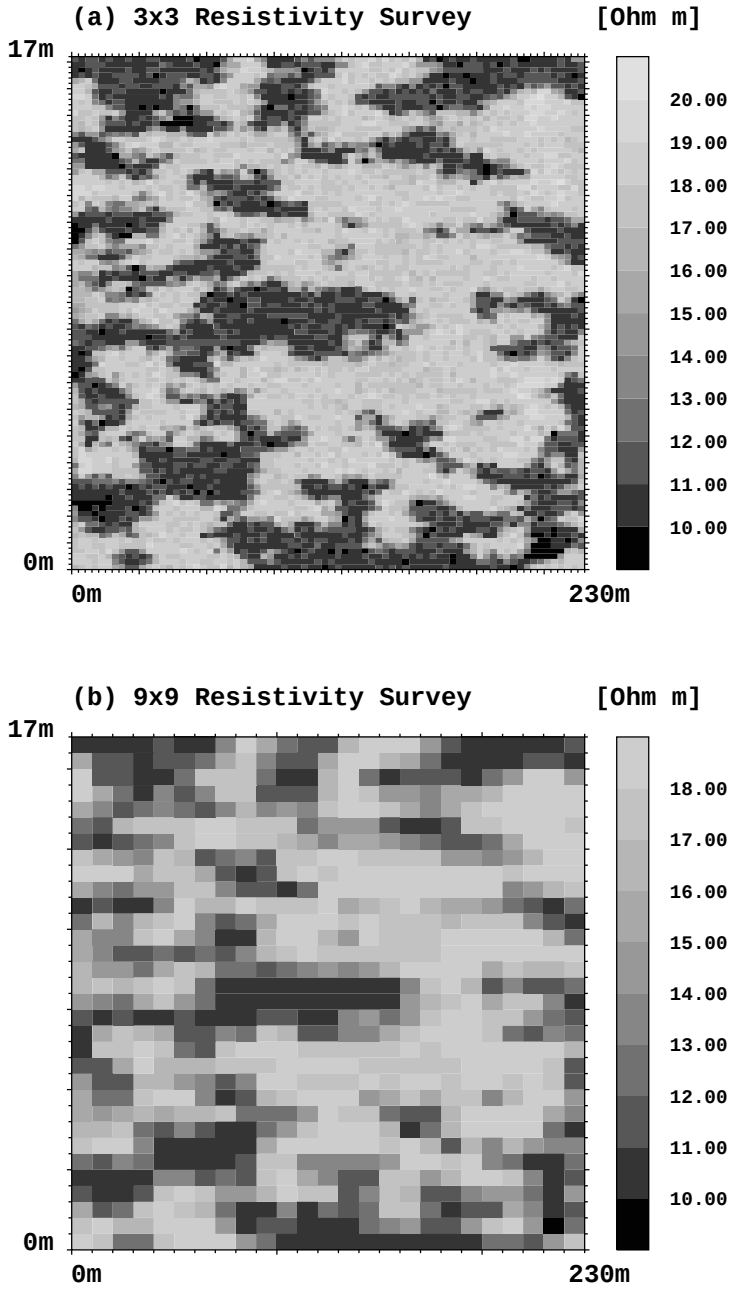


Figure 3.11: Examples of the resistivity surveys obtained by geometric averaging of the "true" resistivity field (Figure 3.9c) over (a) three and (b) nine small-scale blocks in the horizontal and vertical directions.

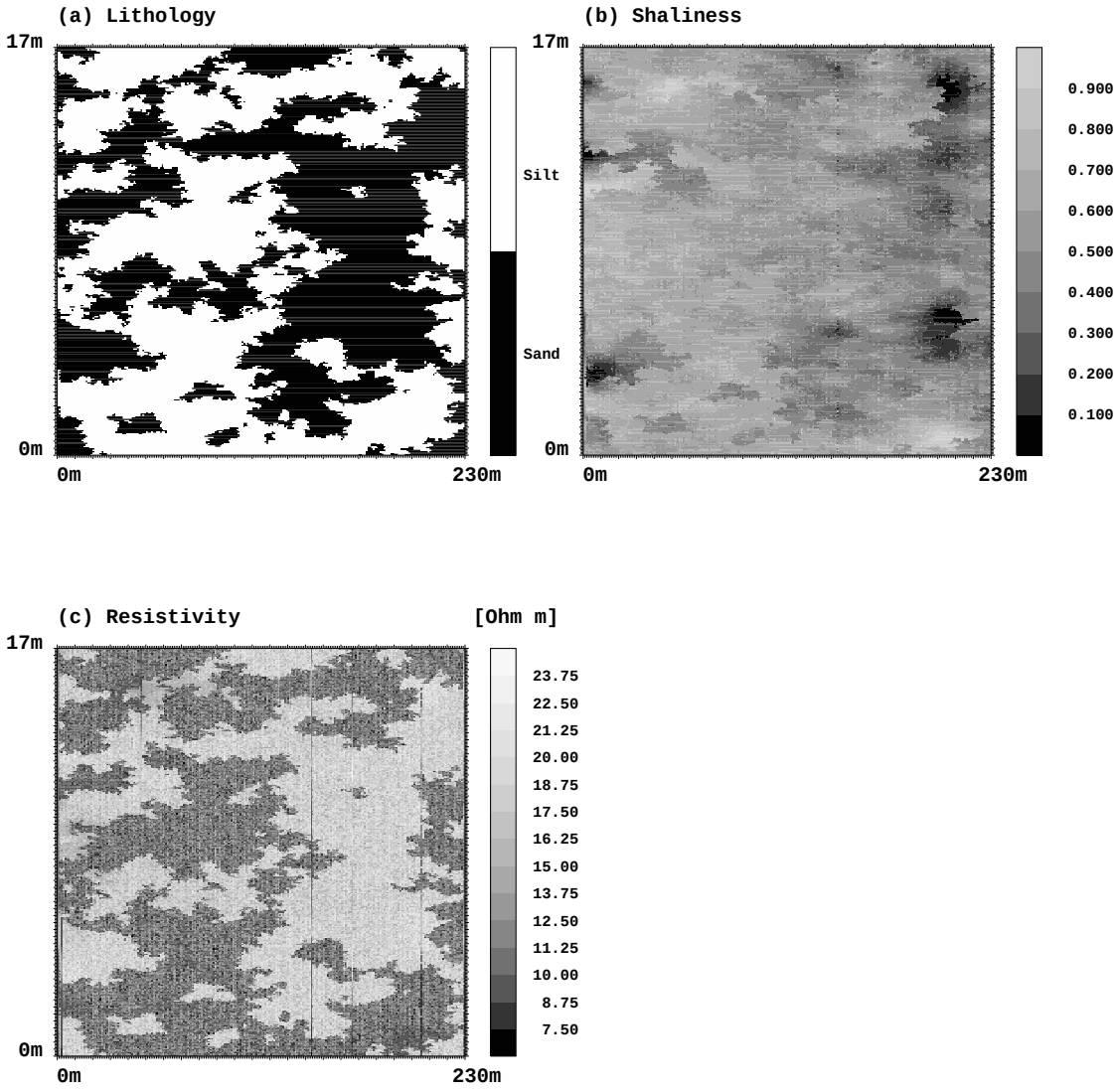


Figure 3.12: (a) Single realization of the lithofacies field obtained by sequential indicator simulation of lithofacies conditional to borehole core data. (b) Single realization of the shaliness field obtained by sequential Gaussian simulation of the shaliness conditional to borehole shaliness measurements. (c) Single realization of the resistivity field built by projecting the shaliness random field (Figure 3.12b) using the petrophysical relationship (Figure 3.6a).

### 3.5.1 Indicator Likelihood Functions and Lithofacies Image Updating

Equation (3.15) requires to infer the likelihood function $L(I|\rho)$. To identify $L$, a "training set" is used. The idea is to identify a portion of the survey area that will be drilled and cored post survey to yield a set of collocated measurements $(\rho, i)$. The dimension of the training set area should be determined such that the survey represents the entire range of conditions expected over the entire surveyed area. The sampled area needs to be ergodic in terms of bivariate $(\rho, i)$ statistics. That usually implies a dimension of several integral scales vertically, along cored wells. $\rho$ measured close to the wells can also be considered as located at the well itself. An alternative that is not pursued here is to derive the likelihood function analytically, based on upscaling rules [*Copty and Rubin*, 1995]. In the present application, the well-sampled area near well 1250 (right-hand side of Figure 3.3) was set to be the "training set", and the much less sampled area near well 1205 (left-hand side of Figure 3.3) was set as the "testing set".

$L(I|\rho)$ is determined for a given $I = i$ and $\rho = \rho_0$ by scanning the set of collocated pairs $(i, \rho_0)$ and computing the conditional probability Prob$[\rho = \rho_0|I = i]$. Equation (3.15) is then used to update the lithofacies image (Figure 3.12a) using different resistivity survey resolutions. Images of the "testing set", obtained based on equation (3.15) for $n_x n_z = 3 \times 3$, $6 \times 6$ and $9 \times 9$ resistivity surveys, are practically of the same


Figure 3.13: Posterior lithofacies image of the "testing set" (left-hand side of Figure 3.12a, using (3.15) and $3 \times 3$ resistivity survey (Figure 3.11a).

quality as without updating and differ only by a fraction less than 1% from the prior lithofacies (Figure 3.12a), even in case of high-resolution resistivity survey ($3 \times 3$) (Figure 3.13). This outcome is a manifestation of the effect of the homogenization, which obscures the resistivity-lithofacies relationship. A large number of resistivity combinations can lead to the same $\rho$ and hence to a non-unique relationship between $\rho$ and the lithofacies.

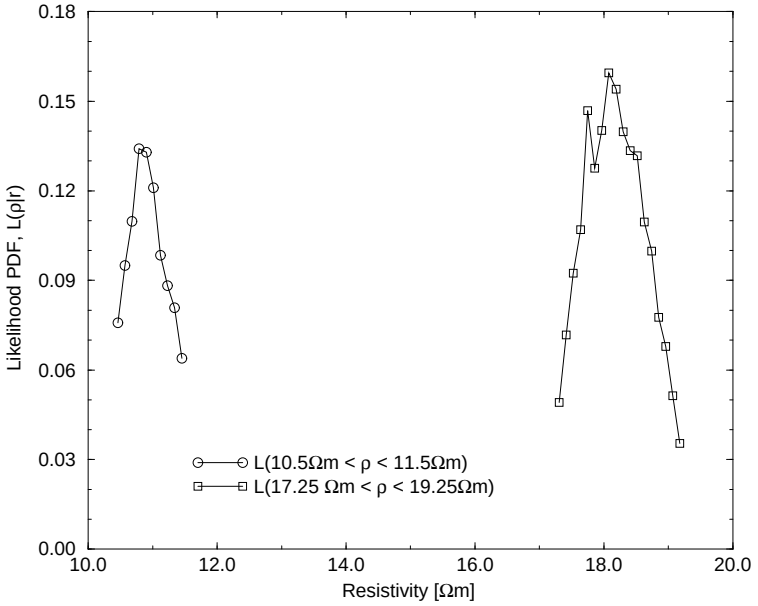


Figure 3.14: Examples of the likelihood function $L(r|\rho)$ inferred from $3 \times 3$ resistivity survey following (3.13).

### 3.5.2 Resistivity Likelihood Functions and Resistivity Image Updating

$L(r|\rho)$ of equation (3.13) is approximated here by $L(r - dr \leq \rho < r + dr|\rho)$ with a relatively small $dr$, due to the data scarcity. A couple of typical examples are shown in Figure 3.14, which are likelihood functions for two ranges of resistivity [10.5, 11.5] Ωm and [17.25,19.25] Ωm.

The effect of updating the resistivity based on surveys with different resolutions

is demonstrated in Figure 3.15. It shows that the prior and posterior resistivity pdfs at arbitrary points within the silt and sand lithofacies for various resolutions of the resistivity survey. The maximum beneficial effect is obtained, not surprisingly, through the high-resolution survey, but the positive impact of conditioning on $\rho$ is discernible even at the low-resolution surveys. The trend of reduction in impact with poorer resolution is evident, and this is an outcome of the diffuse and non-informative nature of the likelihood function as the discrepancy between the survey scale and the desired resolution scale increases.

Note that conditioning $R$ on $\rho$ does not imply that the randomly generated values will average exactly to yield $\rho$ unless special measures are taken. To honor precisely the surveyed value $\rho$, a constraint on the generated value is introduced so that the generated $r$ values over any volume corresponding to $\rho$ will average exactly to yield $\rho$. The procedure is outlined in Appendix B.

Figures 3.16a and 3.16b depict the updated resistivity fields for $n_x n_z = 3 \times 3$ and $9 \times 9$ resistivity survey. These figures should be compared with the "true" image (Figure 3.9c, left part) and with the image generated based on the prior pdfs (Figure 3.11c). It is obvious that the resistivity surveys have a significant positive impact, particularly at the high resolution.

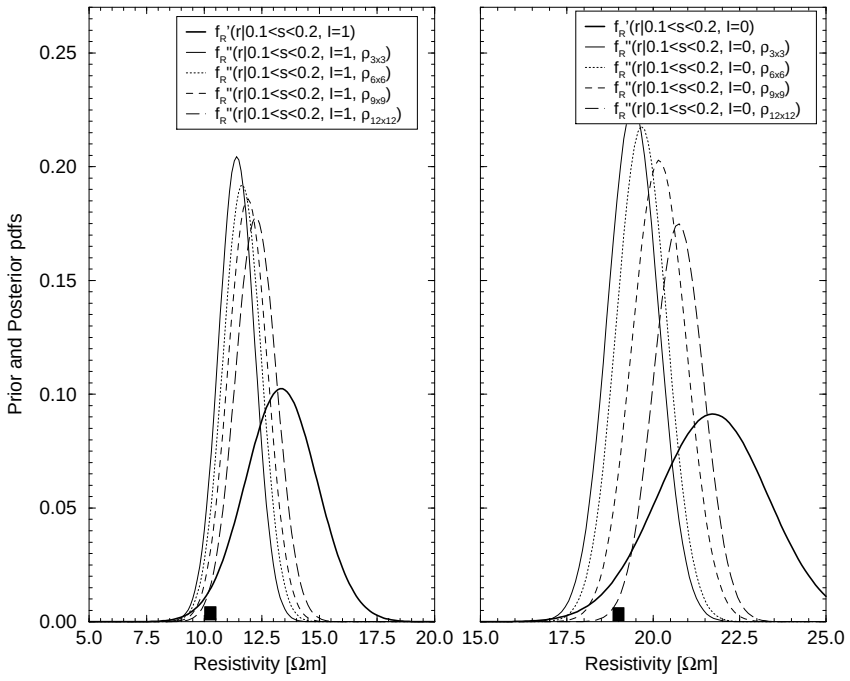


Figure 3.15: Effect of the resolution of the $n_x n_z$ resistivity survey on the posterior pdf's (prior pdf's are also plotted). The bias in the variance and the mean decreases with the increase of the resolution of resistivity survey (from $12 \times 12$, $9 \times 9$, $6 \times 6$, to $3 \times 3$). The black box denotes "true" resistivity values. Prior and posterior pdf's for shaliness between 0.1 and 0.2 in silt (left, $I = 1$), and sand (right, $I = 0$).

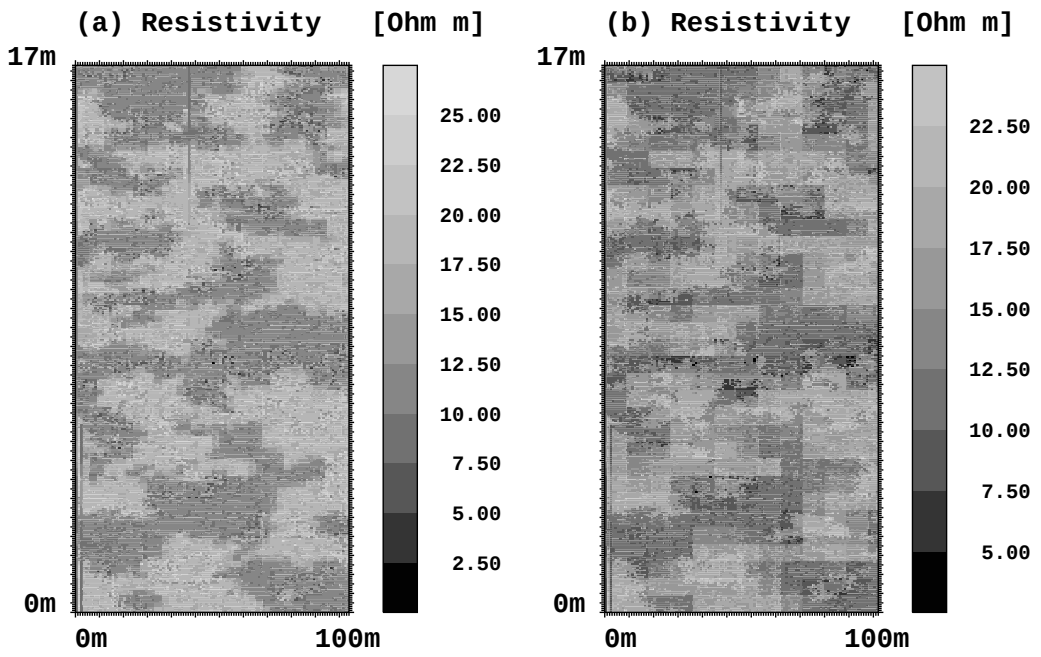


Figure 3.16: Posterior resistivity images of the testing set (left side of Figure 3.12c, using (3.13) and the resistivity surveys: (a) $3 \times 3$ resistivity survey (Figure 3.11a) and the posterior lithofacies (Figure 3.13), and (b) $9 \times 9$ resistivity survey (Figure 3.11c) and the posterior lithofacies (Figure 3.13).

### 3.5.3 Effectiveness of the Bayesian Updating

To evaluate the effectiveness of the updating procedure, the following statistic is analyzed:

$$U_k = \frac{|r_k - m''|}{|r_k - m'|} \begin{cases} U_k < 1, & \text{successeful updating} \\ \text{otherwise}, & \text{unsucesseful}, \end{cases} \tag{3.17}$$

where $k$ is a running index over all the points outside the wells, $r$ is the actual resistivity (Figure 3.9c), $m''$ is the mean of the posterior pdf $f''_{R(\mathbf{x})}(\mathbf{x})$, and $m'$ the mean of the prior pdf $f'_{R(\mathbf{x})}(\mathbf{x})$. The ratio $U$ compares the performance of the posterior and the prior pdfs. $U$ smaller than 1 indicates a successful updating procedure; otherwise, it is a diffuse likelihood and hence a non-informative survey. Figure 3.17 depicts the variation of $U$, as a function of the survey resolution. For completeness, statistics were also computed for resistivity surveys of $2 \times 2$ and $12 \times 12$ block resolution. It has been found that $U$ decreases with decrease in resolution, in line with Figures 3.15a and 3.15b.

## 3.6 Summary

This study investigated some problems associated with combining resistivity tomography and resistivity well logging. The focus on resistivity rather than on hydrogeological properties stems from the observation that properties such as permeability and porosity can be derived from the resistivity based on theoretical or empirical,

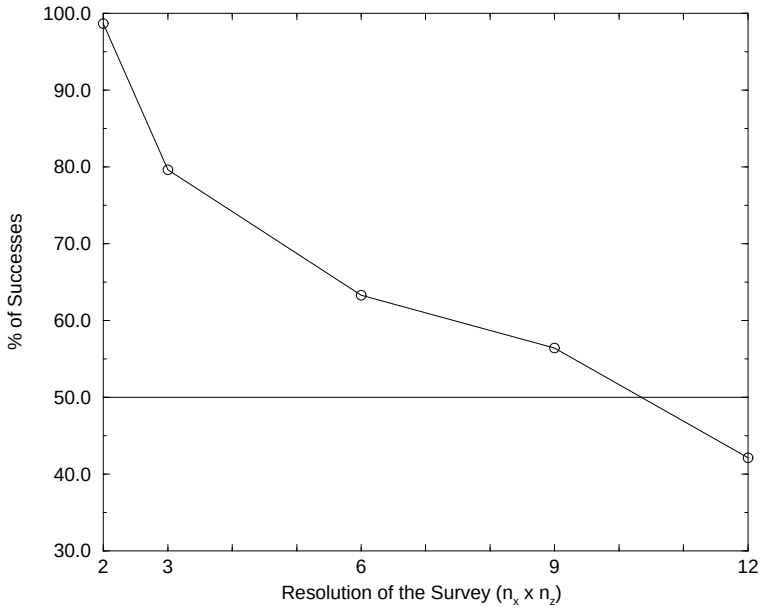


Figure 3.17: Percentage of number of successes (equation (3.17)) of the Bayesian updating approach for different survey resolutions and different errors in the surveys.

site-specific models. The study was conducted by considering conditions as realistic as possible, and the real-life data collected at the LLNL site and the synthetic data along some cross sections were analyzed.

A Bayesian model was used for data assimilation in this study, and the method allows for conditioning on what is clearly a set of complex and nonlinear petrophysical models relating between different geological attributes. This method comprises several steps, each of which intended to explore, model and utilize the aspects of the data that are needed for relating between the tomographic data and the well logs. Although such method is universal in all its components, the complexity of the geophysical surveying and interpretation makes several of its aspects site-specific.

This study employed several relationships between induction resistivity, lithofacies, shaliness and tomographic resistivity. The relationships reflect, to a large degree, properties that are well understood and quite general in terms of trends. However, these relationships may not be transported to other sites; in this sense, the method used in the study does not replace nor alleviate the tedious task of data exploration. It is helpful at the data exploration stage to identify "common-factors": the attributes or parameters that can be used for projecting areally the well log data and act as surrogates for hydrogeological and geophysical properties. At the LLNL site, that key element is the shaliness, due to its well-defined spatial structure and its sensitivity to resistivity. Once the "common factors" are identified, the Bayesian model becomes the key for data assimilation at this stage.

The Bayesian approach used in the study clearly can be changed and improved. The pdfs in this application follow Gaussian models. One may argue that pdfs should be accurate reflections of data and not models or conjectures. This issue is particularly significant for the tails of the pdfs. However, modeling decision here does not impinge on the fundamentals of the approach, which by no means require Gaussian pdfs [*Woodbury and Ulrych*, 1993]. Another issue concerns the development of the petrophysical models. The search leading to Figure 3.6a was based on visual inspection. Nevertheless, there is clearly a need to implement a more systematic approach, especially given that the relationship can be more complex in terms of the number of clusters and the number of the parameters involved.

Under the conditions explored here, a significant reduction of the estimation accuracy was observed in the presence of realistic error levels in the geophysical surveys. The benefits in estimating high-resolution subsurface resistivity are more significant than those gained in estimating lithofacies given a low-resolution resistivity survey. The LLNL data showed good correlation between resistivity and lithofacies at the small scale, but the correlation deteriorates at lower resolution. This observation is supported by the fact that the resistivity surveys were non-informative for updating the lithofacies images. Resistivity-shaliness-lithofacies relations may show perfect correlation at a fine scale but can appear to have large scatter when using a larger observation scale.

# Chapter 4

# Estimating Lithofacies from Borehole and Crosshole Geophysical Data Using the Bayesian Model Coupled with a Fuzzy Neural Network

## 4.1 Introduction

Heterogeneity of lithofacies has an important effect on the determination of hydrogeological and geochemical parameters in flow and contaminant transport. Modeling of this heterogeneity requires the joint use of different sources of information, especially less-invasive and cost-effective geophysical data.

Borehole geophysical data have been used for decades to map spatial variability of lithofacies [*Doveton*, 1986]. The main focus of the mapping is on the inference of the cross correlation between the lithofacies and the multiple geophysical attributes. Many methods have been developed for the purpose, such as graphical models [*Doveton*, 1986], multivariate analysis [*Delfiner et al.*, 1987], neural networks [*Rogers et*

*al.*, 1992], and fuzzy neural networks [*Chang et al.*, 1997]. Among those methods, fuzzy neural networks have been shown most attractive because they take advantages of neural networks in fitting nonlinear functions and fuzzy logic in including human knowledge into the fitting. The methods, however, are ignorant of spatial correlation of lithofacies, which has been shown very useful in many situations [*Deutsch and Journel*, 1998].

Surface and crosshole geophysical data have also been used to improve lithofacies estimation, such as surface or crosshole seismic data [*Lortzer and Berkhout*, 1992; *Copty and Rubin*, 1995; *Hyndman and Gorelick*, 1996], surface gravity and magnetic data [*Bosch*, 1999; *Bosch et al.*, 2001], and surface electrical resistivity data [*Salem*, 2001]. To jointly use geophysical data collected from surface or crosshole surveys and from borehole logging, several problems exist: (1) the scale disparity between surface or crosshole surveys and borehole logging [*Ezzedine et al.*, 1999], (2) spatial correlation of lithofacies, and (3) cross correlation between lithofacies and geophysical attributes. A simple yet practical method to solve the problems is geostatistical indicator cokriging [*Rosenbaum et al.*, 1997; *Deutsch and Journel*, 1998]. This method, however, is limited as the cross correlation between lithofacies and geophysical attributes is highly nonlinear.

Bayesian methods have been used for many years to incorporate geophysical data into lithofacies estimation [*Copty and Rubin*, 1995]. The methods provide a general framework for data assimilation and allow various types of information to be inte-

grated in a hierarchical manner [*Box and Tiao*, 1973; *Bernardo et al.*, 1994]. Using a carefully built Bayesian model, surface or crosshole geophysical data as well as borehole lithofacies and geophysical logs can be jointly used to estimate lithofacies, provided that the relations between the various types of information are available. The parameters in the Bayesian models can be identified using the maximum of a-posteriori probability density function (MAP) under some conditions [*Lortzer and Berkhout*, 1992], but sampling-based methods, such as the Markov Chain Monto Carlo (MCMC) method, are needed for more general situations [*Gilks et al.*, 1998; *Bosch*, 1999; *Bosch et al.*, 2001]. A key focus of the Bayesian model is on the inference of the cross correlation between lithofacies and geophysical attributes. Since the relation is often nonlinear, site-specific, and difficult to obtain, a data-driven model is developed in this study to infer the cross correlation directly from training data sets, without making any assumption about the form of the function.

This chapter develops a Bayesian model coupled with a fuzzy neural network (BFNN) for lithofacies estimation, using spatial correlation of the lithofacies as well as the nonlinear cross correlation between the lithofacies and the geophysical attributes. Section 4.2 describes the Bayesian model, and section 4.3 and 4.4 present two case studies using the model. Discussion is given in sections 4.5.

## 4.2 Bayesian Model Coupled with a Fuzzy Neural Network

This section describes the Bayesian model coupled with a fuzzy neural network for lithofacies estimation. A general Bayesian framework for combining borehole and crosshole geophysical data is given first, and then indicator kriging for estimating the prior and the fuzzy neural network for inferring the likelihood function are introduced. The focus is on the inference of the likelihood function using the fuzzy neural network.

### 4.2.1 Bayesian Framework

The developed Bayesian model is based on a typical situation of subsurface characterization, such as at the Lawrence Livemore National Laboratory (LLNL) site [*Ezzedine et al.*, 1999] and at the Oyster site [*Chen et al.*, 2001]. Suppose lithofacies at any location $\mathbf{x}$ needs to be estimated, given lithofacies measurements at location $\mathbf{x}_i$, $i \in A = \{1, 2, \cdots, n\}$, and collocated geophysical data $g_1(\mathbf{x}), g_2(\mathbf{x}), \cdots, g_t(\mathbf{x})$, where $t$ is the total number of geophysical attributes. The geophysical data can be obtained from crosshole tomography surveys or estimated from intensive borehole geophysical measurements using kriging. Let $Z(\mathbf{x})$ be a discrete random variable taking number $1, 2, \cdots$, or, $q$, where $q$ is the total number of possible lithofacies at a site. Let $z(\mathbf{x}_i)$ be the lithofacies measurement, a number between 1 and $q$, at location $\mathbf{x}_i$, $i \in A$. The conditional probability of the $k$th lithofacies occurring at location $\mathbf{x}$ can

be determined by the following Bayesian formula [*Bernardo and Smith*, 1994],

$$\begin{aligned} P(Z(\mathbf{x}) = k | g_1(\mathbf{x}), g_2(\mathbf{x}), \cdots, g_t(\mathbf{x}), z(\mathbf{x}_i), i \in A) = \\ Cf(g_1(\mathbf{x}), g_2(\mathbf{x}), \cdots, g_t(\mathbf{x}) | Z(\mathbf{x}) = k, z(\mathbf{x}_i), i \in A) \cdot \\ P(Z(\mathbf{x}) = k | z(\mathbf{x}_i), i \in A), \end{aligned} \tag{4.1}$$

where $C$ is a normalizing constant and $f$ is a joint conditional density function, referred to as a likelihood function. The most likely estimate of lithofacies at location $\mathbf{x}$ is the one that maximizes the conditional probability.

The previous Bayesian model can be simplified under certain conditions. Consider the Markov assumption that cross correlation between lithofacies and geophysical attributes at location $\mathbf{x}$ does not depend on lithofacies measurements at other locations, $z(\mathbf{x}_i)$, $i \in A$, can be dropped from the likelihood function. Consequently, the Bayesian formula becomes

$$P_{post}(Z(\mathbf{x}) = k) = [Cf(g_1(\mathbf{x}), g_2(\mathbf{x}), \cdots, g_t(\mathbf{x}) | Z(\mathbf{x}) = k)] \cdot P_{prior}(Z(\mathbf{x}) = k), \tag{4.2}$$

where $P_{post}(Z(\mathbf{x}) = k) = P(Z(\mathbf{x}) = k | g_1(\mathbf{x}), g_2(\mathbf{x}), \cdots, g_t(\mathbf{x}), z(\mathbf{x}_i), i \in A)$ referred to as posterior probability, and $P_{prior}(Z(\mathbf{x}) = k) = P(Z(\mathbf{x}) = k | z(\mathbf{x}_i), i \in A)$ referred to as prior probability.

### 4.2.2 Prior Estimate

The prior probability is estimated only from lithofacies measurements $z(\mathbf{x}_i), i \in A$, using indicator kriging [*Rosenbaum et al.*, 1997; *Deutsch and Journel*, 1998]. Let $I_k(\mathbf{x})$

be an indicator random variable defined by

$$I_k(\mathbf{x}) = \begin{cases} 1 & \text{if } Z(\mathbf{x}) = k \\ 0 & \text{otherwise} \end{cases}, \tag{4.3}$$

and therefore $\{Z(\mathbf{x}) = k\}$ is equivalent to $\{I_k(\mathbf{x}) = 1\}$. Let $p_k$ be the unconditional probability of the $k$th lithofacies occurring at location $\mathbf{x}$. The conditional probability is thus given by

$$P_{prior}(Z(\mathbf{x}) = k) = p_k + \sum_{i \in A} \lambda_i(\mathbf{x})(I_k(\mathbf{x}_i) - p_k) \tag{4.4}$$

and

$$\sum_{i \in A} \lambda_i(\mathbf{x}) C_I(\mathbf{x}_i, \mathbf{x}_j) \quad = \quad C_I(\mathbf{x}, \mathbf{x}_j), \qquad j \in A,$$

where $C_I(\mathbf{x}_i, \mathbf{x}_j)$ is the covariance of the indicator variables at location $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\lambda_i(\mathbf{x})$, $i \in A$, are the unknown parameters.

The indicator kriging takes advantages of spatial correlation of lithofacies. As the distances between location $\mathbf{x}$ and $\mathbf{x}_i, i \in A$, become large, however, the conditional probability will be close to the unconditional probability $p_k$, and lithofacies measurements at boreholes will not influence lithofacies estimation at location $\mathbf{x}$.

### 4.2.3 Likelihood Function

The likelihood function in Equation 4.2 is a connection between lithofacies and geophysical attributes, and it can be inferred from both lithofacies and geophysical

data. The inference, however, is very challenging because cross correlation between the lithofacies and geophysical attributes is often nonlinear and site-specific. Many previous applications of Bayesian methods follow a forward approach [*Kitanidis*, 1986; *Copty and Rubin*, 1995]. They first assumed the form of the likelihood function and then estimated the parameters associated with the form. This approach is straightforward yet efficient in certain circumstances, but giving the form of the likelihood function is very difficult, especially as data quality is low and the cross correlation is complicated. This study follows another line and uses a fuzzy neural network as a computing model to learn the likelihood function directly from a training data set.

The paradigm of the computing model is shown in Figure 4.1, where $h(\mathbf{x}, \boldsymbol{\theta})$ is the function determined by the neural network structure and $\boldsymbol{\theta}$ is a parameter vector associated with the structure. Let $(g_1(\mathbf{x}_i), g_2(\mathbf{x}_i), \cdots, g_t(\mathbf{x}_i), Cf(g_1(\mathbf{x}_i), g_2(\mathbf{x}_i), \cdots, g_t(\mathbf{x}_i) \mid Z(\mathbf{x}_i) = k))$, $i \in A$, be a training data set. Parameter vector $\boldsymbol{\theta}$ can be estimated by minimizing the following objective function:

$$\sum_{i \in A} [h(\mathbf{x}_i, \boldsymbol{\theta}) - log(Cf(g_1(\mathbf{x}_i), g_2(\mathbf{x}_i), \cdots, g_t(\mathbf{x}_i) | Z(\mathbf{x}_i) = k))]^2. \tag{4.5}$$

Fitting $log(Cf)$ instead of $Cf$ is because the log value of $Cf$ has a smaller range that leads to a more efficient computation.

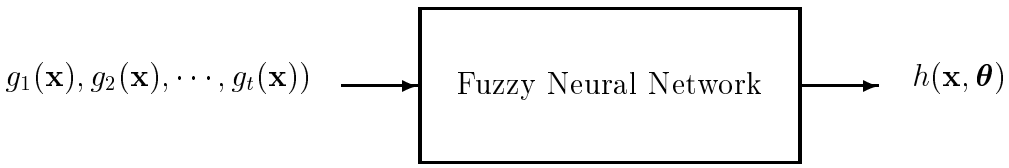


Figure 4.1: Computing model of the likelihood function

### 4.2.4 Structure of the Fuzzy Neural Network

The fuzzy neural network used in the study is similar to the one given by *Takagi and Sugeno* [1985] and *Jang* [1993], which consists of several fuzzy inference rules. To simplify illustration of the system, only two types of geophysical attributes are considered; the model, however, can be directly applied to a more general case. Let $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ be geophysical data at location $\mathbf{x}$. An example of the system with two fuzzy rules is given by

**Rule-1:** If $g_1(\mathbf{x})$ is $A_1$ and $g_2(\mathbf{x})$ is $B_1$, then the output of the fuzzy rule is $o_1$,

**Rule-2:** If $g_1(\mathbf{x})$ is $A_2$ and $g_2(\mathbf{x})$ is $B_2$, then the output of the fuzzy rule is $o_2$.

$A_i$ and $B_i$ $(i = 1, 2)$ in the premises of those rules are linguistic labels, such as *small* and *large*, or fuzzy sets. A fuzzy set is a class of objects with a continuum of grades of membership; such a set is characterized by a membership function that assigns to each object a grade of membership ranging between zero and one [*Zadeh.* 1965]. Fuzzy sets and their linguistic labels are often used interchangeably in literatures, and

this allows interpreting fuzzy inference rules intuitively. $o_1$ and $o_2$ in the consequences of the rules are constants in the study, but they can be linear functions of input $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$, as given by *Takagi and Sugeno* [1985].

Figure 4.2 shows the fuzzy reasoning of the fuzzy neural network. For given geophysical data $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ at location $\mathbf{x}$, the grades of membership of $g_1(\mathbf{x})$ in fuzzy sets $A_1$ and $A_2$, i.e. $\mu_{A_1}[g_1(\mathbf{x})]$ and $\mu_{A_2}[g_1(\mathbf{x})]$, and the grades of membership of $g_2(\mathbf{x})$ in fuzzy sets $B_1$ and $B_2$, i.e. $\mu_{B_1}[g_2(\mathbf{x})]$ and $\mu_{B_2}[g_2(\mathbf{x})]$ are calculated. Using these membership values, the weights of rule-1 and rule-2 for the inference can be determined by following

$$w_i(\mathbf{x}) = \mu_{A_i}[g_1(\mathbf{x})] \cdot \mu_{B_i}[g_2(\mathbf{x})], \quad i = 1, 2. \tag{4.6}$$

Other methods can also be used to compute the weights, such as *minimum* or *maximum*, but the current method is most computationally efficient [*Takagi and Sugeno*, 1985]. The final result of this inference is a weighted linear combination of the outputs of all the fuzzy rules as given by

$$h(\mathbf{x}, \boldsymbol{\theta}) = \frac{w_1(\mathbf{x})o_1 + w_2(\mathbf{x})o_2}{w_1(\mathbf{x}) + w_2(\mathbf{x})}, \tag{4.7}$$

where $\boldsymbol{\theta}$ is a parameter vector and $h(\mathbf{x}, \boldsymbol{\theta})$ is a function of both location $\mathbf{x}$ and parameter $\boldsymbol{\theta}$.

Figure 4.3 shows the structure of the fuzzy neural network as given by [*Jang*, 1993]. The input to each node in layer 1 is $g_1(\mathbf{x})$ or $g_2(\mathbf{x})$; the output is the corresponding

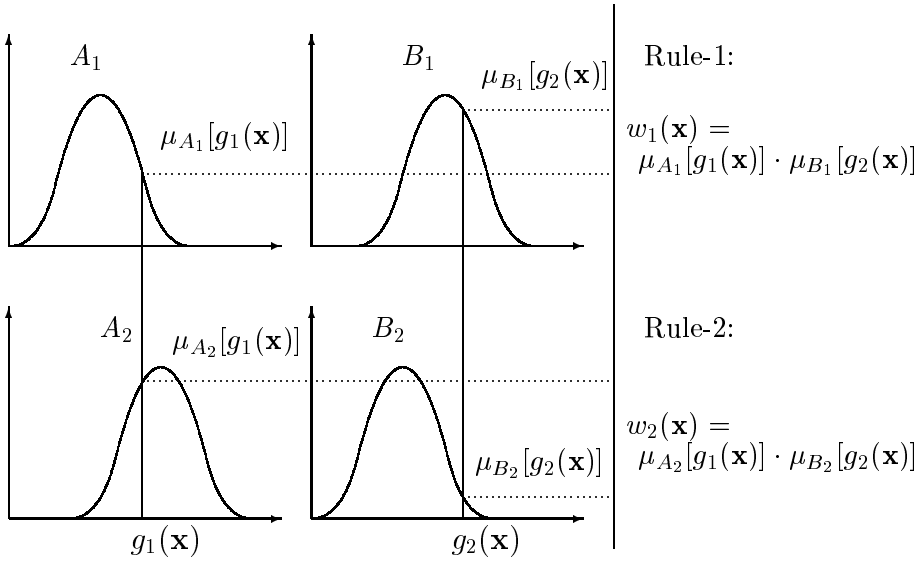


Figure 4.2: Fuzzy reasoning of the fuzzy neural network

grade of membership $\mu_{A_i}[g_1(\mathbf{x})]$ or $\mu_{B_i}[g_2(\mathbf{x})]$ as given by

$$\begin{aligned} \mu_{A_i}[g_1(\mathbf{x})] &= \exp\left(-\left(\frac{g_1(\mathbf{x}) - c_{i1}}{\sigma_{i1}}\right)^2\right), \\ \mu_{B_i}[g_2(\mathbf{x})] &= \exp\left(-\left(\frac{g_2(\mathbf{x}) - c_{i2}}{\sigma_{i2}}\right)^2\right), \end{aligned} \tag{4.8}$$

where $i = 1, 2$, $c_{i1}$ and $c_{i2}$ are the centers of fuzzy sets $A_i$ and $B_i$, and $\sigma_{i1}$ and $\sigma_{i2}$ are their corresponding bases. The outputs of the node in layer 2 and 3 are the weights $w_i(\mathbf{x})$, $i = 1, 2$, and their relative weights $\bar{w}_i(\mathbf{x}) = w_i(\mathbf{x})/(w_1(\mathbf{x}) + w_2(\mathbf{x}))$, $i = 1, 2$, respectively. The input to each node in layer 4 is the relative weight $\bar{w}_i(\mathbf{x})$ and the consequences $o_i$, $i = 1, 2$; the output is the product of its corresponding input. The node in the last hidden layer is the summation of all incoming signals to the node as

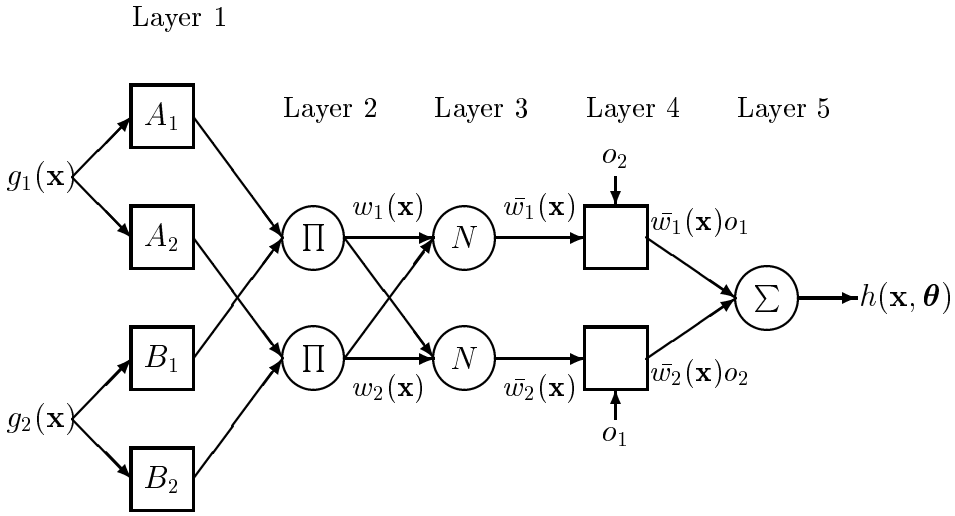


Figure 4.3: Structure of the fuzzy neural network

given by

$$h(\mathbf{x}, \boldsymbol{\theta}) = \bar{w}_1(\mathbf{x})o_1 + \bar{w}_2(\mathbf{x})o_2, \tag{4.9}$$

which is equivalent to Equation 4.7.

This system seems very simple and arbitrary, but it has been demonstrated to be very efficient in fitting multivariate nonlinear functions by *Takagi and Sugeno* [1985], *Horikawa et al* [1991], and *Jang* [1993]. Actually, this system can be used to approximate any continuous nonlinear function defined on a bounded domain as shown by *Jang* [1993] and *Rojas* [1996].

### 4.2.5 Learning Algorithm of the Fuzzy Neural Network

The learning of the fuzzy neural network includes the determination of the number of fuzzy inference rules and the identification of the parameters associated with the rules from a given training data set.

#### Structure Identification

Structure identification is performed using fuzzy c-mean (FCM) cluster analysis. The number of fuzzy rules and the initial values of the centers and bases of each fuzzy set in the premises of those rules are determined by identifying possible patterns inherent in the input data.

The algorithm used in the study was given by *Bezdek* [1981], which is a very efficient iterative procedure. Let $(g_1(\mathbf{x}_j), g_2(\mathbf{x}_j))$, $j \in A$, be input data, which can be divided into $m$ fuzzy clusters. Let $\mathbf{c}_i = (\mathbf{c}_{i1}, \mathbf{c}_{i2})^T$, $i = 1, 2, \cdots, m$, be the centers of the fuzzy sets, $\mathbf{y}_j = (g_1(\mathbf{x}_j), g_2(\mathbf{x}_j))^T$ be the $j$th input data, and $u_{ij}$ be the grade of membership of $\mathbf{y}_j$ in the $i$th fuzzy cluster. The algorithm becomes:

1. Randomly assign the initial grades of membership $u_{ij} \in [0, 1]$ and $\sum_{i=1}^{m} u_{ij} = 1$.

2. Compute the center of each fuzzy cluster using

$$\mathbf{c}_i = \frac{\sum_{j \in A} u_{ij}^2 \mathbf{y}_j}{\sum_{j \in A} u_{ij}^2}, \qquad i = 1, 2, \cdots, m. \tag{4.10}$$

3. Update the grades of membership using the newly obtained centers $\mathbf{c}_i$, $i = 1, 2, \cdots, m$. Let $d_{ij} = \| \mathbf{c}_i - \mathbf{y}_j \|$ and $d_{kj} = \| \mathbf{c}_k - \mathbf{y}_j \|$. If $d_{ij} = 0$, $u_{ij} = 1$ and

$u_{kj} = 0$ for $k \neq i$; otherwise,

$$u_{ij} = \frac{1/d_{ij}^2}{\sum_{k=1}^{m} 1/d_{kj}^2}. \tag{4.11}$$

4. Compare the newly updated grades of membership with the old ones. If they are close to each other, stop the iteration; otherwise, go back to step 2.

The number of fuzzy rules is determined by considering both compactness and separation of fuzzy clusters [*Xie and Beni*, 1991]. The compactness of the clusters is defined as the weighted squared distance as given by

$$J(\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_m) = \sum_{i=1}^{m} \sum_{j \in A} u_{ij}^2 \parallel \mathbf{c}_i - \mathbf{y}_j \parallel^2. \tag{4.12}$$

The separation of the clusters is defined as the minimum distance between the centers of each pair of the clusters as follows

$$d_{min} = \min_{i \neq j} \parallel \mathbf{c}_i - \mathbf{c}_j \parallel, \qquad i, j = 1, 2, \cdots, m. \tag{4.13}$$

Both the compactness and the separation decrease with the increasing of the number of clusters. For a good cluster analysis, data points within clusters are expected to be compact (small compactness), whereas the cluster centers are expected to be well separate (large separation). Consequently, the number of clusters can be determined by minimizing the following function:

$$S = \frac{J(\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_m)/n}{d_{min}^2}, \tag{4.14}$$

where $n$ is the total number of pairs in the input data.

**Parameter Identification**

Parameter identification plays an important role in the learning of the fuzzy neural network. The consequence of each rule and the centers and bases of the fuzzy sets in the premises of the rules, specifically $o_i$, $c_{i1}$, $c_{i2}$, $\sigma_{i1}$ and $\sigma_{i2}$ for $i = 1, 2$, are estimated from a giving training data set.

The training data set used in the study originally is in the form of $(g_1(\mathbf{x}_i), g_2(\mathbf{x}_i), z(\mathbf{x}_i))$, $i \in A$, where $g_1(\mathbf{x}_i)$, $g_2(\mathbf{x}_i)$ and $z(\mathbf{x}_i)$ are geophysical data and lithofacies at location $\mathbf{x}_i$ and $A$ is an index set as given early in the chapter. Since the fuzzy neural network is used to approximate log likelihood rather than lithofacies in this study, transformation from lithofacies measurement $z(\mathbf{x}_i)$ into the corresponding log likelihood is needed. Let $l(\mathbf{x}_i) = log[Cf(g_1(\mathbf{x}_i), g_2(\mathbf{x}_i)|Z(\mathbf{x}_i) = k)]$. Using Equation 4.2, the log likelihood is given by

$$l(\mathbf{x}_i) = \log[P_{post}(Z(\mathbf{x}_i) = k)] - \log[P_{prior}(Z(\mathbf{x}_i) = k)]. \tag{4.15}$$

The prior probability $P_{prior}(Z(\mathbf{x}_i) = k)$ is estimated from lithofacies measurements at other wells. To avoid the prior probability being zero, let $P_{prior}(Z(\mathbf{x}_i) = k) = \epsilon$, where $\epsilon$ is a small value between 0.01 and 0.05, if $P_{prior}(Z(\mathbf{x}_i) = k) < \epsilon$. Likewise, assign $P_{post}(Z(\mathbf{x}_i) = k) = 1 - \epsilon$ if $z(\mathbf{x}_i) = k$, and $P_{post}(Z(\mathbf{x}_i) = k) = \epsilon$ otherwise. The value of $\epsilon$ seems to be arbitrary, but the estimating results are not sensitive to the choice of $\epsilon$.

Several methods have been used to identify those parameters in the fuzzy neural

network. *Horikawa et al.* [1991] considered the system as a general neural network and used the back-propagation method. In light of the fact $h(\mathbf{x}, \boldsymbol{\theta})$ in Equation 4.7 is a linear function of $o_1$ and $o_2$ if other parameters are fixed, *Takagi and Sugeno* [1985], *Jang* [1991] and *Nikraviesh* [1998] employed hybrid learning methods. They used the least squares method to estimate $o_1$ and $o_2$ and other nonlinear optimization methods to estimate $c_{i1}, c_{i2}, \sigma_{i1}$ and $\sigma_{i2}$ for $i = 1, 2$.

The learning algorithm is an iterative process as shown in Figure 4.4. Let $\mathbf{o} = (o_1, o_2)^T$, $\boldsymbol{\alpha} = (c_{11}, c_{12}, c_{21}, c_{22}, \sigma_{11}, \sigma_{12}, \sigma_{21}, \sigma_{22})^T$, and $RSS$ be the residual sum of squares as given below:

$$RSS(\mathbf{o}, \boldsymbol{\alpha}) = \sum_{i \in A} [h(\mathbf{x}_i, \mathbf{o}, \boldsymbol{\alpha}) - l(\mathbf{x}_i)]^2. \tag{4.16}$$

Firstly, parameter $\mathbf{o}^{(1)}$, where the superscript denotes the number of iteration, is estimated using the least squares method, given the initial value $\boldsymbol{\alpha}^{(0)}$ obtained from previous structure identification. Secondly, parameter $\boldsymbol{\alpha}^{(1)}$ is estimated using a nonlinear optimization method, given the newly updated $\mathbf{o}^{(1)}$. Finally, the most recently obtained residual sum of squares ($RSS$) is compared to the previous one. If they are very close, the iteration is terminated; otherwise, it is not, as shown in Figure 4.4.

The least squares method used in the hybrid learning is straightforward. Given parameter $\boldsymbol{\alpha}$, relative weights $\bar{w_1}(\mathbf{x}_i)$ and $\bar{w_2}(\mathbf{x}_i)$ for each input $(g_1(\mathbf{x}_i), g_2(\mathbf{x}_i))$, $i \in A$,

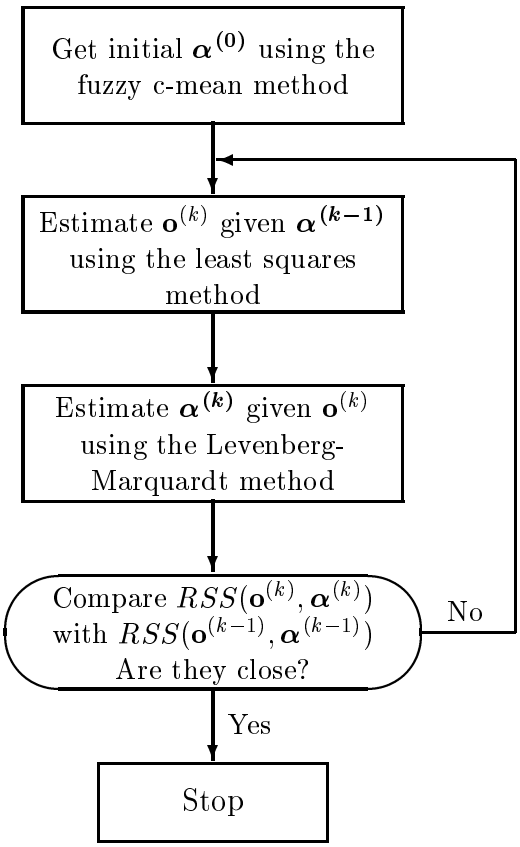


Figure 4.4: Learning procedure of the fuzzy neural network

can be calculated to get a matrix $\mathbf{w}$ as follows,

$$\begin{bmatrix} \bar{w_1}(\mathbf{x}_1) & \bar{w_2}(\mathbf{x}_1) \\ \bar{w_1}(\mathbf{x}_2) & \bar{w_2}(\mathbf{x}_2) \\ \vdots & \vdots \\ \bar{w_1}(\mathbf{x}_n) & \bar{w_2}(\mathbf{x}_n) \end{bmatrix}. \tag{4.17}$$

Let $\mathbf{l} = (l(\mathbf{x}_1), l(\mathbf{x}_2), \cdots, l(\mathbf{x}_n))^T$. The relation between the consequence of each fuzzy rule and the output of the system is given by $\mathbf{wo} = \mathbf{l}$, and the least squares estimate of parameter $\mathbf{o}$ is given by $(\mathbf{w}^T\mathbf{w})^{-1}\mathbf{w}^T\mathbf{l}$ [*Stone*, 1995].

The Levenberg-Marquardt algorithm as a nonlinear optimization method is used to estimate parameter $\boldsymbol{\alpha}$ in the study for a given parameter $\mathbf{o}$. It is a revised Gaussian-Newton method and requires iteratively evaluating residuals $\mathbf{e} = (e_1, e_2, \cdots, e_n)^T$, where $e_i = h(\mathbf{x}_i, \mathbf{o}, \boldsymbol{\alpha}) - l(\mathbf{x}_i)$, $i \in A$, and Jacobian matrix $\mathbf{J}$ given by [*McKeown*, 1980]

$$\begin{bmatrix} \frac{\partial e_1}{\partial c_{11}} & \frac{\partial e_1}{\partial c_{12}} & \frac{\partial e_1}{\partial c_{21}} & \frac{\partial e_1}{\partial c_{22}} & \frac{\partial e_1}{\partial \sigma_{11}} & \frac{\partial e_1}{\partial \sigma_{12}} & \frac{\partial e_1}{\partial \sigma_{21}} & \frac{\partial e_1}{\partial \sigma_{22}} \\ \frac{\partial e_2}{\partial c_{11}} & \frac{\partial e_2}{\partial c_{12}} & \frac{\partial e_2}{\partial c_{21}} & \frac{\partial e_2}{\partial c_{22}} & \frac{\partial e_2}{\partial \sigma_{11}} & \frac{\partial e_2}{\partial \sigma_{12}} & \frac{\partial e_2}{\partial \sigma_{21}} & \frac{\partial e_2}{\partial \sigma_{22}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial e_n}{\partial c_{11}} & \frac{\partial e_n}{\partial c_{12}} & \frac{\partial e_n}{\partial c_{21}} & \frac{\partial e_n}{\partial c_{22}} & \frac{\partial e_n}{\partial \sigma_{11}} & \frac{\partial e_n}{\partial \sigma_{12}} & \frac{\partial e_n}{\partial \sigma_{21}} & \frac{\partial e_n}{\partial \sigma_{22}} \end{bmatrix},$$

where

$$\begin{aligned} \frac{\partial e_i}{\partial c_{ks}} &= 2\bar{w}_k(\mathbf{x}_i)[o_k - h(\mathbf{x}_i, \mathbf{o}, \boldsymbol{\alpha})] \cdot \frac{[g_s(\mathbf{x}_i) - c_{ks}]}{\sigma_{ks}^2}, \qquad k, s = 1, 2, \\ \frac{\partial e_i}{\partial \sigma_{ks}} &= 2\bar{w}_k(\mathbf{x}_i)[o_k - h(\mathbf{x}_i, \mathbf{o}, \boldsymbol{\alpha})] \cdot \frac{[g_s(\mathbf{x}_i) - c_{ks}]^2}{\sigma_{ks}^3}, \qquad k, s = 1, 2, \end{aligned}$$

which are obtained using the chain rule of differentiation and the membership functions given in Equation 4.8. The detailed algorithm is given below

1. Compute residual $\mathbf{e}$ and Jacobian matrix $\mathbf{J}$ for given parameters $\boldsymbol{\alpha}^{(0)}$ and set $k = 0$.

2. Check gradient $\mathbf{J}^T\mathbf{e}$. If $|\mathbf{J}^T\mathbf{e}| < \epsilon$, stop; otherwise, go to next step.

3. For a given positive value $\mu$, compute $\boldsymbol{\alpha}^{(k)} = \boldsymbol{\alpha}^{(k-1)} - (\mathbf{J}^T\mathbf{J} + \mu I)^{-1}\mathbf{J}^T\mathbf{e}$ and check whether $RSS(\mathbf{o}^{(k)}, \boldsymbol{\alpha}^{(k)}) < RSS(\mathbf{o}^{(k)}, \boldsymbol{\alpha}^{(k-1)})$. If it does, go to next stop; otherwise increase $\mu$ value and repeat step 3 until $RSS$ is reduced.

4. Check $|\boldsymbol{\alpha}^{(k)} - \boldsymbol{\alpha}^{(k-1)}| < \epsilon$. If it does, stop; otherwise, set k=k+1 and repeat step 1.

## 4.3 Case Study 1

This case study demonstrates the ability of the Bayesian model coupled with a fuzzy neural network to combine borehole and crosshole data for lithofacies estimation, using a synthetic data set generated from field measurements at the Lawrence Livemore National Laboratory (LLNL) site.

Table 4.1: Spatial structures of lithofacies and geophysical data

| | Lithofacies | | Gamma-ray (Shaliness) | Resistivity ($\Omega m$) |
|---|---|---|---|---|
| | Sand | Silt | | |
| Proportion | 0.52 | 0.48 | N/A | N/A |
| Horizontal Range(m) | 30.0 | 30.0 | 25.0 | $+\infty$ |
| Vertical Range(m) | 1.50 | 1.50 | 2.50 | $+\infty$ |
| Nugget | 0.0 | 0.0 | 0.011 | $\sigma_R^2$ |
| Sill | 0.25 | 0.25 | 0.040 | $\sigma_R^2$ |
| Models | Exponential | Exponential | Gaussian | N/A |

### 4.3.1 Synthetic Data

This study focuses on a small portion of the LLNL site shown in Figure 4.5. Among many geophysical logs collected at the site, gamma-ray and electrical resistivity logs have been found to be most informative for lithofacies identification. Table 4.1 summarizes the spatial structures of lithofacies, gamma-ray and resistivity, inferred from borehole measurements within hydrostratigraphic unit 2 (HSU2). The gamma-ray was converted into shaliness to remove inconsistencies associated with data acquisition [*Doveton*, 1986] using the method described by *Ezzedine et al.* [1999]. Both lithofacies and gamma-ray shaliness have large spatial correlations compared to resistivity, which is spatially uncorrelated along both vertical and horizontal directions.

Synthetic data are generated along the profile from w1205 to w1251, shown as a solid line in Figure 4.5, using the parameters listed in Table 4.1. Firstly, a two-dimensional lithofacies field is generated, which are conditioned to lithofacies measure-

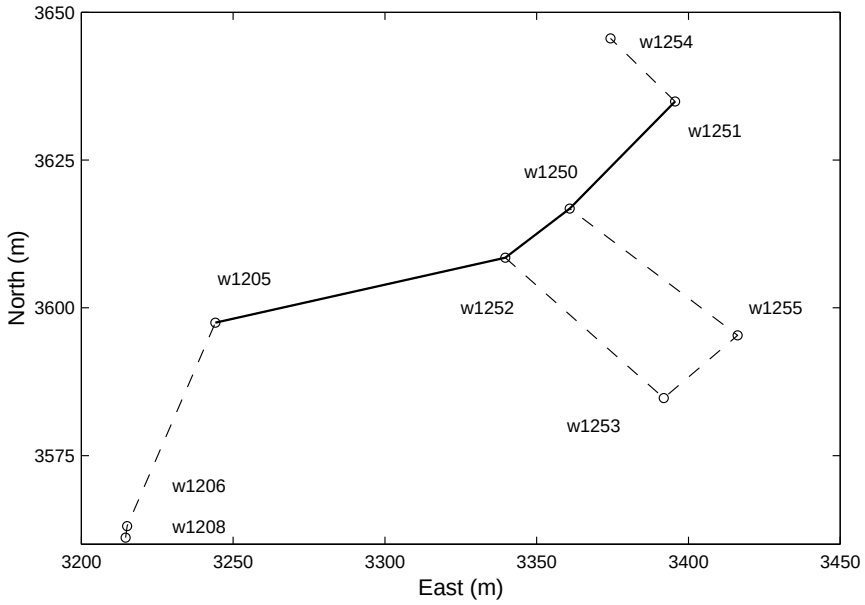


Figure 4.5: Schematic map of the LLNL site. The circles denote wellbores, and the solid lines denote the profile along which the synthetic data are generated.

ments at the wellbores, using the sequential indicator simulation method. Secondly, a two-dimensional gamma-ray shaliness field is generated, which are conditioned to the previously generated collocated lithofacies as well as gamma-ray shaliness at the wellbores, using the sequential Gaussian simulation method [*Deutsch and Journel*, 1998]. Finally, a two-dimensional resistivity field along the same profile is generated, which are only conditioned to collocated lithofacies and gamma-ray shaliness, not resistivity at the wellbores.

The previously generated data are divided into two parts, one for training and the other for testing. The training set is constructed by randomly selecting eight wellbores along the lateral direction as shown in Figure 4.6, which mimic boreholes

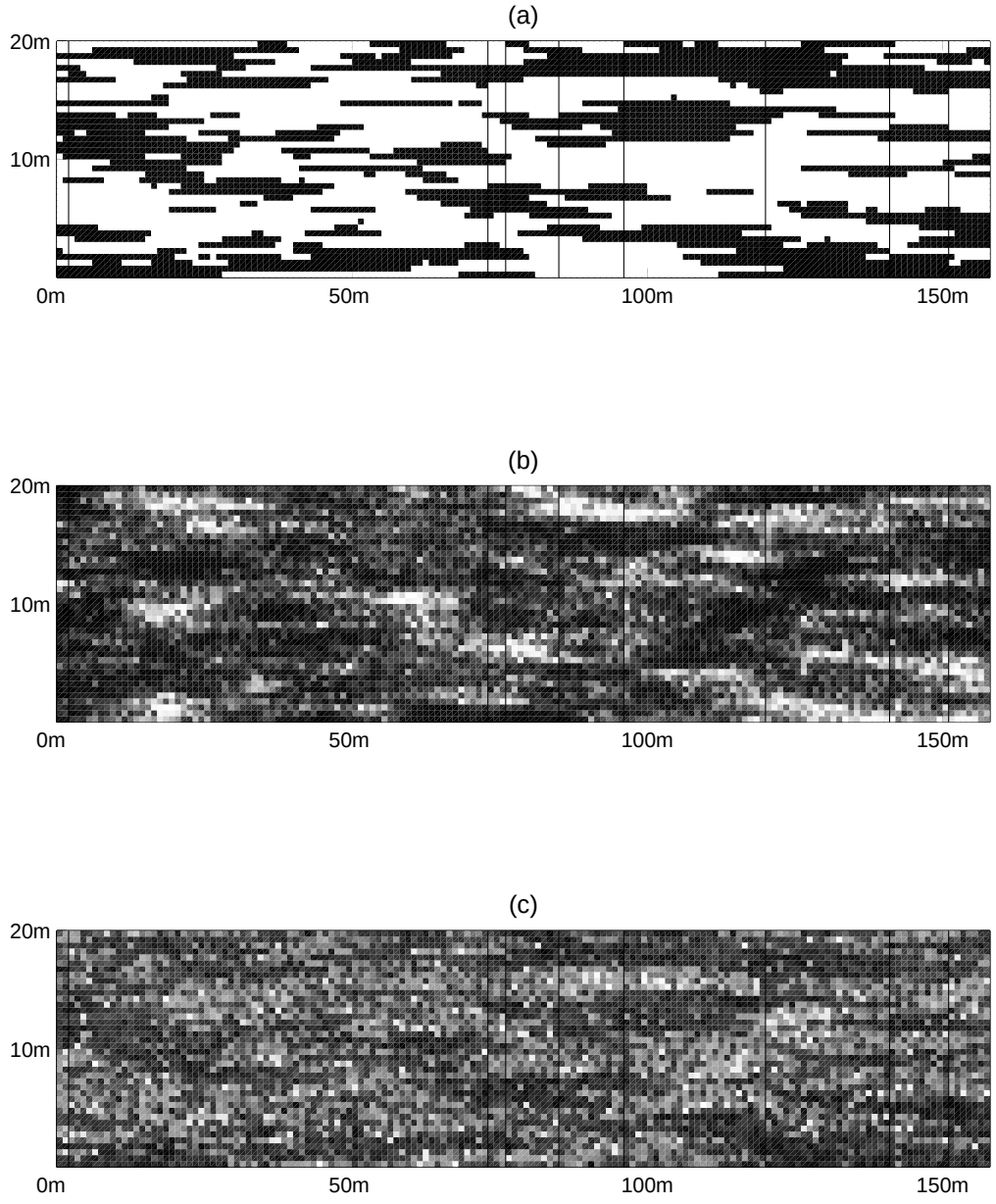


Figure 4.6: (a) Lithofacies field with silt (black) and sand (white); (b) Gamma-ray shaliness field with values between 0 (black) and 1 (white); (c) Electrical resistivity field with values between $5\Omega m$ (black) and $30\Omega m$ (white).

in real situations, and the testing set consists of the remaining of the synthetic data. Figure 4.7 compares the normalized indicator experimental variograms, shown as the dashed lines with circles and inferred from the data at the eight wellbores, with their corresponding theoretical ones, shown as the solid lines and used for generating the synthetic lithofacies field. The consistency between the experimental and theoretical variograms confirms that the generated lithofacies field is a realization of the random field with the spatial structure given in Table 4.1. Figure 4.8 shows the scatter-plot of gamma-ray shaliness versus resistivity using the data at the eight wellbores. It is clear that the cross correlation between gamma-ray shaliness and resistivity is nonlinear, non-unique and of a considerable uncertainty.

### 4.3.2 Approach

The advantages of using the BFNN model for lithofacies estimation are demonstrated by comparing it with several other models commonly used in site characterization. Each model is first trained using the data at the eight wellbores and then the trained model is used to estimate lithofacies at any location in crosshole areas. The estimated results are compared with their corresponding true values and the total number of misclassification is counted in terms of the minimum distances between the testing locations and the locations whose measurements have been used for estimating the current lithofacies. From the relations between the percentages of misclassification and the minimum distances, the efficiency of each model and the

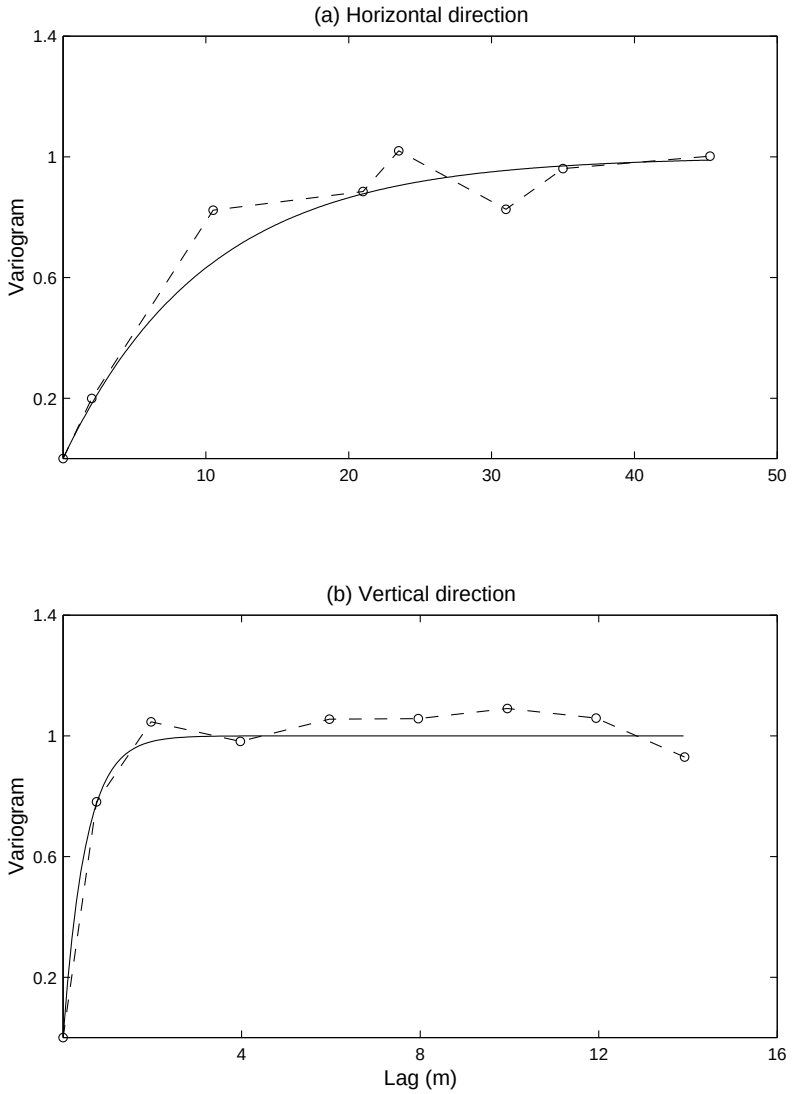


Figure 4.7: Variograms (a) along the horizontal direction and (b) along the vertical direction. The dashed lines with circles are the experimental variograms based on the data at the eight wellbores, and the solid curves are the theoretical ones used for generating the synthetic data.

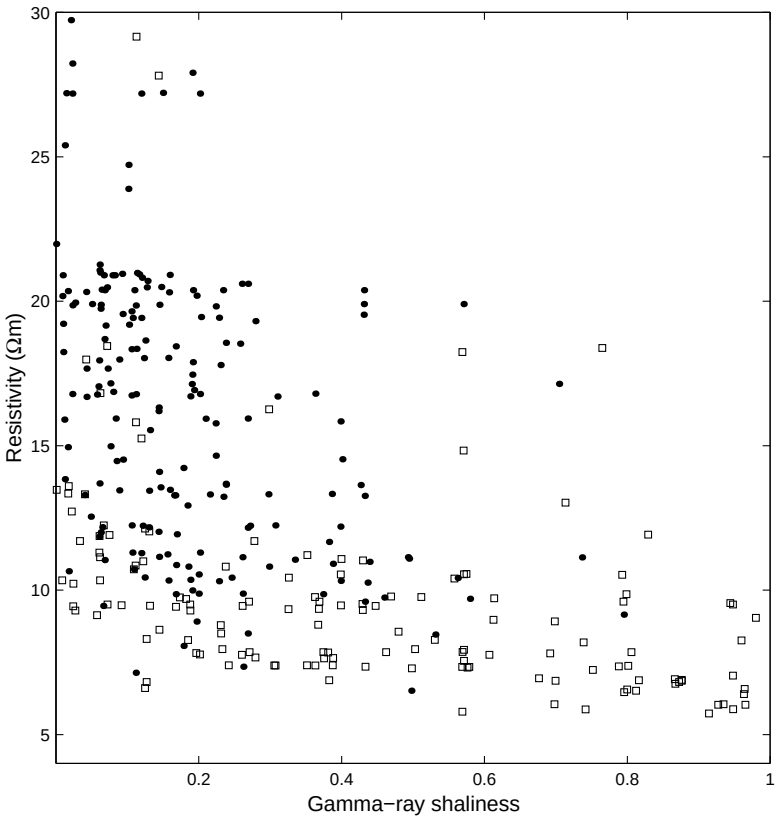


Figure 4.8: Scatter-plot of gamma-ray shaliness versus resistivity based on the data at the eight wellbores. The solid dots denote sand and the squares denote silt.

effects of the spatial correlation can be evaluated.

The models used for comparison include indicator kriging, indicator cokriging, and the fuzzy neural network without considering spatial correlation of lithofacies, which is referred to as FNN later in the chapter. The indicator kriging has been described in section 4.2.2, and the fuzzy neural network without considering spatial correlation is similar to the one given in the previous section, except that the output of the fuzzy neural network is the conditional probability rather than the log likelihood.

The indicator cokriging is similar to the one given by *Almeida and Journel* [1994], where only collocated geophysical data are used. Prior to using the model, resistivity data need to be normalized to a range between 0 and 1 by first subtracting their minimum value and then dividing by their range. Let $g_1(\mathbf{x})$ be the gamma-ray shaliness and $g_2(\mathbf{x})$ be the normalized resistivity at location $\mathbf{x}$. The conditional probability of the $k$th lithofacies occurring at location $\mathbf{x}$ is given by

$$\begin{aligned} P(Z(\mathbf{x}) = k | g_1(\mathbf{x}), g_2(\mathbf{x})) = p_k + \sum_{i \in A} \lambda_i(\mathbf{x})(I_k(\mathbf{x}_i) - p_k) + \\ + s_1(\mathbf{x})(g_1(\mathbf{x}) - m_1) + s_2(\mathbf{x})(g_2(\mathbf{x}) - m_2), \end{aligned} \tag{4.18}$$

and

$$\begin{aligned} \sum_{i \in A} \lambda_i(\mathbf{x}) \rho_I(\mathbf{x}_i, \mathbf{x}_j) + s_1(\mathbf{x}) \rho_{Ig1}(\mathbf{x}, \mathbf{x}_j) + s_2(\mathbf{x}) \rho_{Ig2}(\mathbf{x}, \mathbf{x}_j) &= \rho_I(\mathbf{x}, \mathbf{x}_j), \quad j \in A, \\ \sum_{i \in A} \lambda_i(\mathbf{x}) \rho_{Ig1}(\mathbf{x}, \mathbf{x}_i) + s_1(\mathbf{x}) \rho_{g1} + s_2(\mathbf{x}) \rho_{g1g2} &= \rho_{Ig1}, \\ \sum_{i \in A} \lambda_i(\mathbf{x}) \rho_{Ig2}(\mathbf{x}, \mathbf{x}_i) + s_1(\mathbf{x}) \rho_{g1g2} + s_2(\mathbf{x}) \rho_{g2} &= \rho_{Ig2}, \end{aligned}$$

where $m_1$ and $m_2$ are the means of the gamma-ray shaliness and the normalized resistivity. $\rho_I$, $\rho_{g1}$ and $\rho_{g2}$ in the equations are the correlation coefficients of the indicator variable, the gamma-ray shaliness and the resistivity, respectively; $\rho_{Ig1}$, $\rho_{Ig2}$, $\rho_{g1g2}$ are the cross correlation coefficients between the indicator variable and the geophysical data, $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$.

### 4.3.3 Results

Figure 4.9 shows the percentage changes of misclassification with the increasing of the minimum distances between the testing and the measurement locations. In terms of the changes, the BFNN model has several advantages over other models. For the FNN model, the percentages of misclassification does not depend on the minimum distances, and lithofacies at any location is estimated only from the corresponding collocated geophysical data. Although the method is as efficient as the BFNN model when the testing locations are far away from the measurement locations, it is limited when the testing locations are close to the measurement locations. In the case, the BFNN model takes advantages of using spatial correlation of lithofacies and reduces the percentages of misclassification significantly.

For the indicator kriging, the percentages of misclassification increase with the increasing of the minimum distances as shown in Figure 4.9. As the distances become large, the lithofacies measurements collected from the wellbores have a weak influence on lithofacies estimation. As a result, the percentages of misclassification tend to a

constant value determined by the unconditional probability of each lithofacies. The BFNN model in the case adds geophysical information to the estimation that leads to a considerable reduction in percentages of misclassification.

The indicator cokriging seems to have a similar performance to the BFNN model in terms of the percentages of misclassification, as shown in Figure 4.9, but this is not a general case. The indicator cokriging, as given in Equation 4.18, is a linear prediction method that relies on correlation and cross correlation coefficients. In this case study, there are only two lithofacies, and the nonlinearity of the cross correlation between the geophysical attributes and lithofacies is not very high. The advantages of using the BFNN model are therefore not obvious compared to the indicator cokriging. When the nonlinearity of the cross correlation increases, however, the indicator cokriging will not be as efficient as the BFNN model, which will be shown in the next case study.

## 4.4 Case Study 2

The second case study demonstrates the effect of the nonlinear cross correlation between lithofacies and geophysical attributes on the performances of the BFNN model and the indicator cokriging using synthetic data sets. Since the nonlinearity of the cross correlation usually increases with the increasing of the number of lithofacies, the BFNN model will be compared with the indicator cokriging using the synthetic

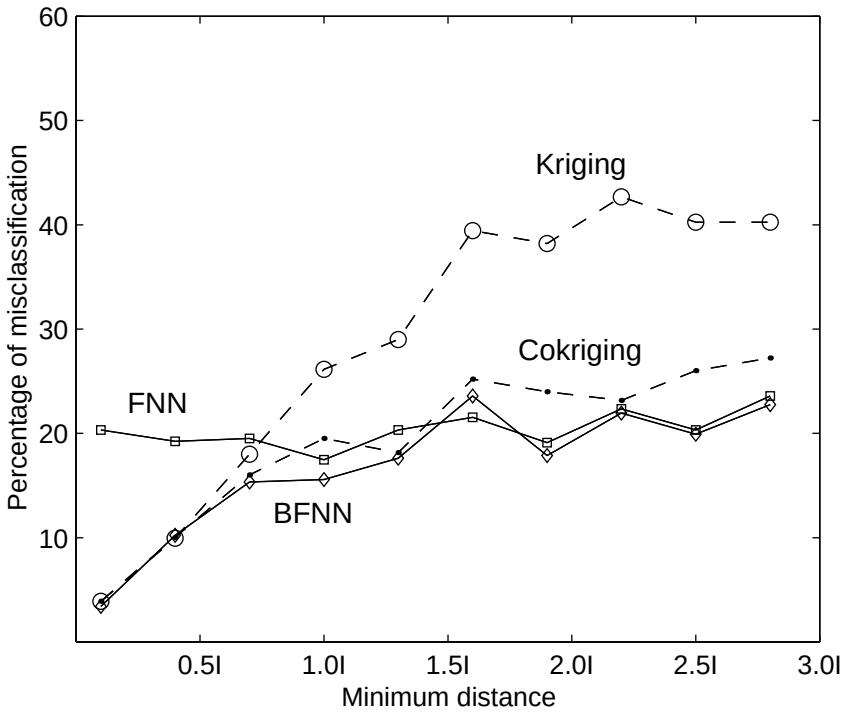


Figure 4.9: Comparison of misclassification using the synthetic data set, where $I$ =10 m is the integral length of sand.

data with two, three, or four lithofacies.

### 4.4.1 Synthetic Data

Synthetic data for the case study include three lithofacies fields, shown in Figure 4.10, and geophysical data along several wellbores, shown as solid lines in the same figure. Each field is generated using the indicator sequential simulation method [*Deutsch and Journel*, 1998] and has a size of 120 m×40 m. The theoretical indicator variograms of different lithofacies used for generating the random fields are the same, which is the exponential model with an integral length of 10 m along the lateral direction and 1 m along the vertical direction.

The hypothetical wellbores were obtained by evenly sampling the previously generated lithofacies fields at 20 m intervals. Along each of the wellbores, electrical resistivity and seismic velocity are randomly generated using the Gaussian random generator. The means and standard deviations of the random generators are determined according to the collocated lithofacies and the data given in Table 4.2. The mean values of resistivity in the table were obtained from resistivity logs collected from w1250 at the LLNL site, and the mean values of seismic velocity were chosen based on published parameter ranges for unconsolidated saturated sediments [*Lankston*, 1990; *Hyndman et al.*, 1994]. The complexity of the cross correlation can be changed by adjusting the standard deviations of those data. Figure 4.11 shows the cross plots of the synthetic data at the five wellbores generated using the parameters given in

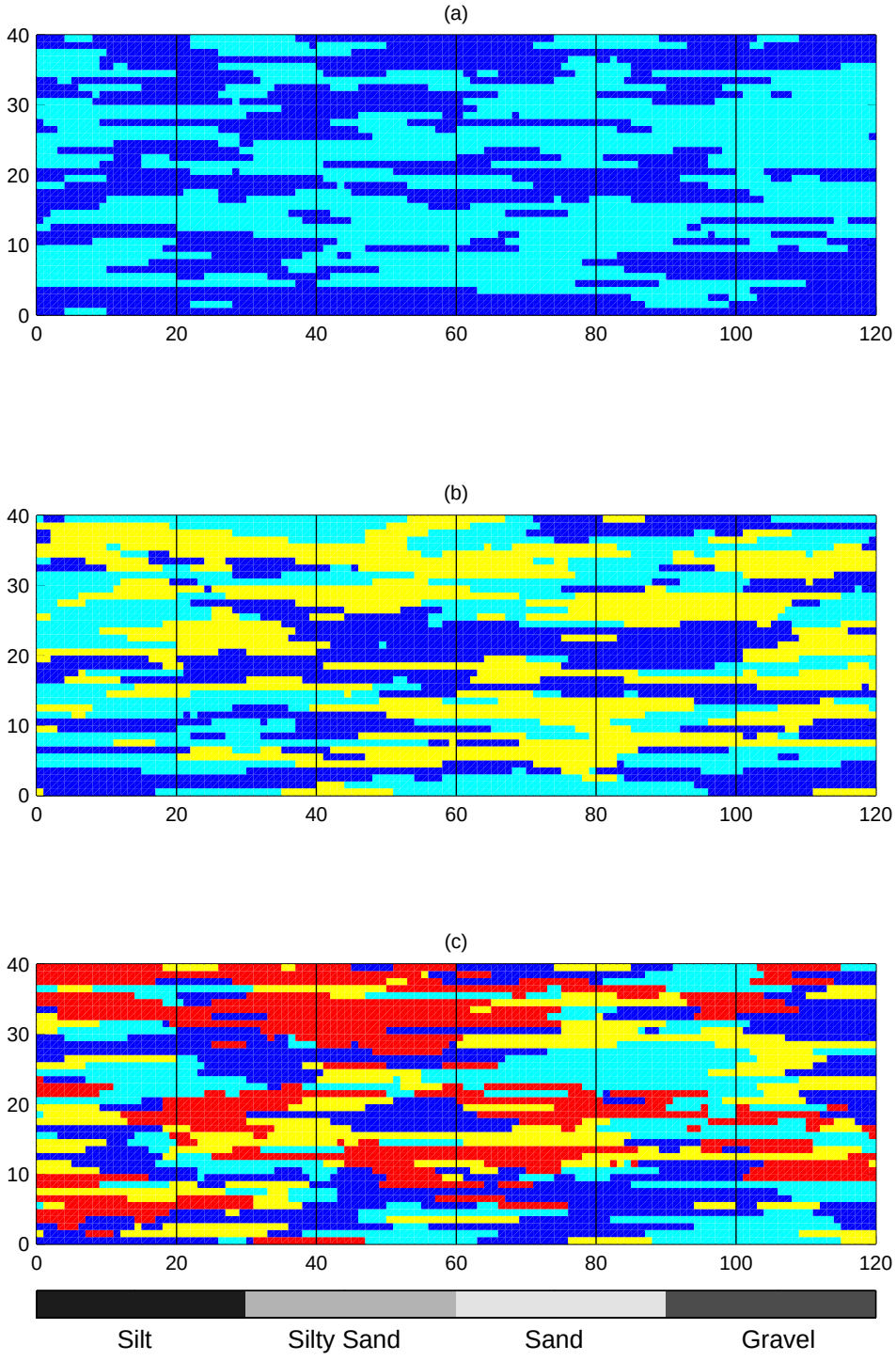


Figure 4.10: Lithofacies fields (a) with silt and silty-sand, (b) with silt, silty-sand and sand, and (c) with silt, silty-sand, sand and gravel.

Table 4.2: Means and standard deviations of geophysical data for each lithofacies

| Lithology | Resistivity ($\Omega m$) | | Seismic Velocity (m/s) | |
|---|---|---|---|---|
| | mean | Standard Deviation | mean | Standard Deviation |
| Silt | 10.43 | 1.23 | 1520 | 130.0 |
| Silty-Sand | 12.89 | 2.10 | 1780 | 105.0 |
| Sand | 18.81 | 2.90 | 1830 | 105.0 |
| Gravel | 25.65 | 1.42 | 1620 | 105.0 |

Table 4.2. It is clear that the nonlinearity of the cross correlation increases as the number of lithofacies increases.

### 4.4.2 Approach

The approach used in the case study is similar to the one used by *Chen et al.* [2001]. Each of the five wellbores is in turn to be considered as a testing well and the corresponding other four wellbores are taken as a training set. The data at the training set are first used to train the BFNN model and the indicator cokriging, and then the trained models are used to estimate lithofacies at each testing location. The estimated values are finally compared with the corresponding true values. The performance of each model is evaluated by analyzing the percentages of misclassification, similar to the previous case study.

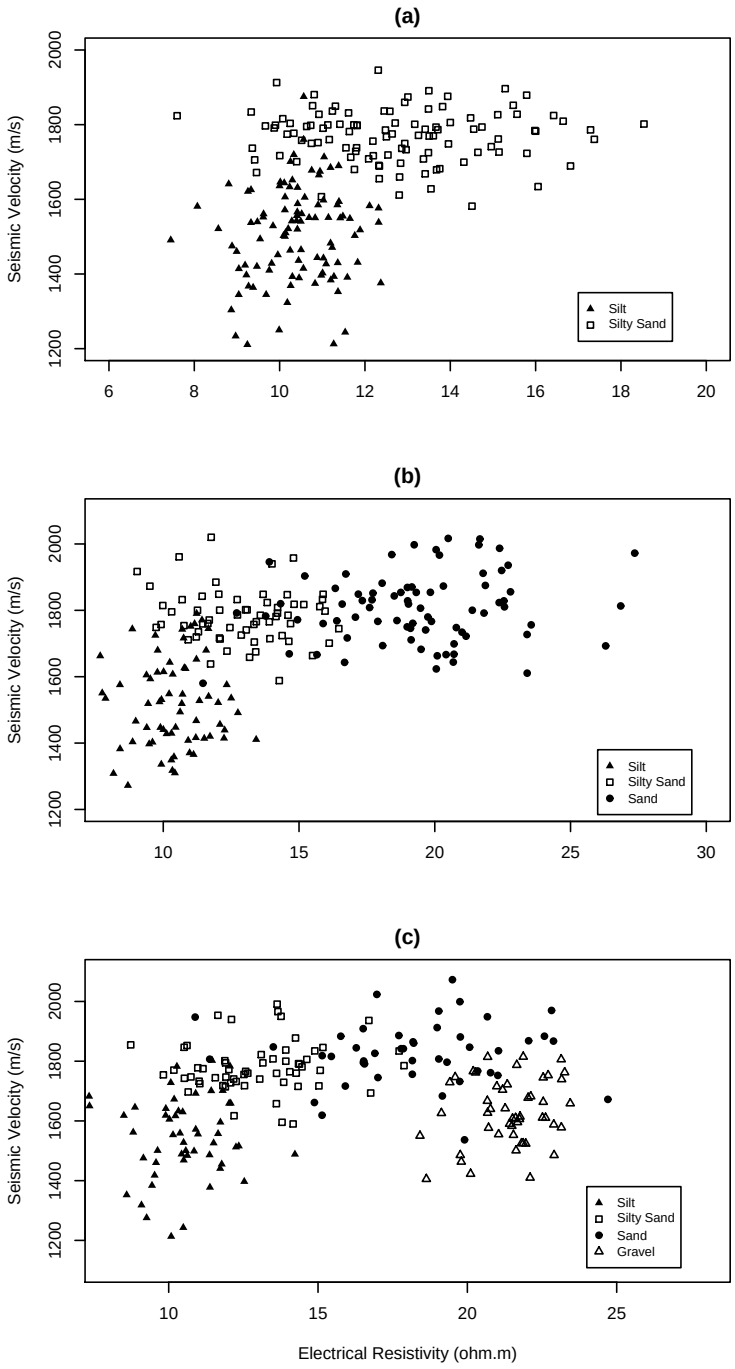


Figure 4.11: Cross-plots of resistivity versus seismic velocity (a) with two lithofacies, (b) with three lithofacies, and (c) with four lithofacies.

### 4.4.3 Results

Table 4.3 is a summary of percentages of misclassification for each data set shown in Figure 4.11. Although the number of testing wells is only five, the effects of nonlinearity on the reduction in percentages of misclassification are observed from Table 4.3. For the lithofacies field with two lithofacies, the nonlinearity of the cross correlation is not high, and the difference in percentages of misclassification between the BFNN model and the indicator cokriging is not significant. That is consistent with the previous case study. However, as the number of lithofacies or the nonlinearity of the cross correlation increases, the differences between the two models become evident. Similar results can also be obtained if the interval between the sampled wellbores are reduced.

## 4.5 Discussion

A Bayesian model coupled with a fuzzy neural network (BFNN) was developed to estimate lithofacies in this chapter using lithofacies core measurements and geophysical data. The prior probability is estimated from lithofacies measurements only using indicator kriging based on spatial correlation of the lithofacies, whereas the posterior probability is updated from the prior using geophysical data through the likelihood function. The efficiency of the model in combining the lithofacies measurements and the geophysical data was demonstrated using two synthetic data sets.

Table 4.3: Percentages of misclassification

| | Testing Well | Kriging (%) | Cokriging (%) | BFNN (%) |
|---|---|---|---|---|
| Pattern-1 | | | | |
| | Well-1 | 54 | 5 | 7 |
| | Well-2 | 39 | 7 | 5 |
| | Well-3 | 39 | 5 | 2 |
| | Well-4 | 36 | 2 | 0 |
| | Well-5 | 46 | 12 | 5 |
| | Average | 43 | 6 | 4 |
| | Standard Deviation | 7 | 6 | 4 |
| Pattern-2 | | | | |
| | Well-1 | 63 | 20 | 24 |
| | Well-2 | 56 | 22 | 7 |
| | Well-3 | 39 | 12 | 10 |
| | Well-4 | 46 | 10 | 7 |
| | Well-5 | 49 | 29 | 15 |
| | Average | 51 | 19 | 13 |
| | Standard Deviation | 9 | 8 | 7 |
| Pattern-3 | | | | |
| | Well-1 | 76 | 37 | 24 |
| | Well-2 | 66 | 39 | 10 |
| | Well-3 | 68 | 24 | 22 |
| | Well-4 | 71 | 34 | 24 |
| | Well-5 | 71 | 41 | 17 |
| | Average | 70 | 35 | 20 |
| | Standard Deviation | 4 | 7 | 6 |

The BFNN is the most effective method among indicator kriging, indicator cokriging and fuzzy neural networks without using spatial correlation (FNN). Each of the alternatives can be considered as a special case of the BFNN in different situations. The BFNN is similar to the indicator kriging when estimating locations are close to boreholes, similar to the FNN when estimating locations are far away from the boreholes, and similar to the indicator cokriging when the number of lithofacies is less than three or the nonlinearity of cross correlation is not high. The BFNN is particularly useful for cases where the nonlinearity of cross correlation between lithofacies and geophysical attributes is very high and estimating locations are within two or three integral lengths of lithofacies. That is a typical situation in subsurface characterization for the purpose of environmental remediation, such as at the bacterial transport site in Oyster (VA) and at the geochemical transport site in Livermore (CA).

The primary focus of the study is on the inference of the likelihood function using a fuzzy neural network from cross correlation between lithofacies and geophysical attributes. This relation is usually very complex and site-specific due to the difference in measurement scales of lithofacies and geophysical data and due to uncertainty associated with acquisition and interpretation of the geophysical data. Making a-priori assumptions about the relation is very difficult, especially when the numbers of lithofacies and geophysical attributes are more than two. The fuzzy neural network as an efficient fitting model, however, allows inferring the likelihood function directly

from training data sets without making any assumption about the form of the nonlinear function. The scale difference between lithofacies and geophysical data and uncertainty related to data collection are implicitly considered in this model.

Although the method is oriented toward the LLNL project where there are two different geophysical attributes that have been shown most informative to lithofacies estimation, it can be directly used to cases where there are more than two types of geophysical data, such as in *Doveton* [1986] and *Rogers et al.* [1992]. The BFNN method is also very efficient in handling multi-dimensional data sets because cross correlation is extracted using a fuzzy neural networks, which allows multiple data as input [*Takagi and Sugeno*, 1985]. With the use of the fuzzy neural network, complex patterns inherent in the multi-dimensional data sets can be extracted, which are very difficult otherwise.

The limitation of the BFNN results from the assumptions that collocated geophysical data are available at any estimating location and geophysical data satisfy the Markov condition. This is applicable when lithofacies along geophysical tomographic profiles need to be estimated or when there are many borehole measurements and crosshole geophysical data so that geophysical data can be interpolated to any estimating location. To estimate lithofacies in other situations, however, more sophisticated models may need to be developed, such as the Markov Chain Monte Carlo (MCMC) method [*Gilks et al.*, 1998], in which the nonlinear cross correlation can be simulated as a mixing distribution and the fuzzy neural network will be used to

explore the structure of the mixing model.

# Chapter 5

# Summary

This dissertation is centered on the joint use of hydrogeological and geophysical data for subsurface characterization. Three case studies were presented, each of which focuses on an important aspect of the data assimilation using Bayesian methods with differences in the inference of likelihood functions.

The first case study explores the use of GPR tomographic velocity, GPR tomographic attenuation and seismic tomographic velocity for hydraulic conductivity estimation at the South Oyster Site, using a Bayesian framework. Since site-specific relations between hydraulic conductivity and geophysical properties are often nonlinear and subject to a large degree of uncertainty such as in this site, a normal linear regression model is developed that allows exploring these relationships systematically. Although the log-conductivity displays a small variation and the geophysical data vary over only a small range, results indicate that the geophysical data improve the estimates of the hydraulic conductivity. The improvement is the most significant where prior information is limited. Among the geophysical data, GPR and seismic velocity are more useful than GPR attenuation.

In the second case study, a Bayesian approach for combining well logs and geophys-

ical surveys is presented to improve subsurface characterization. The main challenge is in creating the bridge to link between ambiguously related geophysical surveys and well logs. The second challenge is imposed by the disparity between the scales of the geophysical surveys and the well logs. This approach intends to transform the well log data so that they can be updated by the geophysical surveys, and this tends to be a convoluted process. The method starts with generating lithofacies images, conditional on well logs. Each of the images is used as the basis for generating a series of shaliness images, again conditional on well logs data. The shaliness images are converted into resistivity images to create an interface with the crosswell resistivity surveys using a site-specific petrophysical model relating between shaliness, resistivity and lithofacies. The lithofacies and resistivity images are then updated using the cross-well resistivity surveys. The limitations of the approach were also explored using synthetic surveys with different resolutions and error levels, which closely mimic the conditions at the LLNL site. Results reveal that the proposed method enhances hydrogeological site characterization even when the resistivity surveys have a relatively low resolution.

In the last case study, a Bayesian model coupled with a fuzzy neural network (BFNN) is developed to alleviate the difficulty of using geophysical data in lithofacies estimation when cross correlation between the lithofacies and the geophysical attributes is nonlinear. The prior estimates are inferred from borehole lithofacies measurements using indicator kriging based on spatial correlation of the lithofacies, and the posterior estimates are obtained from updating of the prior using the geo-

physical data. The novelty of the study lies in the use of a fuzzy neural network for the inference of the likelihood function. This allows incorporating the spatial correlation as well as the nonlinear cross correlation into lithofacies estimation. The effectiveness of the BFNN is demonstrated using synthetic data generated from measurements at LLNL site.

# Bibliography

[1] Abramovich, B. and Indelman, P. (1995). Effective permitivity of log-normal isotropic random media. *J. Phys. A Math. Gen.*, 28(3):693–700.

[2] Ahmed, S., De Marsily, G., and Talbot, A. (1988). Combined use of hydraulic and electrical propeties of an aquifer in a geostatistical estimation of transmissivity. *Ground Water*, 26(1):78–86.

[3] Bachrach, R., Dvorkin, J., and Nur, A. (1998). High-resolution shallow-seismic experiments in sand, part ii: Velocities in shallow unconsolidated sand. *Geophysics*, 63(4):1234–1240.

[4] Bachrach, R., Dvorkin, J., and Nur, A. (2000). Seismic velocities and poisson's ratio of shallow unconsolidated sands. *Geophysics*, 65(2):559–564.

[5] Bachrach, R. and Nur, A. (1998). High-resolution shallow-seismic experiments in sand, part i: Water table, fluid flow, and saturation. *Geophysics*, 63(4):1225–1233.

[6] Beard, L., Hohmann, G. W., and Tripp, A. C. (1996). Fast resistivity/ip inversion using a low-constant approximation. *Geophysics*, 61(1):169–179.

[7] Benaouda, D., Wadge, G., Whitmarsh, R., Rothwell, R., and Macleod, C. (1999). Inferring the lithology of borehole rocks by applying neural network classifiers to

downhole logs: An example from the ocean drilling program. *Geophysics Journal International*, 136:477–491.

[8] Bernardo, J. M. and Smith, A. F. (1994). *Bayesian Theory.* John Wiley and Sons, New York.

[9] Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York and London.

[10] Blake, R., Maley, M., and Noyes, C. (1995). *Characterization of VOC Transport in Heterogeneous Sediments at Lawrence Livemore National Laboratory and its Impact on Remediation Decisions.* Assoc. of Eng. Geol., Sacramento, California.

[11] Bosch, M. (1999). Lithologic tomography: From plural geophysical data to lithology estimation. *Journal of Geophysical Research*, 104(B1):749–766.

[12] Bosch, M., Guillen, A., and Ledru, P. (2001). Lithologic tomography: an applicattion to geophysical data from the cadomian belt of northern brittany, france. *Tectonophysics*, 31:197–227.

[13] Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, New York.

[14] Burger, H. R. (1992). *Exploration Geophysics of the Shallow Subsurface.* Prentice Hall, Englewood Cliffs, NJ.

[15] Buselli, G., Barber, C., Davis, G. B., and Salama, R. B. (1990). Detection of groundwater contamination near waste disposal sites with transient electromagnetic and electrical methods. In Ward, S. H., editor, *Geotechnical and Environmental Geophysics*, Tulsa, Okla, Society of Exploration Geophysics (SEG).

[16] Carman, P. C. (1956). *Flow of Gases Through Porous Media.* Academic Press Inc., New York.

[17] Cassiani, G., Bohm, G., Vesnaver, A., and Nicolich, R. (1998). A geostatistical framework for incorporating seismic tomography auxiliary data into hydraulic conductivity estimation. *Journal of Hydrology*, 206.

[18] Chang, H.-C., Chen, H.-C., and Fang, J.-H. (1997). Lithology determination from well logs with fuzzy associative memory neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 35(3):773–780.

[19] Chen, J., Hubbard, S., and Rubin, Y. (2001). Estimating the hydraulic conductivity at the south oyster site from geophysical tomographic data using bayesian techniques based on the normal linear regression model. *Water Resources Research*, 37(6).

[20] Chew, W. C. (1995). *Waves and Fields in Inhomogeneous Media, Ser. Electromagn. Waves Theory.* IEEE Press, Piscataway, NJ.

[21] Copty, N. and Rubin, Y. (1995). A stochastic approach to the characteriza-

tion of lithofacies from surface seismic and well data. *Water Resources Research*, 31(7):1673–1686.

[22] Copty, N., Rubin, Y., and Mavko, G. (1993). Geophysical-hydrological identification of field permeabilities through bayesian updating. *Water Resources Research*, 29(8):2813–2825.

[23] Dagan, G. (1982). Stochastic modeling of groundwater flow by unconditional and conditional probabilities 2. the solute transport. *Water Resources Research*, 18(4):835–848.

[24] Dagan, G. (1985). Stochastic modeling of groundwater flow by unconditional and conditional probabilities: The inverse problem. *Water Resources Research*, 21(1):67–72.

[25] Dagan, G. (1989). *Flow and Transport in Porous Formations*. Springer-Verlag, New York.

[26] Daily, W., Ramirez, A., LeBrecque, D., and Nitao, J. (1992). Electrical resistivity tomography of vadose water movement. *Water Resources Research*, 28(5):1429–1442.

[27] Davis, J. and Annan, A. (1989). Ground penetrating radar for high resolution mapping of soil and rock stratigraphy. *Geophysical Prospect*, 37:531–551.

[28] DeFlaun, M. F., Murray, C. J., Holben, W., Scheibe, T., Mills, A., Ginn, T.,

Griffin, T., Majer, E., Wilson, and L., J. (2000). Transport of bacteria in a coastal plain aquifer under forced gradient conditions. *EOS, submitted.*

[29] Delfiner, P., Peyret, O., and Serra, O. (1987). Automatic determination of lithology from well logs. In *SPE Formation Evaluation*, volume 2, pages 303–310. The Society of Petroleum Engineers.

[30] Deutsch, C. and Journel, A. G. (1998). *GSLIB: Geostatistical Software Library and User's Guide.* Oxford University Press, New York.

[31] Doveton, J. H. (1986). *Log Analysis of Subsurface Geology Concepts and Computer Methods.* John Wiley & Sons, New York.

[32] Doyen, P. M. (1988). Porosity from seismic data: A geostatistical approach. *Geophysics*, 53(10):1263–1276.

[33] Ezzedine, S., Rubin, Y., and Chen, J. (1999). Hydrological-geophysical bayesian method for subsurface site charaterization: Theory and application to llnl superfund site. *Water Resources Research*, 35(9):2671–2683.

[34] Gassmann, F. (1951). Elastic waves through a packing of spheres. *Geophysics*, 16:673–685.

[35] Gelhar, L. W. and Axness, C. L. (1983). Three-dimensional stochastic analysis of macrodispersion in aquifer. *Water Resources Research*, 19(1):161–180.

[36] Geller, J. T. and Myer, L. R. (1995). Ultrasonic imaging of organic liquid contaminants in unconsolidated porous media. *Journal of Contaminant Hydrology*, 19:85–104.

[37] Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1998). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, New York.

[38] Ginn, T. R. and Cushman, J. H. (1990). Inverse methods for subsurface flow: a critical review of stochastic techniques. *Stochastic Hydrology and Hydraulics*, 4:1–26.

[39] Han, D. H., Nur, A., and Morgan, D. (1986). Effects of porosity and clay content on wave velocities in sandstones. *Geophysics*, 51:2093–2107.

[40] Hill, D. G. (1986). Geophysical well log calibration and quality control. In *Borehole Geophysics for Mining and Geotechnical Applications, Publ. 85-27*, pages 379–392, Geol. Surv. of Can.

[41] Horikawa, S. et al. (1991). A study on fuzzy modeling using fuzzy neural networks. *Fuzzy Engineering toward Human Friendly Systems*, pages 562–573.

[42] Hubbard, S. and Rubin, Y. (2000). Hydrogeological parameter estimation using geological data: A review of selected techniques. *Journal of Contaminant Hydrology*, 45:3–34.

[43] Hubbard, S., Rubin, Y., and Majer, E. (1997). Ground penetrating radar assisted saturation and permeability estimation in bimodal systems. *Water Resources Research*, 33(5):971–990.

[44] Hubbard, S., Rubin, Y., and Majer, E. (1999). Spatial correlation structure using geophysical and hydrogeological data. *Water Resources Research*, 35(6):1809–1825.

[45] Hyndman, D. W. and Gorelick, S. M. (1996). Estimating lithologic transport properties in three dimensions using seismic and tracer data: The kesterson aquifer. *Water Resources Research*, 32(9):2659–2670.

[46] Hyndman, D. W., Harris, J. M., and Gorelick, S. M. (1994). Coupled seismic and tracer test inversion for aquifer property characterization. *Water Resources Research*, 30(7):1965–1977.

[47] Isaaks, E. H. and Srivastava, R. M. (1989). *Applied Geostatistics*. Oxfoed University Press, New York.

[48] Jang, J.-S. R. (1993). Anfis: Adaptive network based fuzzy inference system. *IEEE Trans. Systems, Man, and Cybernetics*.

[49] Johnson, W., Zhang, P., Fuller, M., Scheibe, T., Mailloux, B., Onstott, T., DeFlaun, M., Hubbard, S., Radtke, J., Kovacik, W., and Holbin, W. (2001). Ferrographic tracking of bacterial transport in the field at the narrow channel focus area, oyster, va. *Environmental Science & Technology*, 35(1):182–191.

[50] Journel, A. G. (1989). *Fundamentals of Geostatistics in Five Lessons*. American Geophysical Union, Washington, D.C.

[51] Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic Press, New York.

[52] Kachanoski, R. G., Gregorich, E. G., and Wesenbeeck, I. J. (1988). Estimating spatial variations of soil water content using noncontacting electromagnetic inductive methods. *Canadian Journal of Soil Science*, 68:715–722.

[53] Keers, H., Johnson, L. R., and Vasco, D. W. (2000). Acoustic crosswell imaging using asymptotic waveforms. *Geophysics*, 65(5):1569–1582.

[54] Kelly, W. (1977). Geoelectrical sounding for estimating hydraulic conductivity. *Ground Water*, 15(6):420–425.

[55] Keys, W. S. (1989). *Borehole Geophysics Applied to Ground-Water Investigations*. National Water Association, Dublin, Ohio.

[56] Keys, W. S. (1997). *A Practical Guide to Borehole Geophysics in Environmental Investigations*. Lewis Publishers, New York.

[57] Kitanidis, P. (1998). Geostatistics: Interpolation and inverse problems. In Delleur, J. W., editor, *The Handbook of Groundwater Engineering*, CRC Press, New York.

[58] Kitanidis, P. K. (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research*, 22(4):499–507.

[59] Kitanidis, P. K. (1991). Orthonomal residuals in geostatistics: Model criticism and parameter estimation. *Mathematical Geology*, 23(5):741–758.

[60] Klimento, T. and McCann, C. (1990). Relationships among compressional wave attenuation, porosity, clay content, and permeability in sandstones. *Geophysics*, 55(8):998–1014.

[61] Knoll, M. (1996). *A Petrophysical Basis for Ground Penetrating Radar and Very Early Time Electromagnetics: Electrical Properties of Sand-Clay Mixtures.* PhD thesis, The University of British Columbia, Canada.

[62] Knoll, M. D., Haeni, F. P., and Knight, R. J. (1991). Characterization of a sand and gravel aquifer using ground penetrating radar, cape code, massachusetts. In *U.S. Geological Survey*, pages 29–35.

[63] Lankston, R. W. (1990). High-resolution refraction seismic data acquisition and interpretation. In Ward, S. H., editor, *Geotechnical and Environmental Geophysics*, Tulsa, Okla, Society of Exploration Geophysics (SEG).

[64] Lortzer, G. J. M. and Berkhout, A. J. (1992). An integrated approach to litho-logic inversion, part i, theory. *Geophysics*, 57:233–244.

[65] Lucet, N. and Mavko, G. (1991). Images of rock properties estimated from a

crosswell seismic velocity tomogram, (abstract). In *Technical Program and Abstracts of Papers*, Eur. Assoc. of Explor. Geophys., Florence, Italy.

[66] Marion, D. (1990). *Acoustical, Mechanical and Transport Properties of Sediments and Granular Materials*. PhD thesis, Stanford University, Stanford, California.

[67] Marion, D., Nur, A., and Han, D. (1992). Compressional velocity and porosity in sand-clay mixtures. *Geophysics*, 57(4):554–563.

[68] Mavko, G., Mukerji, T., and Dvorkin, J. (1998). *The Rock Physics Handbook Tools for Seismic Analysis in Porous Media*. Cambridge University Press, New York.

[69] Mazac, O., Kelly, W. E., and Landa, I. (1985). A hydrogeophysical model for relations between electrical and hydraulic properties of aquifers. *Journal of Hydrology*, 79:1–19.

[70] McLaughlin, D. and Townley, L. R. (1996). A reassessment of the groundwater inverse problem. *Water Resources Research*, 32(5):1131–1161.

[71] Mixon, R. B. (1985). Stratigraphic and geologic framework of the upper cenozoic deposits in the southern delmarva peninsula, paper 1067-g. *U.S. Geol. Surv. Prof. Paper*, page 53p.

[72] Molz, F. and Young, S. (1993). Development and application of borehole flowmeters for environmental assessment. *The Log Analyst*, 3:13–23.

[73] Nikravesh, M. (1998). Neural network knowledge-based modeling of rock properties based on well log databases. *BISC-SIG-ES Seminar.*

[74] Noyes, C. D. (1991). *Hydrostratigraphic Analysis of the Pilot Remediation Test Area, Lawrence Livemore National Laboratory, Livemore, CA*. PhD thesis, University of California at Davis, California.

[75] Parker, R. L. (1994). *Geophysical Inverse Theory.* Princeton University Press, Princeton, NJ.

[76] Peterson, J. E. (2000). Pre-inversion processing and analysis of tomographic radar data. *Journal of Environmental and Engineering Geophysics (accepted).*

[77] Peterson, J. E., Paulsson, B. N., and McEvilly, T. V. (1985). Applications of algebraic reconstruction techniques to crosshole seismic data. *Geophysics*, 50:1566–1580.

[78] Rector, J. W. (1995). Crosswell methods: Where are we, where are we going? *Geophysics*, 60(6):629–630.

[79] Rogers, S. J., Fang, J., Karr, C., and Stanley, D. (1992). Determination of lithology from well logs using a neural network. *The American Association of Petroleum Bulletin*, 76(5):731–739.

[80] Rojas, R. (1996). *Neural Networks: A Systematic Introduction.* Springer, New York.

[81] Rosenbaum, M., Rosen, L., and Gustafson, G. (1997). Probabilistic models for estimating lithology. *Engineering Geology*, 47:43–55.

[82] Rubin, Y. (1995). Flow and transport in bimodel heterogeneous formations. *Water Resources Research*, 31(10):2461–2468.

[83] Rubin, Y. and Dagan, G. (1987). Stochastic identification of transmissivity and effective recharge in steady groundwater flow, i. theory. *Water Resources Research*, 23(7):1185–1192.

[84] Rubin, Y., Hubbard, S., Wilson, A., and Cushey, M. (1998). Aquifer characterization. In Delleur, J. W., editor, *The Handbook of Groundwater Engineering*. CRC Press, New York.

[85] Rubin, Y., Mavko, G., and Harris, J. (1992). Mapping permeability in heterogeneous aquifers using hydrologic and seismic data. *Water Resources Research*, 28(7):1809–1816.

[86] Salem, H. S. (2001). Modelling of lithology and hydraulic conductivity of shallow sediments from resistivity measurements using schlumberger vertical electric soundings. *Energy Sources*, 23:599–618.

[87] Scheibe, T. D., Chien, Y., Ginn, T., Murray, C., and Hubbard, S. S. (1999). Heterogeneity characterization for field-scale bacterial transport modeling. *EOS*, 80(17).

[88] Serra, O. (1984). Fundamentals of well-log interpretation. *Developments in Petroleum Science*, 15A-B.

[89] Shapiro, S. A., Audigane, P., and Royer, J.-J. (1999). Large-scale in site permeability tensor of rocks from induced microseismicity. *Geophysical Journal International*, 137:207–213.

[90] Sheets, K. R. and Hendricks, J. M. H. (1995). Noninvasive soil water content measurement using electromagnetic induction. *Water Resources Research*, 31(10):2401–2409.

[91] Stone, C. J. (1995). *A Course in Probability and Statistics*. Duxbury Press, New York.

[92] Sun, N. Z. (1994). *Inverse Problems in Groundwater Modeling*. Kluwer Academic Publishers, Dordrecht, Netherlands.

[93] Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(1).

[94] Telford, W. M., Geldart, L. P., and Sheriff, R. E. (1990). *Applied Geophysics*. Cambridge Univ. Press, New York.

[95] Topp, G. C., Davis, J. L., and Anna, A. P. (1980). Electromagnetic determination

of soil-water content: Measurement in co-axial transmission lines. *Water Resources Research*, 16:574–582.

[96] Urish, D. W. (1981). Electrical resistivity-hydraulic conductivity relationships in glacial outwash aquifers. *Water Resources Research*, 17(5):1401–1408.

[97] Vasco, D. W., Peterson, J. E., and Major, E. L. (1996). A simultaneous inversion of seismic traveltimes and amplitudes for velocity and attenuation. *Geophysics*, 61(6):1738–1757.

[98] Vernik, L. and Nur, A. (1992). Petrophysical classification of siliciclastics for lithology and porosity prediction from seismic velocities. *The American Association of Petroleum Geologist Bulletin*, 76(9):1295–1309.

[99] Waldrop, W. and Hamby, J. (1998). Results of field tests with the electromagnetic borehole flowmeter at the narrow channel focus area, oyster, va. Technical report, Quantum Engineering Corporation, Loudon, Tennessee.

[100] Welhan, J. and Reed, M. (1997). Geological analysis of regional hydraulic conductivity variations in the snake river plain aquifer, eastern idaho. *Geological Society of America*, 109(7):855–868.

[101] Williamson, P. R. and Worthington, M. H. (1993). Resolution limits in ray tomography due to wave behavior: Numerical experiments. *Geophysics*, 58(5):727–735.

[102] Woodbury, A. D. and Ulrych, T. J. (1993). Minimum relative entropy: Forward probabilistic modeling. *Water Resources Research*, 29(8):2847–2860.

[103] Wyatt, D. E., Waddell, M. G., and Sexton, G. B. (1996). Geophysics and shallow faults in unconsolidated sediments. *Ground Water*, 34(2):326–334.

[104] Xie, X. and Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[105] Yeh, W. W. G. (1986). Review of parameter identification procedures in groundwater hydrology: the inverse problem. *Water Resources Research*, 22(2):95–108.

[106] Zadeh, L. A. (1965). Fuzzy sets. *Inform. ontrol.*, 8:338–353.

[107] Zodhy, A. A., Eaton, G. P., and Mayhey, D. R. (1974). Applications of surface geophysics to groundwater investigations. In *Techniques of Water Resources Investigations of the U.S. Geological Survey.*

# Appendix A

# Synthetic Electromagnetic Survey

The theoretical foundation of the EM survey is based on Maxwell's wave propagation equations. These latter couple the electric field $\mathbf{E}$ to the magnetic field $\mathbf{H}$ as follows:

$$\nabla \times \mathbf{E}(\mathbf{x}, t) = -\frac{\partial}{\partial t}\mathbf{B}(\mathbf{x}, t) \tag{A.1}$$

$$\nabla \times \mathbf{H}(\mathbf{x}, t) = \frac{\partial}{\partial t}\mathbf{D}(\mathbf{x}, t) + \mathbf{J}(\mathbf{x}, t) \tag{A.2}$$

$$\nabla \cdot \mathbf{B}(\mathbf{x}, t) = 0; \text{ and } \nabla \cdot \mathbf{D}(\mathbf{x}, t) = Q(\mathbf{x}, t) \tag{A.3}$$

where $\mathbf{D}$ is the electric flux, $\mathbf{B}$ the magnetic flux, $\mathbf{J}$ is the current density, and $Q$ is the charge density [*Chew*, 1995]. For time varying EM fields, Equations A.3 can be derived from the first two equations using the continuity equation given by:

$$\nabla \mathbf{J} = -\frac{\partial Q}{\partial t} = -\frac{\partial}{\partial t}\left(\nabla \cdot \mathbf{D}\right) \tag{A.4}$$

In this static case, the electric field and the magnetic one are decoupled, and the electric field equations can be solved independently from the magnetic ones. This approximation should be viewed as an approximation, applicable in cases where the spatial variability of the conductivity is weak. This approximation has already been used in previous studies [*Beard et al.*,1996].

In the case of the HSU2 at the LLNL, the mean of the resistivity is $11\Omega m$ and $17\Omega m$, in silt and sand, respectively. The variance of the resistivity in silt and sand are 2 and $3.5(\Omega m)^2$. These statistics show that the contrast of variability between and within each silt and sand bodies are similar, and assumptions similar to those of *Beard et al.* [1996] can be made.

Defining the current density $\mathbf{J}$ as [*Telford et al.*, 1990]:

$$\mathbf{J} = R^{-1}\mathbf{E};\ \mathbf{E} = -\nabla\mathbf{V} \tag{A.5}$$

where $V$ is the potential, it satisfies the continuity equation (A.4), which leads to

$$\nabla(\kappa\nabla V) = 0 \tag{A.6}$$

where $\kappa = 1/R$ is the electric conductivity.

Equation (A.6) describes the response of the domain at the smallest scale over which $\kappa$ can be defined. In a geophysical survey, the small-scale variability cannot be detected. Instead, large parts of the domain are homogenized, and respond as a homogeneous block. The geophysical survey defines the block conductivity, $\kappa_b = 1/\rho$, for which the continuity equation becomes:

$$\nabla(\kappa_b\nabla V_b) = 0 \tag{A.7}$$

where $V_b$ describes the potential field in the homogenized resistivity field, subject to the same boundary conditions as in (A.6).

# Appendix B

# Conditional Mean Sampling

For an arbitrary covariance matrix, generating random samples from a joint normal distribution with given values of some linear combinations is not difficult, because all linear combinations of jointly normal random variables are jointly normal, which means that the conditional distributions are also jointly normal. The procedure is to subtract the regression of the various variables on the constraints, i.e., assuming the $n$ random variables $r_1, \cdots, r_i, \cdots, r_n$ are generated from different Gaussian pdf $f_1''(r; m_1'', \sigma_1'')$, $\cdots$, $f_n''(r; m_n'', \sigma_n'')$, respectively, all $r_i$ have to average to the resistivity survey $\rho$. For an arithmetic average, the difference between $\rho$ and the sampled $\bar{r} = \sum_l r_l/n$ is then subtracted from each $r_l$ as follows:

$$r_l'' = r_l - (\bar{r} - \rho) \tag{B.1}$$

where $r''$ is the posterior resistivity. For a geometric mean the procedure remains valid, but the resistivity should be replaced by its logarithm: $\ln(r)$.

To take into account the variability for each pdf, one should weight the corrected resistivity value with respect to their relative inertia, as follows:

$$r_l'' = r_l - n\sigma_l^2(\bar{r} - \rho)/\sum_m \sigma_m^2 \tag{B.2}$$

Since corrections are deterministic, statistics of each distribution remain the same.