The Thesis Committee for Morteza Elahi Naraghi
Certifies that this is the approved version of the following thesis :

Geostatistical Data Integration in Complex Reservoirs

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Sanjay Srinivasan

Mrinal K. Sen

Geostatistical Data Integration in Complex Reservoirs

By

Morteza Elahi Naraghi, B.S.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

December 2014

Dedication

To my mom, the one who I owe my entire life.

To my dad, for his endless love, support, and believe in me.

To Bahar, for her love, patience, and understanding.

To Ali, for cheering me up in the toughest situations.

To my grandparents, for teaching me the importance of hard work.

Acknowledgements

Completion of this thesis would not have been possible without the support and contribution of many people. It is a great honor for me to thank some of those many, to whom I owe my deepest gratitude.

I would like to express my deepest appreciation to my supervisor, Dr. Sanjay Srinivasan for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. He continually and convincingly conveyed a spirit of adventure in regard to research, and courage and versatility to battle challenging problems that I face in life. His scientific intuition, genius, invaluable insights, and continuous involvement were indispensable in developing my academic merits. Additional gratitude goes to Dr. Mrinal Sen for agreeing to be my secondary reader on such short notice.

It gives me great pleasure in acknowledging the support of Dr. Kamy Sepehrnoori for helping me get through difficult times, and for all the emotional and technical guidance he provided during my life in Austin.

I would also like to acknowledge the staff of the Department of Petroleum and Geosystems Engineering, Dr. Roger Terzian, Frankie Hart, Cheryl Kruzie, Jin Lee, Joanna Castillo, Glen Baum, Gary Miscoe, John Cassibry, and Mary Pettengill for their technical and administrative support.

I would like to express my gratitude to all friends who helped and supported me throughout my study and made my life rewarding at Austin. I am especially thankful to Reza Parsi, Arjang Shahriari, Amir Khalighi, and Mohammad Lotfollahi for their support during tough times. I am very grateful to my friends, Dr. Amir Reza Rahmani, Dr. Aboulghasem Kazemi, Mohammad Reza Beigi, Mahmood Shakiba, Mohsen Rezaveisi,

Amir Kianinejad, Haiyan Zhou, Baehyun Min, Lianping Li, Harpreet Singh, Travis Payton Hampton, Dhananjay Kumar, Hariharan Ramachandran, Odeji Olaluna, Chiazor Nwachukwu, Mayuri Murugesu, Henry Li, Kwangjin Lee, Krupa Kannan, Hoonyoung Jeong, Jamses Choi, Colin Schroeder, Arindam Raina, Williams Ozowe, Sang Hyon, and many other friends that I have not mentioned their names for all their support and making my life so pleasant during these years.

Last but not least, I wish to express my sincere appreciation to my beloved parents, Sheila and Jafar, for their endless love and incredible support through my entire life. I owe them more than they know. I would like to thank Ali and Bahar for their indispensable love and believe in me. I also wish to express my gratitude to my uncle Shahrokh Hejazi and his beloved wife, Poopak Fahimi, for their indispensable support during my life in US. I am also thankful to my grandparents for their valuable advices and encouragement during my life.

Abstract

Geostatistical Data Integration in Complex Reservoirs

Morteza Elahi Naraghi, M.S.E.

The University of Texas at Austin, 2014

Supervisor: Sanjay Srinivasan

One of the most challenging issues in reservoir modeling is to integrate information coming from different sources at disparate scales and precision. The primary data are borehole measurements, but in most cases, these are too sparse to construct accurate reservoir models. Therefore, in most cases, the information from borehole measurements has to be supplemented with other secondary data. The secondary data for reservoir modeling could be static data such as seismic data or dynamic data such as production history, well test data or time-lapse seismic data.

Several algorithms for integrating different types of data have been developed. A novel method for data integration based on the permanence of ratio hypothesis was proposed by Journel in 2002. The premise of the permanence of ratio hypothesis is to assess the information from each data source separately and then merge the information accounting for the redundancy between the information sources. The redundancy between the information from different sources is accounted for using parameters (tau or nu parameters, Krishnan, 2004).

The primary goal of this thesis is to derive a practical expression for the tau parameters and demonstrate the procedure for calibrating these parameters using the available data. This thesis presents two new algorithms for data integration in reservoir modeling. The algorithms proposed in this thesis overcome some of the limitations of the current methods for data integration.

We present an extension to the direct sampling based multiple-point statistics method. We present a methodology for integrating secondary soft data in that framwork. The algorithm is based on direct pattern search through an ensemble of realizations. We show that the proposed methodology is sutiable for modeling complex channelized reservoirs and reduces the uncertainty associated with production performance due to integration of secondary data. We subsequently present the permanence of ratio hypothesis for data integration in great detail. We present analytical equations for calculating the redundancy factor for discrete or continuous variable modeling. Then, we show how this factor can be infered using available data for different scenarios. We implement the method to model a carbonate reservoir in the Gulf of Mexico. We show that the method has a better performance than when primary hard and secondary soft data are used within the traditional geostatistical framework.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

One of the most challenging issues in reservoir modeling is to integrate information coming from different sources at different scales and precision. The primary data are borehole measurements, but in general, these are too sparse to construct accurate reservoir models if used by themselves. Therefore, in most cases, the information from borehole measurements have to be supplemented with a secondary data set. The secondary data for reservoir modeling could be static data such as seismic data or dynamic data such as production history or well test data.

Many plausible reservoir models can be constructed to match the primary borehole measurement (non-uniqueness issue in reservoir modeling), not all of them represent the underlying geology. Complex curvilinear features such as channels, fractures etc. significantly impact reservoir production and inaccuracies in representing the connectivity of such features lead to inaccurate reservoir performance prediction thereby undermining the reliability of the reservoir management scenarios. Secondary data can also provide valuable information regarding the connectivity of such reservoir features.

Due to the sparseness of primary data, integration of secondary data in reservoir models is a very important and challenging task. Most of the time, secondary data such as seismic data that provide indirect information about spatial variation of reservoir attributes are available exhaustively. A method by which knowledge coming from these different sources can be combined is therefore required so that models for spatial variability of reservoir attributes can be constructed reliably. It is shown in several studies that soft data integration can tremendously improve the accuracy of the reservoir model and decrease the uncertainty associated with reservoir performance predictions.

1.1 PREVIOUS APPROACHES AND PROBLEM STATEMENT

Several algorithms for integrating different types of data have been developed. There are several ways to integrate the seismic data in geostatistical reservoir models such as kriging with an external drift (Goovaerts, 1997) or more robustly cokriging (Dubrule O. , 2003). Cokriging and cosimulation (Deutsch, C. and A. Journel, 1998; Goovaerts, 1997) perform spatial modeling by integrating the information from different data sources in order to construct the local conditional distribution quantifying the uncertainty in attribute value at the estimation or simulation location. Cokriging is an extension of kriging in which different types of data are linearly weighted in order to arrive at the estimate. The spatial variability of each type of data is considered through the appropriate autocovariance of the variable, while the redundancy between the data is considered through the cross-covariance between pairs of data. One of the major tasks in cokriging is to ensure that the matrix of auto and cross-covariances is positive definite in order to ensure a unique solution using co-kriging. One of the approaches to ensure positive-definiteness is to assume a linear model of corregionalization (Goovaerts, 1998) which prescribes that the random function being modeled can be linearly decomposed into basis random functions that are auto and cross correlated. Under this assump0tion of linear decomposition, the uniquesness of the cokriging solution is ensured provided the matrices of sill contributions of various structures making up the auto and cross-covariances are positive definite.

A spatial interpolation algorithm such as kriging or co-kriging yields a unique map where the local estimate at every location is identified with the conditional mean at that location. One of the deficiencies of these models is that the joint uncertainty at several estimated locations might not be accurate.. In addition, the maps obtained from kriging are smooth and may not represent the spatial variability of the data. These issues are addressed in stochastic simulation

where multiple possible realizations of the random function model are obtained, each honoring the same set of data constraints (Goovaerts, 1997). However, stochastic simulation algorithms based on kriging or co-kriging still require legitimate models for the auto and cross-covariances. In addition, all of these algorithms are constrained to the variogram or covariance function which is a measure of two-point statistics spatial continuity. It has been shown in several studies that covariance-based algorithms are incapable of representing the spatial connectivity of complex geological features such as channels.

A novel method for data integration based on the permanence of ratio hypothesis was proposed by Journel in 2002. The problem of integrating data from different sources can be posed as finding the conditional probability distribution of $P(A|D_1, D_2, \dots D_n)$ where A is the event being estimated and D_is are the data from different sources. In order to model the conditional probability, it would be convenient if the information from each data source can be assessed independently in order to find $P(A|D_i)$, and then merge these joint probabilities to calculate $P(A|D_1, D_2, \dots D_n)$ accounting for the redundancy between different data sources. This is precisely the premise of the permanence of ratio hypothesis. The odds-ratio measures the relative distance to the occurrence of A given D_i. If A is guranteed not to occur given D_i, then the relative distance to occurrence of A is infinity, while if A is guaranteed to occur because of the occurrence of D_i, then the relative distance is zero. The permanence of ratio hypothesis then states that the merging of the conditional probabilities from the individual sources of information can be accomplished using odd ratios computed for each source assuming that the ratio computed corresponding to one source of information remains invariant regardless of the occurrence of another data event. The redundancy between the information from different sources is subsequently accounted for using parameters (tau or nu parameters, Krishnan, 2004).

The concept of data redundancy arises because the information inferred about event A from data source D_i may exhibit some overlap with the information derived from data source D_j. Traditional data integration techniques address the concept of data redundancy through measures such as cross-covariance and mutual information (McEliece, 2002). The tau model (Krishnan s. , 2008), can accommodate any complex description of data redundancy and an exact expression for the tau parameters can be derived from a tautology. However, techniques for practical calibration of tau weights from the statistics of data are lacking, which makes the theory difficult to apply. The primary goal of this thesis is to derive a practical expression for the tau parameters and demonstrate the procedure for calibrating these parameters using the available data.

1.2 THESIS OUTLINE

This thesis presents two new algorithms for data integration in reservoir modeling. The algorithms proposed in this thesis overcome some of the limitations of the current methods for data integration. The thesis is organized as follows.

Chapter 2 presents a literature review of past works in the area of multiple point statistics and discusses current data integration algotihms.

In chapter 3, we present an extension to the direct sampling based multiple-point statistics method. We present a methodology for integrating secondary soft data in that framwork. The algorithm is based on direct pattern search through an ensemble of realizations. We show that the proposed methodology is sutiable for modeling complex channelized reservoirs and reduces the uncertainty associated with production performance due to integration of secondary data.

In chapter 4, we present the permanence of ratio hypothesis for data integration in great detail. We present analytical equations for calculating the redundancy factor for discrete or

continuous variable modeling. Then, we show how this factor can be infered using available data for different scenarios.

In chapter 5, we present a case study where permanence of ratio is used for modeling a carbonate reservoir in the Gulf of Mexico. We show that the method has a better performance than when primary hard and secondary soft data are used within the traditional geostatistical framework.

We conclude the thesis in chapter 6 with a review of key research conclusions and ideas for future research.

Chapter 2: Literature Review

In this research, we focus on extending the capability of existing tools for data integration. The primary source of information for reservoir modeling is borehole measurements that are usually too sparse to yield to a reliable model if used solely. The secondary source of information in many cases is seismic data, which is generally available exhaustively in the reservoir. This data can be static (3D seismic attributes) or may yield information about the dynamic variability of attributes such as fluid saturations (4-D seismic).

In terms of modifying the current approaches to data integration, two areas are focused in this research: 1- Using multiple point statistics algorithm and using a pattern search algorithm for data integration. 2- Focusing on the correlation between the seismic data and reservoir attributes and accounting for the redundancy between these sources of information using the tau model. This chapter provides a literature review on these two important aspects of the data integration problem.

2.1 GEOSTATISTICAL MODELING

As previously mentioned, multiple-point statistics based reservoir modeling methods utilize conditioning data and a training image in order to synthesize a model for the target reservoir. Over the past two decades, several multi-point statistical methods have been developed for application in the geosciences.

Kriging is one of the earliest of the geostatistics methods, developed by Matheron (1976). The technique generates a spatial map by interpolating sparse data. The interpolation is exact and accounts for the proximity of the data, the redundancy between the data and the magnitude of the data values. Indicator kriging modifies the original algorithm to utilize a set of binary data which

may be obtained by specifying a threshold in order to classify the original data (Journel A. , 1986). The indicator variables are defined as:

$$I(z_c; \boldsymbol{x_\alpha}) = \begin{cases} 1, & if\ z(x_\alpha) \le z_c \\ 0, & otherwise. \end{cases} \quad (2\text{-}1)$$

z_c is the threshold, $z(\boldsymbol{x_\alpha})$ are the data values at various spatial locations $\boldsymbol{x_\alpha}$. It calculates the conditional cumulative distribution function (ccdf) at each location estimating the probability of a given outcome at each location. The expected value of the indicator is the cumulative probability function of the variable at the threshold value at that location, i.e.:

$$E\{I(z_c; \boldsymbol{x_\alpha})\} = Prob\{z(x_\alpha) \le z_c\} = F_Z(z_c) \quad (2\text{-}2)$$

The expected value of the indicator conditioned to the indicator data in the vicinity of the estimation location is estimated using a linear estimator:

$$E\{I(z_c; \boldsymbol{x_\alpha})\} = F_Z(z_c) = \sum_{i=1}^{n} \lambda_i(\boldsymbol{x_\alpha}) I(z_c; \boldsymbol{x_i}) \quad (2\text{-}3)$$

Where $\lambda_i(x_\alpha)$s are determined by the following system of equations:

$$\begin{cases} \sum_{i=1}^{n} \lambda_i(\boldsymbol{x_i}) C_I(z_c; \boldsymbol{x_i} - \boldsymbol{x_\beta}) + \mu = C_I(z_c; \boldsymbol{x_\alpha} - \boldsymbol{x_\beta}), \qquad \beta = 1, \dots, n \\ \sum_{i=1}^{n} \lambda_i(\boldsymbol{x_i}) = 1, \end{cases} \quad (2\text{-}4)$$

Where $C_I(z_c; \boldsymbol{x_i} - \boldsymbol{x_\beta})$ is the covariance of indicator variables at a pair of locations x_i and x_β evaluated corresponding to the threshold z_c .

First inspired by Journel and Alabert (1990), sequential indicator simulation (*sisim*) builds upon indicator kriging and introduces an algorithm for spatial simulation based on sequential sampling from the conditional distributions obtained by kriging. The process begins by defining a path through all unknown locations. Then, known data points in the vicinity of a

simulation node are retrieved. Indicator kriging is performed at the simulation location to generate a conditional probability density function (cpdf) for the variable. A value is drawn from the cpdf and assigned to the simulation node. The process is repeated at each location along the path until the sparse map is filled. Although the *sisim* algorithm doesn't assume a normal distribution at each point and evaluates the empirical conditional distribution at each point, it is still based on two-point statistics and fails to reproduce complex geologic features.

Several algorithms for integrating different types of data within the kriging framework have also been developed There are several ways to integrate the seismic data in reservoir models such as kriging with an external drift (Goovaerts, 1997) or more robustly cokriging (Dubrule O. , 2003). Cokriging is an extension of kriging in which different types of data are linearly weighted in order to arrive at the estimate. The expected value in cokriging is estimated as follows:

$$z^*(\boldsymbol{u}) = \sum_{i=1}^{n} \lambda_{i0} z(\boldsymbol{u_i}) + \sum_{i=1}^{K} \lambda_{i1} y(\boldsymbol{u_i}) \tag{2-5}$$

Where y is the secondary variable, z is the primary variable, n is the number of primary conditioning data, K is the number of secondary conditioning data, and the coefficients are determined by solving the following system of equations;

$$\begin{cases} \sum_{i=1}^{n} \lambda_{i0} c_{11}(\boldsymbol{u_i} - \boldsymbol{u_j}) + \sum_{i=1}^{K} \lambda_{i1} c_{12}(\boldsymbol{u_i} - \boldsymbol{u_j}) + \mu_1 = c_{11}(\boldsymbol{u} - \boldsymbol{u_j}) & ,j = 1, \dots, n \\ \sum_{i=1}^{n} \lambda_{i0} c_{12}(\boldsymbol{u_i} - \boldsymbol{u_j}) + \sum_{i=1}^{K} \lambda_{i1} c_{22}(\boldsymbol{u_i} - \boldsymbol{u_j}) + \mu_2 = c_{12}(\boldsymbol{u} - \boldsymbol{u_j}) & ,j = 1, \dots, K \\ \sum_{i=1}^{n} \lambda_{i0} = 1 \\ \sum_{i=1}^{K} \lambda_{i1} = 0 \end{cases} \tag{2-6}$$

$c_{11}, c_{12},$ and c_{22} are autocovariace of primary variable, cross-covariance, and autocovariance of the secondary variable respectively.

The spatial variability of each type of data is considered through the appropriate autocovariance of the variable, while the redundancy between the data is considered through the

cross-covariance between pairs of variables. One of the major tasks in cokriging is to ensure that the matrix of auto and cross-covariances is positive definite in order to ensure a unique solution using co-kriging. One of the approaches to ensure positive-definiteness is to assume a linear model of corregionalization (Goovaerts, 1997) which prescribes that the matrices of sill contributions of various structures making up the auto and cross-covariances are positive definite. Almeida and Journel (1994) proposed a simpler Markov corregionalization model which reduces the task of modeling cross-covariances to the rescaling of primary or secondary variable covariance or correlogram as follows:

If we assume:

$$E\{y(\boldsymbol{u})|z(\boldsymbol{u}) = z, z(\boldsymbol{u}+\boldsymbol{h}) = z'\} = E\{y(\boldsymbol{u})|z(\boldsymbol{u}) = z\} \quad (2\text{-}7)$$

In other words, only the collocated hard and soft data at location u is needed for the modeling and the data at a neighboring location $z(\mathbf{u}+\mathbf{h})$ is screened by the datum $z(\mathbf{u})$. The cross-covariance function can then be calculated as:

$$c_{12}(\boldsymbol{h}) = \rho_0 c_{11}(\boldsymbol{h}) \quad (2\text{-}8)$$

ρ_0 is the correlation coefficient of collocated hard and soft data.

A spatial interpolation algorithm such as kriging or co-kriging yields a unique map where the local estimate at every location is identified with the conditional mean at that location. One of the deficiencies of these models is that they do not provide an assessment of global uncertainty. In addition, the maps obtained from kriging are smooth and may not represent the spatial variability of the data. These issues are addressed in stochastic simulation where multiple possible realizations of the random function model are obtained, each honoring the same set of data constraints (Goovaerts, 1997). However, stochastic simulation algorithms based on kriging or co-kriging still require legitimate models for the auto and cross-covariances and suffer from

the same drawback in that the models of spatial heterogeneity are constrained only to spatial two–point covariances or variograms.

2.2 SEISMIC INTEGRATION TECHNIQUES

Current seismic integration techniques focus on developing extensive workflows that utilize large amounts of qualitative geologic information to guide facies interpretation (Lindseth, 1979; Riddiford & Goupillot, 1994; Doyen P. , 2007). Other approaches invoke well and production data for constraining the inversion of seismic attributes (Russell et al., 2001; Hampson et al., 2001; Andersen et al., 2006). These methods have high accuracy, but suffer from high computational time/cost and computer memory. Therefore, several research studies have focused on stochastic integration of seismic data into reservoir models (Haas and Durbule, 1994; Durbule, 2003; Doyen, 1998; Caers and Srinivasan, 2002; Eidsvik et al., 2004; John et al., 2008; Zhang et al., 2012; Srinivasan and Sen, 2009; Srinivasan and Sen, 2010). The main idea of these methods is to statistically correlate the seismic attributes to reservoir parameters, and relate the spatial continuity of the reservoir parameters to the spatial correlation of seismic data.

Caers and Srinivasan (2002) present a method for merging the probability of being in each facies given seismic data with the probability computed on the basis of well data and prior geologic model using the permanence of ratio hypothesis (Journel, 2002). They assumed conditional independence between the seismic attributes and well log measurements. The details of this hypothesis and potential usage in the field of data integration will be discussed thoroughly in the subsequent chapters.

John et al. (2008) calculated the probability of different facies given seismic attribute using a mixture of Gaussian models. They assumed that the seismic attributes follow a normal distribution within each rock type. They subsequently performed an optimization algorithm to fit

the final multi-modal normal distribution to the experimental distribution. Then, they used Bayes' theorem to calculate the final facies probabilities at each location.

Srinivasan and Sen (2009) generated the probability map corresponding to different lithofacies in a carbonate reservoir in the Gulf of Mexico. They combined the facies probabilities based on well data computed using the indicator simulation method with the probability computed by performing multivariate analysis of seismic data using permanence of ratio hypothesis (Journel A. , 2002). They also implemented this methodology for mapping diagenesis in a carbonate reservoir in the Gulf of Mexico (Srinivasan and Sen, 2010). The same methodology was also applied in Calabrese et al. (2011).

Lecante et al. (2013) provided a detailed workflow for the statistical integration of 3D seismic information in order to further constrain a facies model derived from well logs and depositional models. At the core of this workflow is the truncated pluriguassian (Mariethoz et al., 2009) method. In this method multiple Gaussian simulation realizations are truncated using surfaces derived on the basis of the depositional models and later combined in order to create categorical maps. Seismic inversion data is used to compute facies proportions and the Gaussian realizations are constructed using variograms and conditioning data values.

Weltje, et al. (2013) developed a quantitative approach to integrate basin-scale geologic knowledge with locally known reservoir specific data in order to create a final static model of the reservoir. In addition to the traditional data - seismic and well data, they also proposed a novel data integration scheme that takes into account sequence stratigraphy information. The technique uses a method similar to that of Lecante et al. (2013) in order to calculate a final output map using calibrated probability density functions. The locations of specific geologic features are not explicitly determined in the static model. Instead, the final volume consists of proportion of a

cell volume occupied by channels. A notable aspect of the approach is the conditioning of the model to several geologic factors such as topography and sediment source.

The methods discussed above mostly rely on traditional geostatistical simulation algorithms i.e. variogram-based algorithms. Variograms are measures of two-point spatial variability. In the indicator context, the variogram represent the probability of transitioning from one category (or rock facies) to another over a certain lag distance. However, a variogram is inadequate to correctly represent the spatial connectivity of complex, curvilinear features such as channels. As a result, simulation algorithms constrained to a variogram model fail to reproduce the complex connectivity of geological features such as channels in fluvial reservoirs. Novel algorithms based on multiple-point statistics have been devised to represent such complex features in reservoir models.

Strebelle (2002) worked on the initial idea regarding multiple point simulation proosed by Guardiano and Srivastava (1994) and developed the first structured multiple point statistics based simulation algorithm - ***snesim*** for simulating categorical variables. His technique was based on scanning an analog model for a reservoir i.e. a Training Image (TI) to infer the multiple point statistics representing the spatial continuity of the reservoir. The method has been subsequently extended for integrating seismic data (Castro et al., 2006; Caers et al., 2006; Strebelle et al., 2003). The studies combine probability maps created during seismic inversion (Andersen et al., 2006) with the probability of different facies from *snesim* algorithm assuming conditional independence. The final reservoir model is chosen by history matching to prior production data.

Chugunova and Hu (2008) proposed another algorithm that accounts for disparity in data support while integrating continuous secondary data. They showed that the secondary source of

data can be incorporated in order to reproduce geometric features accurately especially in non-stationary cases.

Despite these extensions to geostatistical procedures, the techniques discussed still rely quite a bit on elaborate workflows customized to an individual user's prior beliefs and preferences. These techniques also required pre-specification of geometric templates for retrieving the pattern of variability exhibited by the attribute of interest. The pattern of variability exhibited by the secondary data such as seismic is unaccounted for in these modeling approaches. Instead, the pattern of variability exhibited only by the primary variable of interest is considered. Also, the conditional dependency between seismic data and reservoir parameters are ignored within the data integration schemes implemented using the permanence of ratio hypothesis.

2.3 MULTIPLE-POINT STATISTICAL MODELING

As mentioned previously, Strebelle (2002) proposed a multiple point simulation algorithm *snesim* that utilizes a training image to develop spatial patterns, Single normal equation simulation (*snesim*) begins by moving a template over each pixel of the training image sequentially and storing the pattern of variability exhibited by the nodes of the template, resulting in a database of all of the pattern configurations for a given template definition. The simulation then proceeds sequentially over the un-informed nodes of a simulation grid by acquiring conditioning data from the sparse grid, finding the pattern matches from the database, and then performing Monte Carlo sampling to assign the value at the central location of the template to the simulation node. The creation of the database is intended to provide rapid recall of a pattern during the simulation process.

The major limitations of *snesim* are the inability to maintain the connectivity of large-scale features and the inability to process continuous data. In addition, the simulation procedure requires substantial computer storage for complex training images. Pattern reproduction is limited to the size of the template, and increasing the size of the template greatly increases computing costs. Efforts to streamline the data storage in snesim and to speed up the pattern recall process are currently underway (Straubhaar et al., 2013).

Growthsim (Eskandari and Srinivasan, 2007; Huang and Srinivasan, 2012) introduced the notion of simulating multiple point simulation event conditioned to multiple point data event in the vicinity of the simulation node. This method is in contrast to traditional multiple point statistics algorithms where the simulation progresses one node at a time. Huang and Srinivasan (2012) demonstrated growthsim algorithm for developing the reservoir model for a deepwater turbidite system. They also showed the capability of growthsim algorithm to represent non-stationary features.

Arpat and Caers (2005) proposed *simpat*, a simulation algorithm that aims at pattern reproduction instead of statistics reproduction while honoring conditioning data. This algorithm relies on several image processing concepts, such as image similarity, to borrow and reproduce patterns from training images constrained to hard and soft data. The method makes use of a new multiple-grid approach by which the scale relations between the training image patterns are better captured and reproduced.

Zhang et al (2006) introduced a multi-point algorithm that operated similar to *snesim*, but pasted entire patterns from the training image into the sparse map. Rather than simulating one grid location at a time, *filtersim* scores patterns from the training image, places the pattern into a

binned database, retrieves the pattern consistent with the pattern of conditioning data during simulation, and places the entire pattern onto the simulation grid.

In all these techniques, multiple point statistics describing the patterns observed in a training image are obtained by scanning the image using a spatial template. During simulation, the data base of pattern statistics is searched for the matching pattern. The direct sampling algorithm by Mariethoz et al. (2010) eliminates the expensive step of first scanning and saving the pattern statistics and instead scans the training image directly during simulation and the first match to a conditioning data pattern is retrieved and applied at the simulation node.

In all the available algorithms, the possibility of integrating a secondary source of data is lacking. The methods are highly reliable on the availability of training images, and there is no way for assessing if the training image is accurately compatible with the underlying geologic formation. As mentioned earlier, some methods have been proposed for integrating seismic data within multiple point statistics framework, but they all assume a conditional independency between the seismic data and primary data. In chapter 4, we will show how this assumption will affect the data integration process.

In the next chapter, this direct sampling algorithm is discussed fully and a pattern search based algorithm for incorporating secondary soft data such as seismic impedance data within this framework is presented.

Chapter 3: Direct Pattern Search Based Seismic Integration.

In this chapter, we present the details of the direct sampling algorithm for multiple point simulation and present an algorithm for seismic data integration in that framework. The Direct Sampling (DS) method was proposed by Mariethoz et al. (2010) as a multi-point geostatistical simulation algorithm that populates a sparse data field conditioned to sparse field data using patterns seen in a fully realized training image. The method differentiates itself from other multiple point simulation algorithms due to its very minimal use of computer memory storage and its potential application to continuous variables.

3.1 DEVELOPMENT OF DIRECT SAMPLING

Direct sampling was initially developed to model heterogeneity in hydrogeologic systems and, shared some features of the multiple point statistics methods such as *snesim and filtersim.* The direct sampling process differentiates itself from the previously introduced algorithms mainly by skipping the entire process of scanning and storing pattern statistics from a training image. Instead, it simulates the outcome at the simulation node by directly finding the first match in the training image, rather than by retrieving a match from the database of prior patterns. The idea is borrowed from the early work of Claude Shannon, more specifically his work on replicating English text using a Markov chain (Shannon, 1948). He demonstrated the ability to mathematically approximate English word and sentence by matching a conditioning string pattern to that in a dictionary or a training text.

In the direct pattern sampling technique by Mariethoz (2010), the outcome at a central node given the pattern of information in the surrounding nodes is directly sampled from the training image. Instead of counting and storing the pattern histogram by scanning the training

image and subsequently retrieving the probability corresponding to a conditioning data pattern during simulation, in the direct sampling algorithm, the training image is scanned at every step of the simulation. The first match with the conditioning data pattern is sampled and the outcome at the central node is directly copied and applied at the simulation node.

The goal of direct sampling method is to simulate a random function $Z(\mathbf{x})$ where the nodes in the simulation grid are denoted as $\mathbf{x}$. In the following, the training image is assumed to have information at nodes denoted as $\mathbf{y}$. In our case, the conditioning data for simulation are borehole measurements. The algorithm proceeds as follows:

1. Assign each conditioning data to the closest grid node on the simulation grid. If there are more than one conditioning data in a grid cell, the one closest to the center of the grid is chosen.
2. Define a path through all other nodes in the simulation grid. This would be a vector containing the indices of the nodes. The path can be random (Strebelle S. , 2002), unilateral (Daly, 2004) or any other structured path.
3. Find the next simulation location $\mathbf{x}$ in the path
4. Find the neighbors of $\mathbf{x}$ at which conditioning data values are available. These could be the original conditioning data or nodes where simulation has already been performed and the value has been assigned. It would consist of maximum number of n nodes (user specified) $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$. If there is no neighbors (in the first step of unconditional simulation) randomly choose a location $\mathbf{y}$ from training image and assign $z(\mathbf{y})$ to $z(\mathbf{x})$.
5. Computing the lag vector $L = \{\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_n\}$ separating the simulation node $\mathbf{x}$ from the remaining n conditioning locations.

6. Define the data event $d_n(\mathbf{x}, \mathbf{L}) = \{z(\mathbf{x} + \mathbf{h_1}, \mathbf{x} + \mathbf{h_2}, \ldots \mathbf{x} + \mathbf{h_n}\}$. It is a vector containing the data values at the neighboring nodes.
7. Randomly choose a node y in the training image and check if the pattern $d_n(\mathbf{y}, \mathbf{L}) = \{z(\mathbf{y} + \mathbf{h_1}, \mathbf{y} + \mathbf{h_2}, \ldots \mathbf{y} + \mathbf{h_n}\}$ matches the conditioning data pattern $d_n(\mathbf{x}, \mathbf{L})$. If $d_n(\mathbf{x}, \mathbf{L}) = d_n(\mathbf{y}, \mathbf{L})$ assign z(**y**) to z(**x**). Add the node to conditioning data set and return to step 3.
8. If $d_n(\mathbf{x}, \mathbf{L}) \neq d_n(\mathbf{y}, \mathbf{L})$ start systematically scanning the training image. Compute the distance $D\{d_n(\boldsymbol{x}, \boldsymbol{L}), d_n(\boldsymbol{y}, \boldsymbol{L})\}$ between the events in the training image and the conditioning pattern. Different options for the distance based on whether the patterns are in terms of discrete or continuous outcomes are discussed in Mariethoz (2010). Find the closest data event to $d_n(\boldsymbol{x}, \boldsymbol{L})$ i.e.$\arg\min D\{d_n(\boldsymbol{x}, \boldsymbol{L}), d_n(\boldsymbol{y}, \boldsymbol{L})\}$ and then assign z(**y**) = z(**x**).
9. If the simulation grid is completely simulated stop, otherwise return to 3.

These steps are conceptually shown in Figure 3-1. Figure 3-1(a) shows the neighboring conditioning data, and question mark is representative of the node to be simulated. Figure 3-1(b) shows a one step of the scanning process where the training pattern does not match the conditioning data completely. Figure 3-1(c) shows a complete match. Hence, the value is assigned to the simulation node (Figure 3-1(d)).

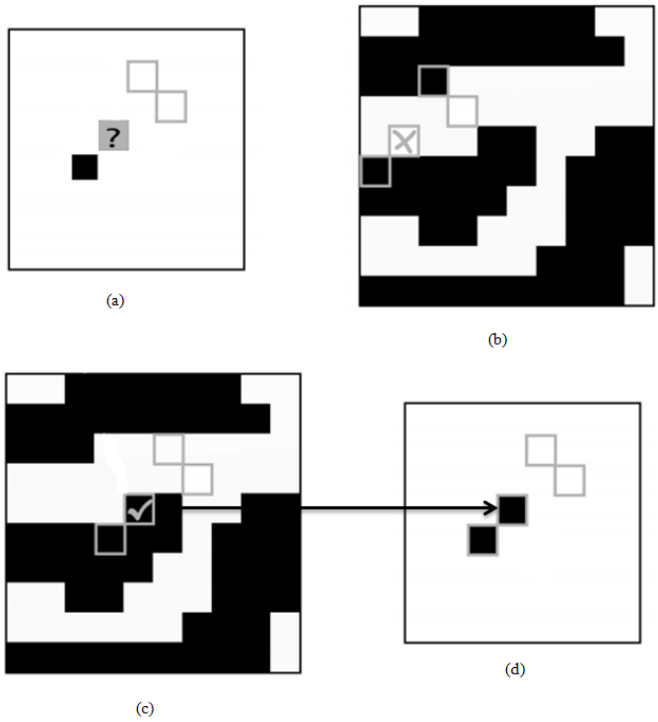


Figure 3-1. Illustration of Direct Sampling method. (a) Defining data event. (b) One of the steps of scanning through training image. (c) The first time the data event template matches the training image completely. (d) Assign the simulated value to the simulation node and add it to conditioning points (Mariethoz, Renard, & Straubhaar, 2010).

The quality of the pattern reproduction in the generated images depends on the specification for the size of the neighborhoods. In general, the continuity of spatial patterns is better reproduced if a large neighborhood is specified. However, specifying a large neighborhood for obtaining the conditioning pattern can add significantly to the computational cost. CPU burden can be alleviated using parallelization. Parallelizing the DS algorithm is straightforward on shared memory machines: each CPU performs the search in a limited portion of the TI (Mariethoz G. , 2010; Mariethoz, Renard, & Straubhaar, 2010).

It should be noted that if the path is random, one will get different results as the random seed is updated. Uncertainty can therefore be assessed by repeating the algorithm several times.

3.2 INTEGRATING SEISMIC DATA

As mentioned earlier, in general the well data are too sparse to make reliable estimation of patterns of variability of reservoir rock properties. Therefore, another source of data that has better areal coverage is required. In most cases seismic data provides exhaustive spatial information about the whole reservoir. However, seismic data is generally of poor spatial resolution and consequently has to be combined with well observations in order to develop reliable models for the reservoir. In the following, a multiple point statistics based algorithm for integrating seismic data is presented and implemented.

Since the seismic impedance is a continuous variable, the first step should be to transform the seismic map to a corresponding categorical map. This transformation is justified, because in the multiple point pattern based method, it is the pattern of variability exhibited at the search nodes rather than the absolute magnitude of the attribute that is important. In order to accomplish this without any user specified cut-offs, a one-dimensional k-means clustering algorithm was implemented with the objective to maximize the following distance measure:

$$d = E(d_k - d_{k'})^2 + \{E(d_i - d_k)^2\}^{-1} \quad 3\text{-}1)$$

d_{ik} is the i^{th} data point classified to be in the k^{th} cluster, and d_k is the corresponding mean of the k^{th} cluster. *k'* is another cluster with mean $d_{k'}$. The objective is therefore to minimize the intra-cluster distance (the second term in RHS) while maximizing the inter-cluster distance (the first term in RHS).

Once the seismic map has been clustered and transformed to a categorical map, a pattern search based algorithm for integrating seismic data could be implemented. The goal of the

algorithm would be to sample the outcome at each node in the simulation grid given the available categorized seismic data and the training image. The inputs will be the simulation grid whose nodes are denoted as **x**, a training image whose node are denoted as **y**, and seismic attribute values $Y(\boldsymbol{x})$ available at each location **x** of the simulation grid. Note that if the seismic data is at a different resolution than the simulation grid, it is assumed that appropriate up or downscaling of the seismic data has been performed.

1. Based on the conditioning data, draw multiple realizations of the simulation grid by running the direct sampling algorithms N times. These realizations constitute a matrix with columns $\mathbf{x}_1$, $\mathbf{x}_2$, … $\mathbf{x}_N$ where each column has *K* rows denoting the number of grid blocks in each model.
2. Define the size of the search neighborhood
3. Define a random path passing through all the nodes in the simulation grid. This would be a vector containing the indices of the nodes. The path can be random (Strebelle S. , 2002), unilateral (Daly, 2004) or any other structured path.
4. Find the next simulation location in **x**.
5. Put the template around the simulation node in such a way that the center of the template would be the simulation node.
6. Find the neighbors of **x** which are within the specified search neighborhood.
7. Computing the lag vector $\mathbf{L} = \{\mathbf{h_1},\mathbf{h_2},\ldots,\mathbf{h_n}\}$
8. Define the data event $d_{ny}(\hat{\mathbf{x}}, \mathbf{L}) = \{y(\hat{\mathbf{x}} + \mathbf{h_1}), y(\hat{\mathbf{x}} + \mathbf{h_2}), \ldots y(\hat{\mathbf{x}} + \mathbf{h_n})\}$. It is a vector containing the categorial seismic value of the nighboring data.
9. Start systematically scanning the seismic map.
10. For each location **u** in the seismic map calculate the seismic data event around that point

11. If the previous template matches $d_n(\hat{\mathbf{x}}, \mathbf{L})$ store the location **u**

12. At each matched seismic location u, scan through the entire suite of N realizations and record the outcome at the location corresponding to the simulation node on the match template. Then calculate the local conditional probability corresponding to the category z_k as:

$$\text{Prob}\left\{Z(\mathbf{x}) = z_k \mid d_{ny}(\hat{x}, L), \text{conditioning data}\right\} = \frac{\text{count}_{\text{N,grid}}(z(\mathbf{u})=z_k)}{\text{count}_{\text{grid}}(\mathbf{u}) \times \text{N}}$$

$\text{count}_{\text{N,grid}}(z(\mathbf{u})=z_k)$: count of matches of category z_k at the node corresponding to the simulation node at the seismic match location **u** over the **N** realizations aggregated over all locations of the seismic match over the **grid**

$\text{count}_{\text{grid}}(\mathbf{u})$: count of locations u where the seismic data pattern matches the pattern at the simulation node

13. If the ccdf is computed at every location on the simulation grid stop, otherwise go to 4.

As can be seen, the algorithm would result in a unique invariant probability map by which one can assess the uncertainty and generate multiple realizations. The advantage of the algorithm is that it honors both the hard observation data as well as soft data. In the next section, the seismic data integration method is implemented and the results are presented.

3.3 RESULTS

A synthetic case study is implemented in order to demonstrate the application of the method. In this case study, the simulation grid size is 50×50. A training image has been

constructed using SGEMS software. As can be seen in Figure 3-2, the training image is that of a channelized reservoir.

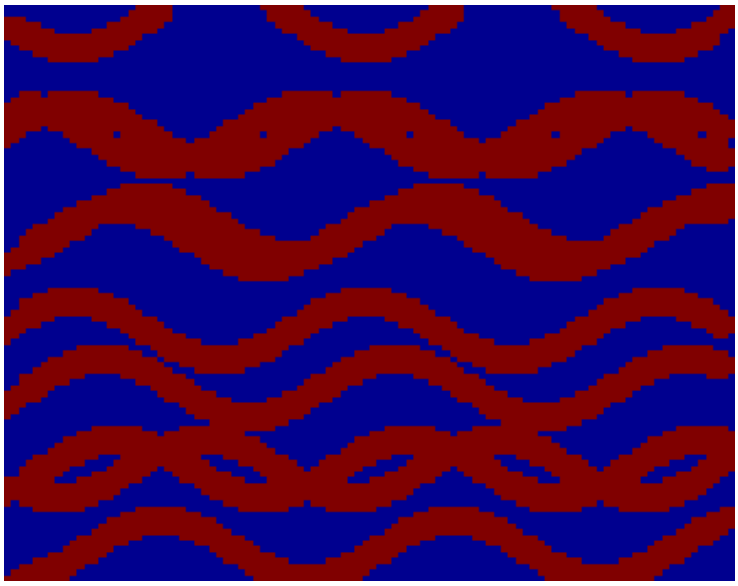
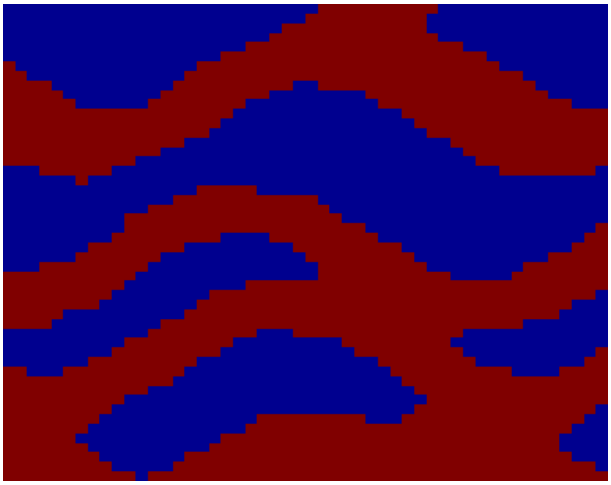
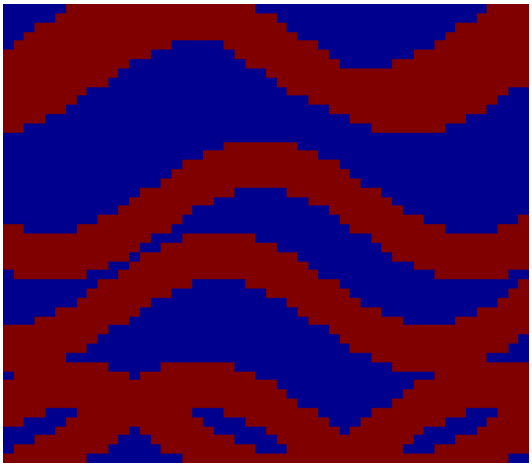


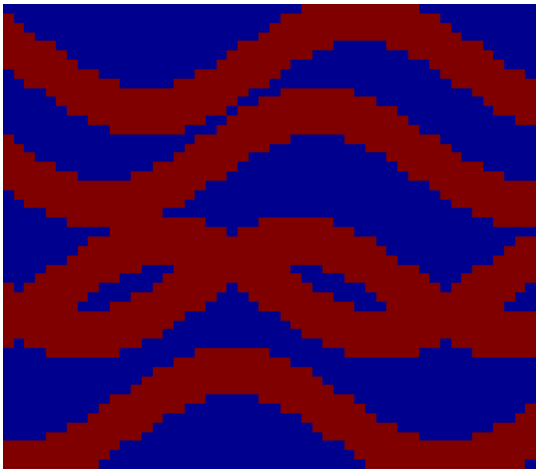
Figure 3-2. The training image of a channelized reservoir generated by SGEMS

51 realizations of the reservoir facies were generated conditioned to five conditioning data randomly sampled from the training image. These realizations were generated applying the direct sampling algorithm described earlier. One of the realizations is considered as a reference model.



(a)





(b) (c)

Figure 3-3. Multiple realization results of Direct Sampling. (a) Reference model 1 as well as two other realizations (b, c)

Figure 3-3 shows reference model as well as two other realizations generated using the direct sampling procedure. The ensemble average of the 51 realizations yields the probability map of each facies using only well data and the training image as shown in Figure 3-4.

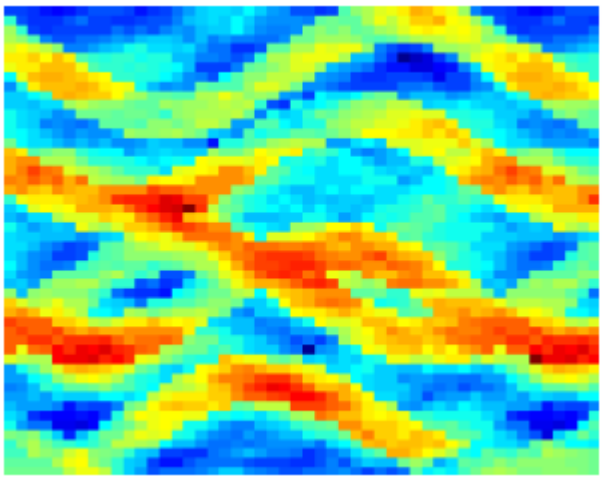


Figure 3-4. The ensemble average of multiple realizations yields the probability of each facies at each location shown in the figure.

After the facies maps are created, a blurring mechanism is used to create synthetic seismic image corresponding to the reference map. This process is similar to forward seismic processing. The blurring algorithm used in this research begins by converting one of the geologic realizations created in the previous section into a corresponding impedance map. In order to accomplish this, a Gaussian filter is applied.. Gaussian filters are commonly used in single dimensional signal processing and two dimensional image editing applications. The Gaussian function in two dimensions, as used in this method, applies weights to all locations surrounding a given point and replaces the original point's value with a weighted average. The filter more heavily weighs locations closer to the central location in a grid according to the parameters for the bi-dimensional Gaussian function given in Eq. (3-2). The main advantages of the Gaussian filter as compared to other convolution matrices is the absence of negative weights

and the heavy weighting of the central location. These attributes allow the blurring mechanism to keep the edges and location of structures in the image constant while applying the loss of information evenly. The Gaussian filter function is:

$$g(x,y) = \frac{1}{2\pi \cdot \sigma} \cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (3\text{-}2)$$

where x and y are the distances from the original point and σ is the standard deviation. Subsequently, Gaussian noise representative of measurement noise is added to the results. The resultant Seismic map for the reference map is shown in Figure 3-5.

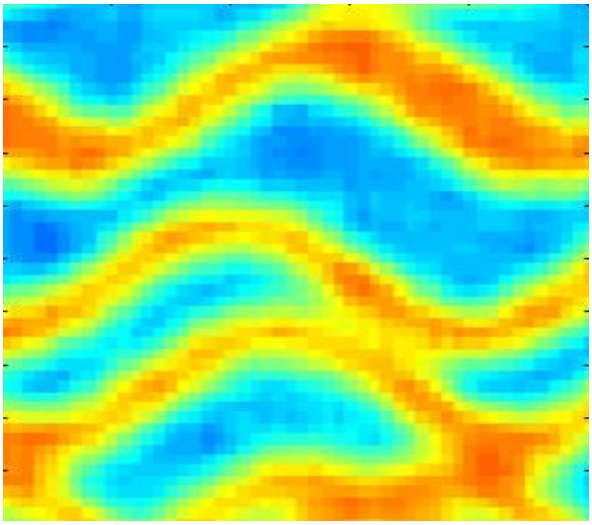


Figure 3-5. Generated synthetic seismic map

Once the seismic map and multiple realizations of direct sampling method are available, the proposed data integration algorithm can be implemented. As explained completely in the previous section, by scanning the seismic map and the ensemble of realizations, the conditional probability map can be constructed. This would be the conditional probability of a particular category to exist at each location given the hard observation data and soft seismic data. The resultant probability map is depicted in Figure 3-6.

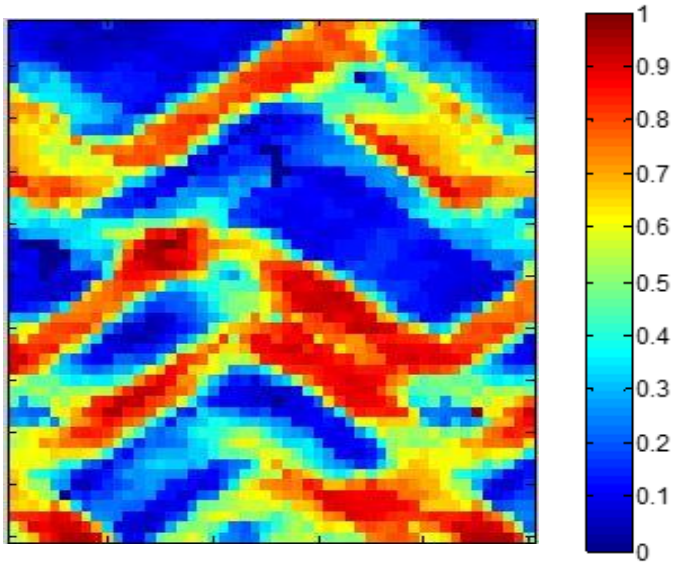
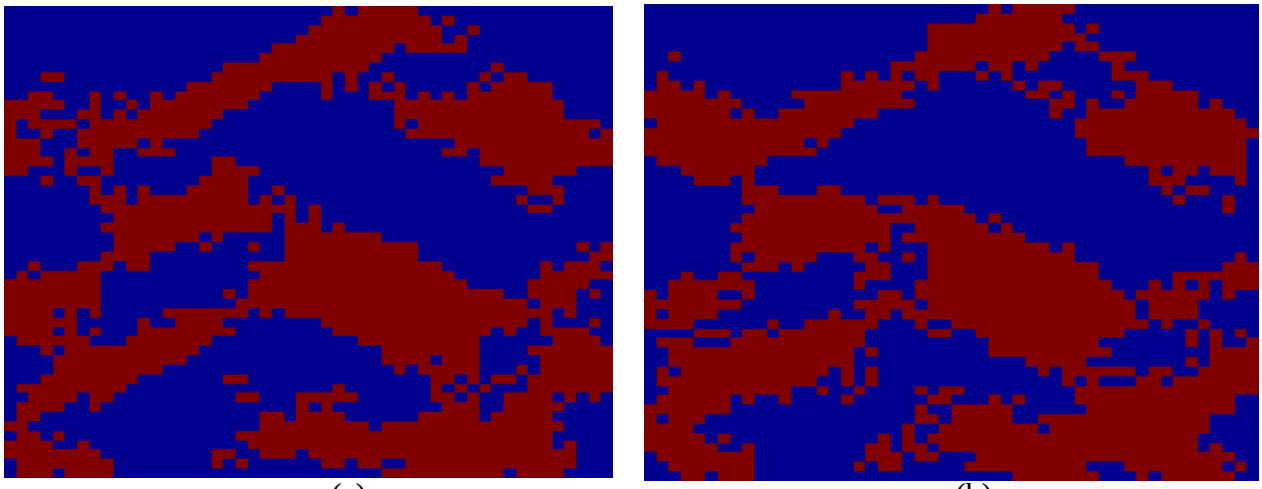


Figure 3-6. Local conditional probability map conditioned to well data and seismic data

As can be seen in the Figure 3-6, the channel shapes can be obviously seen in the probability map. The map shows the characteristics observed in the seismic map but also exhibits some differences. Comparing Figure 3-6 and Figure 3-4 clearly shows the improvement of the accuracy of the results by integrating seismic data into reservoir model. The algorithm yields the probability map corresponding to facies at each location which is a representation of the geologic uncertainty at each location.

Once the probability map is available, we can draw multiple realizations. In order to draw a realization a map of random numbers in required. We generate multiple unconditional realizations of random numbers using the direct sampling algorithm with the probability map (Figure 3-6) and transform the obtained map to uniform distribution between 0 and 1. The random seed map is then used to generate a realization map with Monte-Carlo sampling technique. This process can be done several times to get multiple realizations. Figure 3-7 shows two realizations that resemble the reference map.



(a) (b)

Figure 3-7. Multiple realizations sampled from the final probability map

3.4 CONCLUSION

In this chapter, we presented a new algorithm which is based on multiple point algorithms and integrates both primary and secondary data. The algorithm is based on a direct pattern search through the training image and seismic map that yields the conditional probability at each location given well measurements and seismic data. The implementation of the algorithm reveals that the integration of secondary soft data improves the accuracy of the final ensemble average map and reduces the uncertainty. It was shown that the method is suitable for complex structures such as channelized reservoirs. In the next chapter, we will present the permanence of ratio hypothesis in details and discuss an alternative approach to integrate secondary data into reservoir models.

Chapter 4: Permanence of Ratio Hypothesis

In this chapter, we present another method for data integration which is based on the permanence of ratio hypothesis. The permanence of ratio hypothesis is a probabilistic scheme to merge information coming from different sources. The aim of this process is to calculate the conditional probability distribution $\mathrm{P(A|B,C)}$ when the conditional probabilities $\mathrm{P(A|B)}$ and $\mathrm{P(A|C)}$ are known. In earth science problems, A usually stands for the rock type or property being modeled, B represents the prior "hard" information (generally in the same units as A) available to model and C is the secondary information such as seismic data that may be indirectly related to attribute being modeled. The permanence of ratio hypothesis is predicated on the calibration of information from each source of data and then merging the information accounting for the redundancy between data. We explore this important paradigm for data integration in this chapter

4.1 PERMANENCE OF RATIO HYPOTHESIS

In order to model the joint conditional probability obtained by integrating information from two different sources of information, it would be convenient if the information from each data source can be assessed independently in order to find $\mathrm{P(A|B)}$ and $\mathrm{P(A|C)}$, and then merge these joint probabilities to calculate $\mathrm{P(A|B,C)}$ accounting for the redundancy between different data sources. This is precisely the premise of the permanence of ratio hypothesis.

The joint probility of $P(A|B,C)$ can be calculated from following equation:

$$P(A|B,C) = \frac{P(A,B,C)}{P(B,C)} \tag{4-1}$$

With

$$P(A,B,C) = P(A)P(B|A)P(C|A,B) = P(A)P(C|A)P(B|A,C) \quad (4\text{-}2)$$

In order to simplify the problem, one can assume that the two data sources are independent, i.e.

$$P(B,C) = P(B)P(C) \quad (4\text{-}3)$$

$$P(C|A,B) = P(C|A), \;\; P(B|A,C) = P(B) \quad (4\text{-}4)$$

Equation (4-1) then simplifies to:

$$P(A|B,C) = P(A).\frac{P(B|A)}{P(B)}.\frac{P(C|A)}{P(C)} \quad (4\text{-}5)$$

i.e.:

$$P(A|B,C) = P(A|B).\frac{P(A|C)}{P(A)} \quad (4\text{-}6)$$

The simplicity of Equation (4-8) is due to the assumption of full independence of data events and leads to some inconsistencies as described below:

1- $P(A|B,C)$ is not guaranteed to be between 0 and 1. This will happen if the conditional probabilities *P(A|B)* and *P(A|C)* are evaluated independently and they are inconsisent with the assumption of independence between *B* and *C*. Consider the case where *P(A|B)=0.6*, *P(A|C) = 0.4*, and *P(A) =0.2*. Then, Equation **Error! Reference source not ound.** leads to *P(A|B,C) = 1.2*,and that cannot happen.

2- If C is fully informative of A i.e *P(A|C) = 1* or *0*, *P(A|B,C) ≠ P(A|C).* This means that by the assumption of full independence between data events, the resultant data integration scheme implies uncertainty even if one of the data sources is rendering the prediction deterministic**.**

Now, if we assume that there is conditional independence between event *B* and event *C*,i.e. we assume that in the presence of a common conditiong event (such as the existence of a reservoir property field A), we can write the following

$$P(C|A,B) = P(C|A) \quad (4\text{-}7)$$

$$P(B|A,C) = P(B|A) \quad (4\text{-}8)$$

Plugging thesein equation (4-2) will lead to the following equation:

$$P(A|B,C) = \frac{P(A)P(B|A)P(C|A)}{P(B,C)} \quad (4\text{-}9)$$

Now consider the complement event $\bar{A}$, such that $P(A)+P(\bar{A})=1$ Doing the same developmens for $P(\bar{A}|B,C)$ will yield:

$$P(\bar{A}|B,C) = \frac{P(\bar{A})P(B|\bar{A})P(C|\bar{A})}{P(B,C)} \quad (4\text{-}10)$$

Thus:

$$\frac{P(A|B,C)}{P(\bar{A}|B,C)} = \frac{P(A)P(B|A)P(C|A)}{P(\bar{A})P(B|\bar{A})P(C|\bar{A})} \quad (4\text{-}11)$$

Recall that:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (4\text{-}12)$$

Hence the above equation can be written as:

$$\frac{P(A|B,C)}{P(\bar{A}|B,C)} = \frac{\left(\frac{P(A|B)}{P(\bar{A}|B)}\right)\left(\frac{P(A|C)}{P(\bar{A}|C)}\right)}{\frac{P(A)}{P(\bar{A})}} \quad (4\text{-}13)$$

Defining the following odds ratios:

$$x = \frac{1-P(A|B,C)}{P(A|B,C)} \quad , \quad a = \frac{1-P(A)}{P(A)} \; , \; b = \frac{1-P(A|B)}{P(A|B)} \; , \; c = \frac{1-P(A|C)}{P(A|C)} \quad (4\text{-}14)$$

We can write Equation 4-14 as:

$$\frac{x}{a} = (\frac{b}{a})(\frac{c}{a}) \quad (4\text{-}15)$$

The odds ratio measure the relative distance to the occurrence of A given each data source e.g. B. If A is guaranteed not to occur given the data, then the relative distance to

occurrence of A is infinity, while if A is guaranteed to occur because of the occurrence of the data source, then the relative distance is zero. Because the odds ratio *x* is in terms of the required conditional probability *P(A|B,C)*, the permanence of ratio hypothesis provides an approach to merge the conditional probabilities from individual sources of information in order to obtain *P(A|B,C)*.

4.2 MATHEMATICAL PROPERTIES

In this section, we investigate some special cases and present the results obtained using this form of the permanence of ratio hypothesis.

4.2.1 Limit Properties

Since all the odd ratios in Equation (4-14) are positive, the resultant odd ratio remains positive, thus:

$$x \geq 0 \Rightarrow 0 \leq P(A|B,C) = \frac{1}{1+x} \leq 1 \tag{4-16}$$

4.2.2 Completely informative data

This is the case when one source of information guarantees the occurrence of an event A, say for example:

$$P(A|B) = 1 \tag{4-17}$$

Therefore

$$b = \frac{1 - P(A|B)}{P(A|B)} = 0 \tag{4-18}$$

Plugging Eq (4-18) into Eq (4-25) leads to:

$$x = b\,(\frac{c}{a}) = 0 \tag{4-19}$$

that leads to

$$P(A|B,C) = \frac{1}{x+1} = 1 \tag{4-20}$$

Therefore, if one of the data sources guarantees occurrence of a simulation event A, integrating a secondary source of information will not introduce any uncertainty in the occurrence of that event and the merged probability also guarantees the occurrence of that event. On the other hand, in case a source of information guarantees that the simulation event does not occur, then:

$$P(A|B) = 0 \tag{4-21}$$

Therefore

$$b = \frac{1 - P(A|B)}{P(A|B)} = \infty \tag{4-22}$$

Plugging Eq (4-22) into Eq (4-25) leads to:

$$x = b\,(\frac{c}{a}) = \infty \tag{4-23}$$

Which leads to

$$P(A|B,C) = \frac{1}{x+1} = 0 \tag{4-24}$$

Therefore, if one of the data sources guarantees that the simulation event does not occur, integrating a secondary source of information will not introduce any uncertainty in the occurrence of that event and the merged probability also guarantees that the event does not occur.

4.3 THE TAU MODEL

It should be noted that the above discussion of the permanence of ratio model assumes conditional independence between events *B* and *C* given the event *A*. One way to reintroduce dependence between the data events is to set the contribution $\frac{c}{a}$ to a power which is itself a function of both *B* and *C*

$$\frac{x}{b} = (\frac{c}{a})^{\tau(B,C)} \tag{4-25}$$

It should be considered that when $\tau = 0$ the data source *C* is completely ignored. If $\tau > 1$ the influence of *C* is increased, and if $\tau < 1$ the influence of *C* is decreased.

As pointed out in Journel (2002), data dependence is different from data redundancy. The elementary data sources B, C could be related to each other and that is indicative of their dependency. However, the concept of data redundancy comes from how much information inferred about event A from data source B overlaps with the information coming from data source C. Traditional data integration techniques address the concept of data redundancy through measures such as data-to-data covariance and mutual information (McEliece, 2002). In order to account for redundancy between data sources, tau model was presented (Krishnan s. , 2008). It is shown that it can accommodate the complex description of data redundancy and that any conditional probability $\mathrm{P}(\mathrm{A}|\mathrm{B}, \mathrm{C})$ can be decomposed exactly into a function of the elementary conditional probabilities $\mathrm{P}(\mathrm{A}|\mathrm{B})$ and $\mathrm{P}(\mathrm{A}|\mathrm{C})$. Rewriting Equation (4-11) for $\bar{A}$ results in:

$$\frac{P(A|B,C)}{P(\bar{A}|B,C)} = \frac{P(A)P(B|A)P(C|A,B)}{P(\bar{A})P(B|\bar{A})P(C|\bar{A},B)} \tag{4-26}$$

Invoking Baye's rule leads to:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}, P(B|\bar{A}) = \frac{P(\bar{A}|B)P(B)}{P(\bar{A})} \quad (4\text{-}27)$$

Hence the above equation can be written as:

Implementing Equation (4-27) in Equation (4-32) and using the definition of odd ratios yields:

$$\frac{x}{b} = \frac{P(C|\bar{A},B)}{P(C|A,B)} \quad (4\text{-}28)$$

Comparing Equation (4-28) with Equation (4-25) results in:

$$\tau = \frac{log(\frac{P(C|\bar{A},B)}{P(C|A,B)})}{log(\frac{c}{a})} \quad (4\text{-}29)$$

Recall that:

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)}, P(C|\bar{A}) = \frac{P(\bar{A}|C)P(C)}{P(\bar{A})} \quad (4\text{-}30)$$

Therefore:

$$\frac{P(C|\bar{A})}{P(C|A)} = \frac{P(A)P(\bar{A}|C)}{P(\bar{A})P(A|C)} = \frac{c}{a} \quad (4\text{-}31)$$

Plugging Equation (4-31) in (4-29) leads to:

$$\tau = \frac{log(\frac{P(C|\bar{A},B)}{P(C|A,B)})}{log(\frac{P(C|\bar{A})}{P(C|A)})} \quad (4\text{-}32)$$

In the next section, the sensitivity of the conditional probability to the Tau parameter is explored.

4.4 SENSITIVITY STUDY

In this section, we present an example from Journel (2002) to show the advantage of the permanence of ratio hypothesis to other integration approaches. We also show the sensitivity of the results to the redundancy factor (τ).

Consider the estimation of the probability of occurrence of an event A conditioned to the data events B, C and D. The three data events B, C, and D are assumed dependent on each other according to the joint probability distribution defined in Table 4-1. It should be noted that events B and D are not fully independent, indeed: although *P(B|D) = P(B)*, when C is known: *P(B|C,D)≠P(B|C)*.

The event A is fully determined by the three data event B, C, and D as:

$$A = BC + BCD \tag{4-33}$$

Since A is fully determined by B, C, and D:

$$P(A = bc + bcd | B = b, C = c, D = d) = 1, \qquad \forall\, b, c, d \tag{4-34}$$

Table 4-1. Distribution of data events (Journel A. , 2002)

B	C	D	A	P(B,C,D)
1	1	1	2	0.175
1	1	0	1	0.175
1	0	1	0	0.15
1	0	0	0	0
0	1	1	0	0.125
0	1	0	0	0.125
0	0	1	0	0.2
0	0	0	0	0.05

We used the distribution table to calculate *P(A|B)* and *P(A|C).* The values are shown in Table 4-2.

Table 4-2. Conditional distribution of *P(A|B)* and *P(A|C).* (Journel A. , 2002)

B	C	D	A	P(A|B)	P(A|C)
1	1	1	2	0.35	0.292
1	1	0	1	0.35	0.292
1	0	1	0	0.3	1
1	0	0	0		
0	1	1	0		0.417
0	1	0	0	1	
0	0	1	0		1
0	0	0	0		

We, then, used the values in Table 4-2 to calculate *P(A|B,C)* assuming fully independence, conditional independence, and different values of redundancy factor. The results are shown in Table 4-3.

Table 4-3. Estimation of A from B and C

B	C	D	A	P(A|B,C) Full independence	P(A|B,C) Conditional independence	P(A|B,C) $\tau = 3$	P(A|B,C) $\tau = \infty$	P(A|B,C) Exact decomposition
1	1	1	2	0.5833	0.4392	0.4926	0.5	0.5
1	1	0	1	0.5833	0.4392	0.4926	0.5	0.5
1	0	1	0	0.4615				
1	0	0	0					
0	1	1	0	0.6410	1	1	1	1
0	1	0	0					
0	0	1	0	**1.5385**				
0	0	0	0					

As can be seen, the full independence assumption results in severe inconsistencies such as probability of greater than 1. Although the assumption of conditional independence is more robust, the results show that the correct interpretation of the redundancy factor leads to a more accurate estimation of conditional probability density function. In the next sections, we present a robust method for calculating the redundancy factor.

4.5 INTERPRETATION OF TAU WEIGHTS

Calculating tau weight is a crucial step for using the permanence of ratio hypothesis. However, a robust methodology for calculating tau has not been proposed yet. In several implementation examples (Caers & Srinivasan, Statistical pattern recognition and geostatistical data integration, 2002; Caers & Srinivasan, Combining geological information with seismic and production data, 2003; Kashib & Srinivasan, 2006; Srinivasan & Sen, Mapping of diagenesis in a carbonate reservoir in the Gulf of Mexico by a stochastic data integration technique, 2010) tau

weight has been assumed to be unity. According to Eq (4-32) τ equal to 1 requires that the ratios $\frac{P(C|\bar{A},B)}{P(C|A,B)}$ and $\frac{P(C|\bar{A})}{P(C|A)}$ be equal to each other. One possibility is:

$$P(C|A,B) = P(C|A) \;\; and \;\; P(C|\bar{A},B) = P(C|\bar{A}) \tag{4-35}$$

This means that the datum C is independent of the previous data B given A, i.e. conditional independence. However, in general, equality of these ratios does not necessarily require conditional independence but only that: $\frac{P(C|\bar{A},B)}{P(C|\bar{A})} = \frac{P(C|A,B)}{P(C|A)}$ = *some constant r.* The value $r = 1$ arises from conditional independence. Any other value of $r \in [0, \infty]$ would also yield to unity tau (Krishnan, Boucher, & Journel, 2004). The interpretation then is that the relative information in data source *C* towards the discrimination of event *A* remains invariant regardless of the event *B.*

Krishnan et al. (2004) proposed the following relationship between the tau parameter and correlation coefficient between data sources:

$$\tau = 1 - \rho^2_{B,C|A} \tag{4-36}$$

Where $\rho^2_{B,C|A}$ is conditional correlation of B and C given A. Intuitively the greater the correlation coefficient, the less informative is data source C, and its weight should decrease. However, the main drawback of Eq. (4-36) is that the tau weight is always less than 1, which means that the sensitivity of datum C to change in A is always diminished by data source B.

4.6 ROBUST CALCULATION OF TAU WEIGHTS

In this section, we present a method for accounting for redundancy between data sources and calculation of tau weights by simple statistical properties inferred from data. In our derivations, we assume A i.e. the property being estimated, is a categorical attribute such as rock type. For instance, A_i could stand for different rock types such as mudstone, grainstone, etc.

We assume that the probability distribution of data *B* and *C* exhibit a multimodal characteristics consistent with the presence of multiple categories of event A. Such multimodal distributions can be modeled employing a mixture of Gaussian hypothesis. The tau weights can then be calculated using the mixture of Gaussian assumption as follows:

$$\tau_i = \frac{log(\frac{P(C|\bar{A}_\iota, B)}{P(C|A_i, B)})}{log(\frac{P(C|\bar{A}_\iota)}{P(C|A_i)})} \tag{4-37}$$

$$P(C|\bar{A}_\iota, B) = \frac{\sum_{j\neq i} P(A_j) P(B, C|A_j)}{\sum_{j\neq i} P(A_j) P(B|A_j)} \tag{4-38}$$

$$P(C|A_i, B) = \frac{P(B, C|A_i)}{P(B|A_i)} \tag{4-39}$$

$$P(C|\bar{A}_\iota) = \frac{\sum_{j\neq i} P(A_j) P(C|A_j)}{1 - P(A_i)} \tag{4-40}$$

Where

$$P(B|A_i) \quad \propto (2\pi)^{-\frac{N_B}{2}} \left|\Sigma_{B_i}\right|^{\frac{-1}{2}} exp(\frac{-1}{2}(B - \mu_{B_i})^t {\Sigma_{B_i}}^{-1}(B - \mu_{B_i})) \tag{4-41}$$

$$P(C|A_i) \quad \propto (2\pi)^{-\frac{N_C}{2}} \left|\Sigma_{C_i}\right|^{\frac{-1}{2}} exp(\frac{-1}{2}(C - \mu_{C_i})^t {\Sigma_{C_i}}^{-1}(C - \mu_{C_i})) \tag{4-42}$$

$$P(B, C|A_i) \propto (2\pi)^{-\frac{N_X}{2}} \left|\Sigma_{X_i}\right|^{\frac{-1}{2}} exp(\frac{-1}{2}(X - \mu_{X_i})^t {\Sigma_{X_i}}^{-1}(X - \mu_{X_i})) \tag{4-43}$$

Where A_i is category i. B is primary data source, N_B is the size of primary data source, Σ_{B_i} and μ_{B_i} are the covariance matrix and mean values of primary data within class i. C is secondary data source, N_c is the size of vector C, Σ_{C_i} and μ_{C_i} are the covariance matrix and mean values of secondary data within class i. $X = \begin{bmatrix} B \\ C \end{bmatrix}$, N_X is the size of vector X, Σ_{X_i} and μ_{X_i} are the covariance matrix and mean values of vector X within class i. By calculating these parameters from data sources, we are able to calculate the tau weights, merge the conditional probabilities, and find $P(A_i|B,C)$ for all categories.

As mentioned above, in earth science problems, B and C usually stand for well measurements and seismic attributes respectively. The conditional probability $P(A_i|B,C)$ can also be computed using co-indicator kriging using seismic data as soft data. Two drawbacks of co-indicator based method in comparison to our proposed method are: 1) cross-covariance modeling in such a way that the covariance matrix would be positive definite is a difficult task to do, and 2) co-indicator algorithms can often result in illegitimate probability values (greater than 1 or less than 0) that have to be corrected for order relations.

It should be noted that the expression (4-37) for calculating tau is data dependent and requires the actual outcomes of B and C data event. However, the outcome of one or both the data events may not be known at every location. In order to overcome this problem the expected value of tau is calculated and can be used:

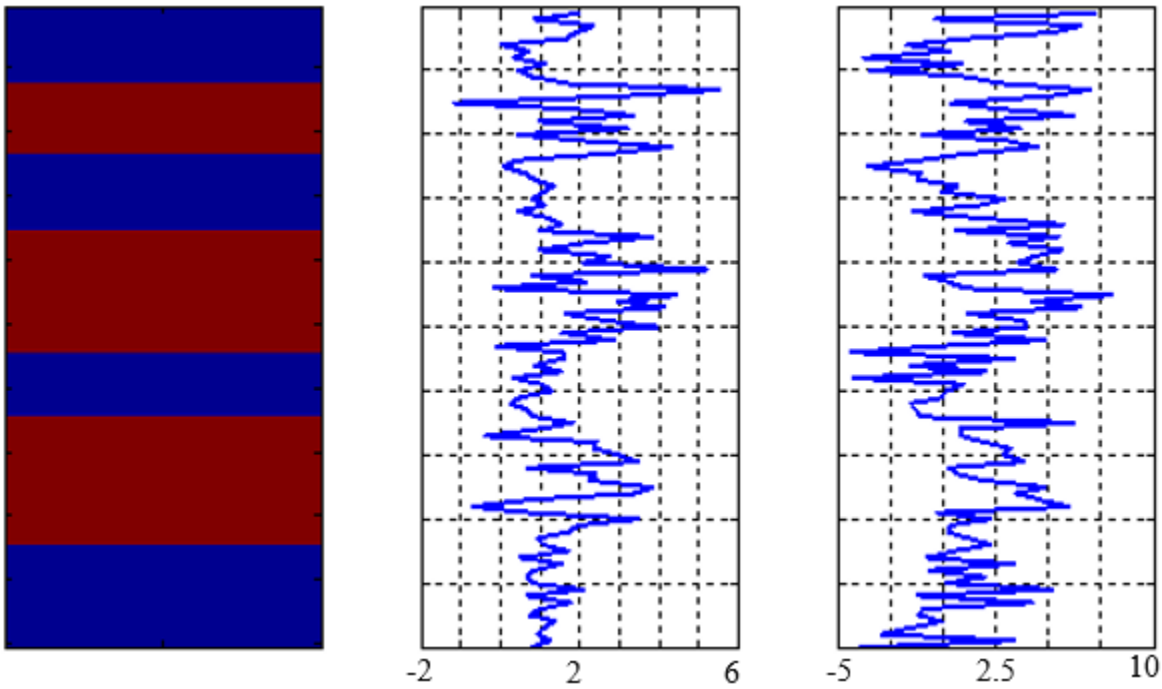
$$\bar{\tau} = \iint \tau(b,c) f_{BC}(b,c) db\, dc \qquad (4\text{-}44)$$

Where $f_{BC}(b,c)$ is the joint pdf of variables B and C. This expected value can be computed by taking various realizations of the data events B and C and computing tau and finally computing the average of all values of tau. It can be proven that the average value of tau can be

related to the mutual information between the data sources and discrimination density expectation and is guaranteed to be positive (Krishnan s. , 2008):

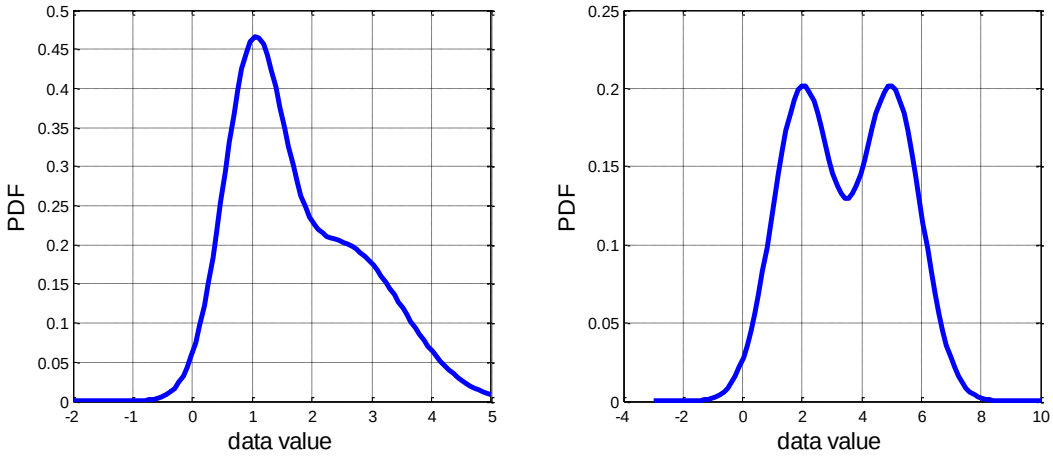
4.7 SYNTHETIC 1-D CASE

In this section, we demonstrate the proposed methodology for a simple 1-D classification problem. We consider a 1-D map of facies along with two different sources of information (see Figure 4-1). This data set could be considered as rock type and two different log attributes along a well. The distribution of each data source is shown in Figure 4-2. As can be seen in Figure 4-2, each data source is showing a bimodal distribution, where each mode corresponds to one type of facies. The scatter plot of these two data sources is shown in Figure 4-3. Figure 4-3 shows the mixture of Gaussian behavior of the data.



(a) (b) (c)

Figure 4-1. (a) The facies map (Hereafter known as **A**) (b) one source of information (Hereafter known as **B**) (c) another source of information (Hereafter known as **C**)



(a) (b)

Figure 4-2. (a) The marginal distribution of random variable **B** (b) The marginal distribution of random variable **C**

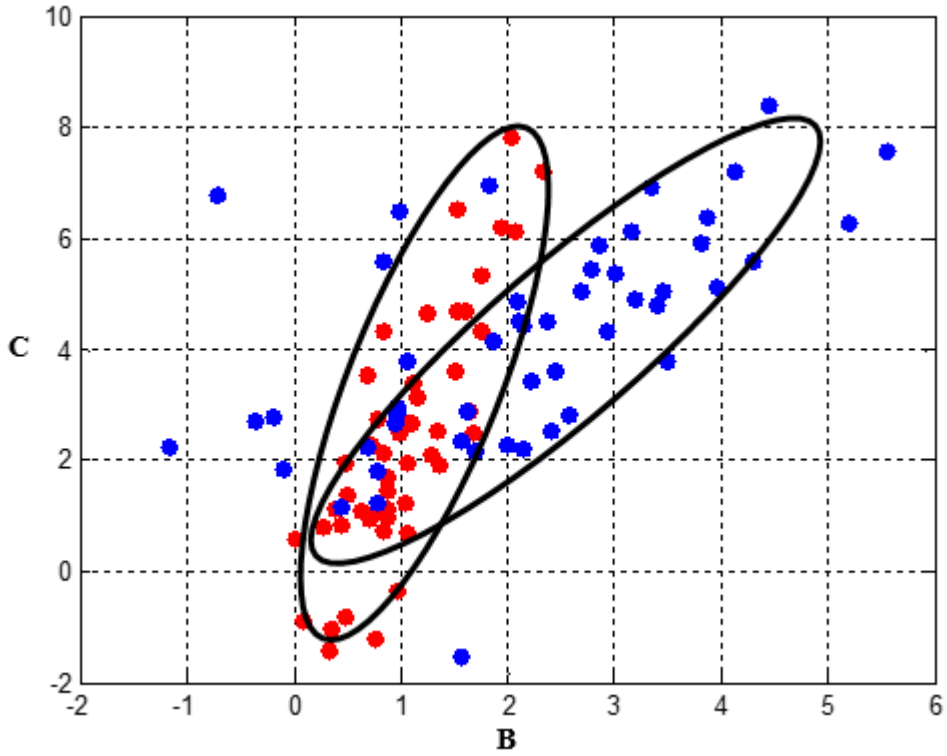


Figure 4-3. The scatter plot of data sources. The mixture of Gaussian behavior can be seen in the plot.

Then, we use the marginal distribution of each data source (shown in Figure 4-2) to calculate *P(A|B)* and *P(A|C)* using Baye's rule:

$$P(A|B) = \frac{f(B|A)P(A)}{\sum_i f(B|A_i)P(A_i)} \tag{4-45}$$

where $f(B|A_i)$ is calculated by:

$$f(B|A_i) = \frac{1}{\sqrt{2\pi}\sigma_{Bi}} \exp(-\frac{|B-\mu_{Bi}|^2}{2{\sigma_{Bi}}^2}) \tag{4-46}$$

where A_i, σ_{Bi}, and μ_{Bi} stand for facies *i*, standard deviation and mean value of **B** within facies *i*. The same set of equations can be applied to calculate *P(A|C).* Figure 4-4 shows the results of classification based on each data source independently.

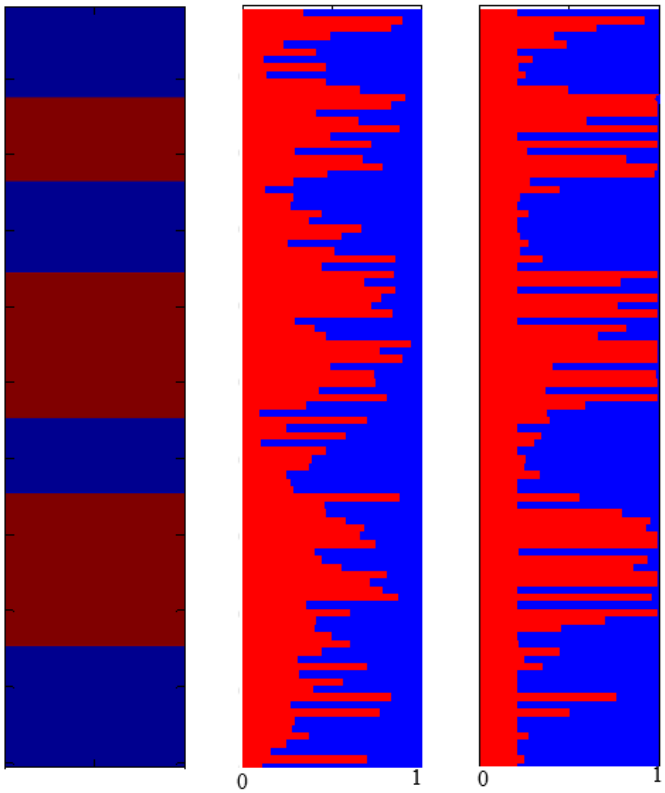


Figure 4-4. (a) The facies map (b) The map of *P(A|B).* (c) The map of *P(A|C).*

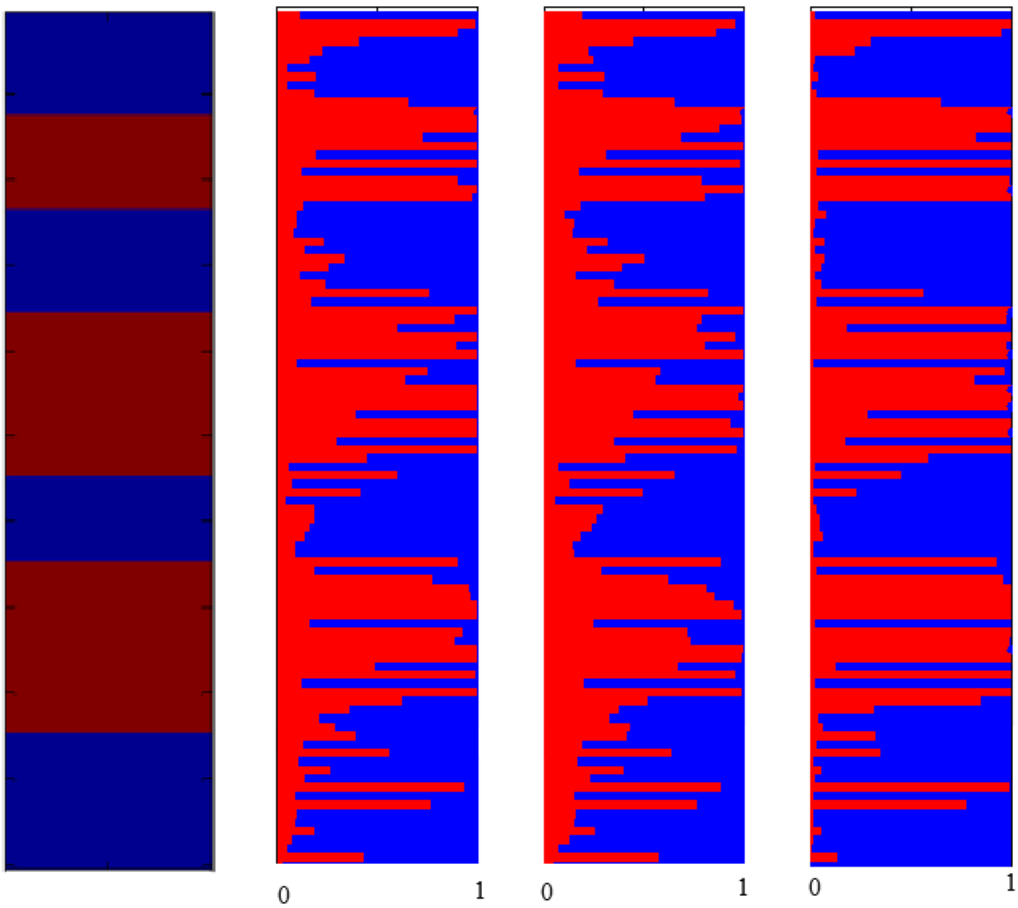
Then, we merge these probability maps with three different approaches:

1- Assuming conditional independence between the data sources ($\tau = 1$)

2- Merging the probabilities with the redundancy Tau weigh proposed by Krishnan et al., (2004) (Equation (4-36)).

3- Merging the probabilities with the proposed method using Equation (4-44).

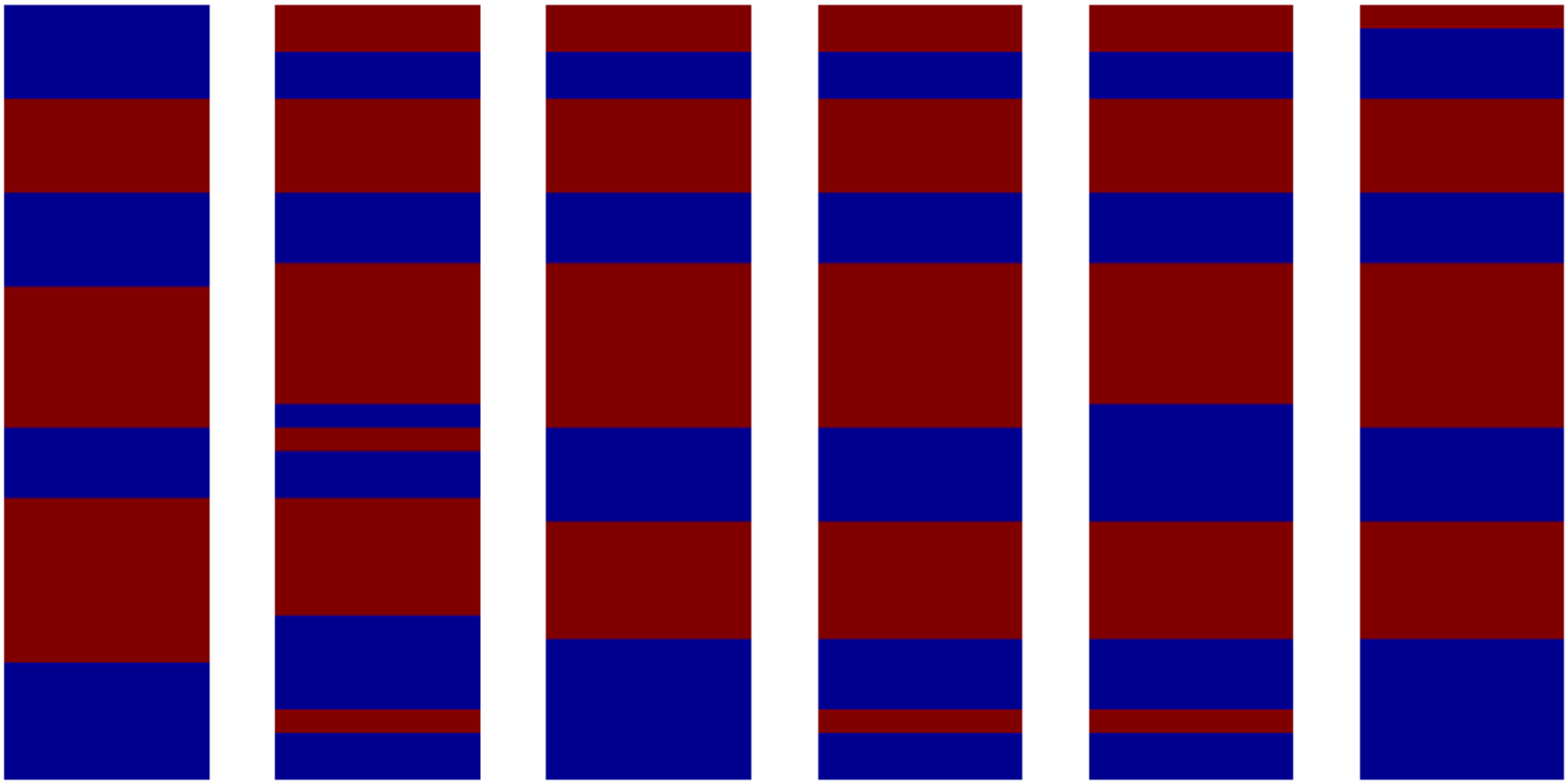
Figure 4-5 shows the results of *P(A|B,C)* with the previous approaches.



(a) (b) (c) (d)

Figure 4-5. (a) The facies map and the results of *P(A|B,C)* using (b) $\tau = 1$ (conditional independence) (c) $\tau = 0.5$ (Krishnan, Boucher, & Journel, 2004) (d) $\tau = 2.43$ (The proposed method)

Then, we calculated the average proportion in a certain volume. If the average value is greater than 0.5, it is assigned to 1, otherwise it will be assigned to 0. The purpose of this process is to compare the accuracy of these different classification method. The results are shown in Figure 4-6.



(a) (b) (c) (d) (e) (f)

Figure 4-6. (a) The up-scaled facies map and the upscaling results of facies using (b) only *P(A|B)* (c) only *P(A|C)* (d) *P(A|B,C)* with $\tau = 1$ (e) *P(A|B,C)* with $\tau = 0.5$ (Krishnan, Boucher, & Journel, 2004) (f) *P(A|B,C)* with $\tau = 2.43$ (The proposed method)

The average value of error between the facies map obtained from the classification method and the original facies map is tabulated in Table 4-4.

Table 4-4. The average value of error between the facies map obtained from the classification method and the original facies map

Classification method	Average error
Using only **B** as the information source	0.2424
Using only **C** as the information source	0.1515
Merging probabilities with $\tau = 1$	0.1818
Merging probabilities with $\tau = 0.5$ (Krishnan, Boucher, & Journel, 2004)	0.2121
Merging probabilities with $\tau = 2.43$ (The proposed method)	0.1212

As can be seen in Table 4-4, since the classification based on only primary source of information (**B**) is erroneous, merging the probabilities results in a more accurate classification method. IT can also be seen that as the value of the tau weight increases, the effect of the secondary information source (**C**) will increase, and the average error decreases.

In the next section, we implement the proposed method in a 2-D section of a 3-D synthetic reservoir. We show that merging the information of different sources and correctly accounting for the redundancy between the data sources greatly improve the results.

4.8 STANFORD V RESERVOIR

In this section, we implement the permanence of ratio hypothesis on a 2D section of a 3D fluvial reservoir. The exhaustive data sets used here are obtained from Stanford V Reservoir data set (Mao & Journel, 1999). The purpose of this section is to show how the algorithm can be applied for facies classification in reservoirs and for probabilistic data integration scheme. Figure 4-7 shows the seismic impedance variations in the reservoir, and Figure 4-8 shows the location of hard data. The color codes red and blue represent the pay and non-pay zone respectively. The reference map of facies is shown in Figure 4-9.

The implementation of the permanence of ratio hypothesis consists of three steps:

1- Classification based on well measurements and calculating *P(A|B)*

2- Classification based on seismic impedance and calculating *P(A|C)*

3- Evaluating the redundancy Tau factor and merging the probabilities

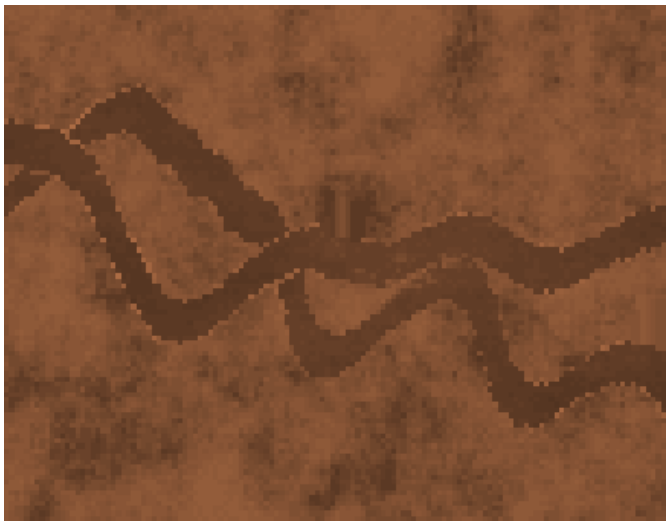


Figure 4-7. The seismic impedance variation over the reservoir

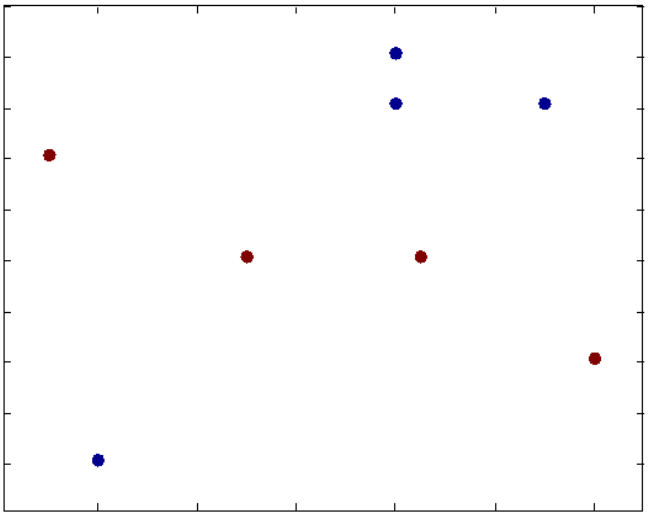


Figure 4-8. The location of hard data. Red and blue points correspond to pay and non-pay facies respectively.

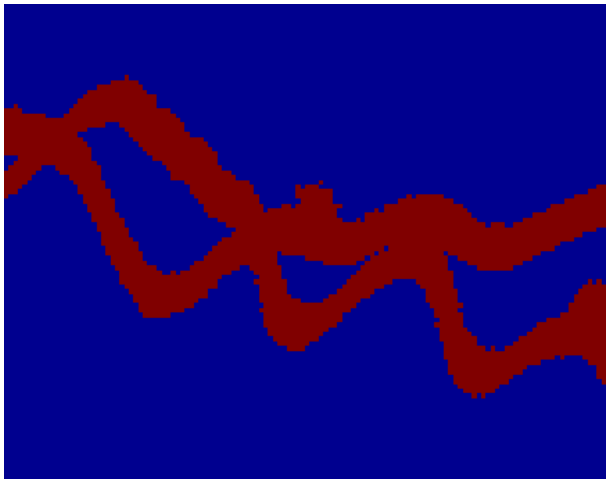


Figure 4-9. The reference map of facies. Red and blue correspond to pay and non-pay facies, respectively.

4.8.1 Classification Based on Well measurements

In order to calculate the *P(A|B)*, we first apply the direct sampling algorithm (discussed in previous chapter) and generate 150 realizations of pay and non-pay facies. Then, the ensemble average of the realizations will be computed and this will yield the *P(A|B)* map. The training image used for the simulations is shown in Figure 4-10

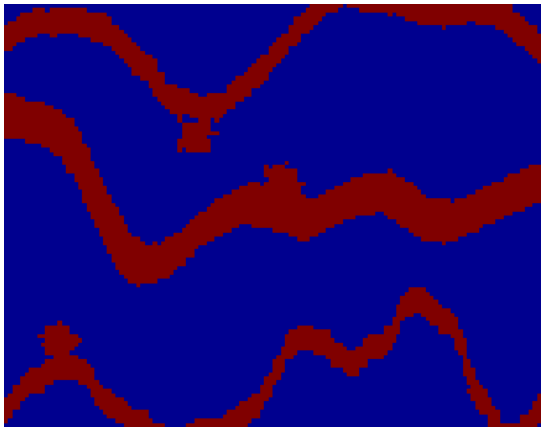


Figure 4-10. The training image used for generation the realizations of facies models using direct sampling algorithm

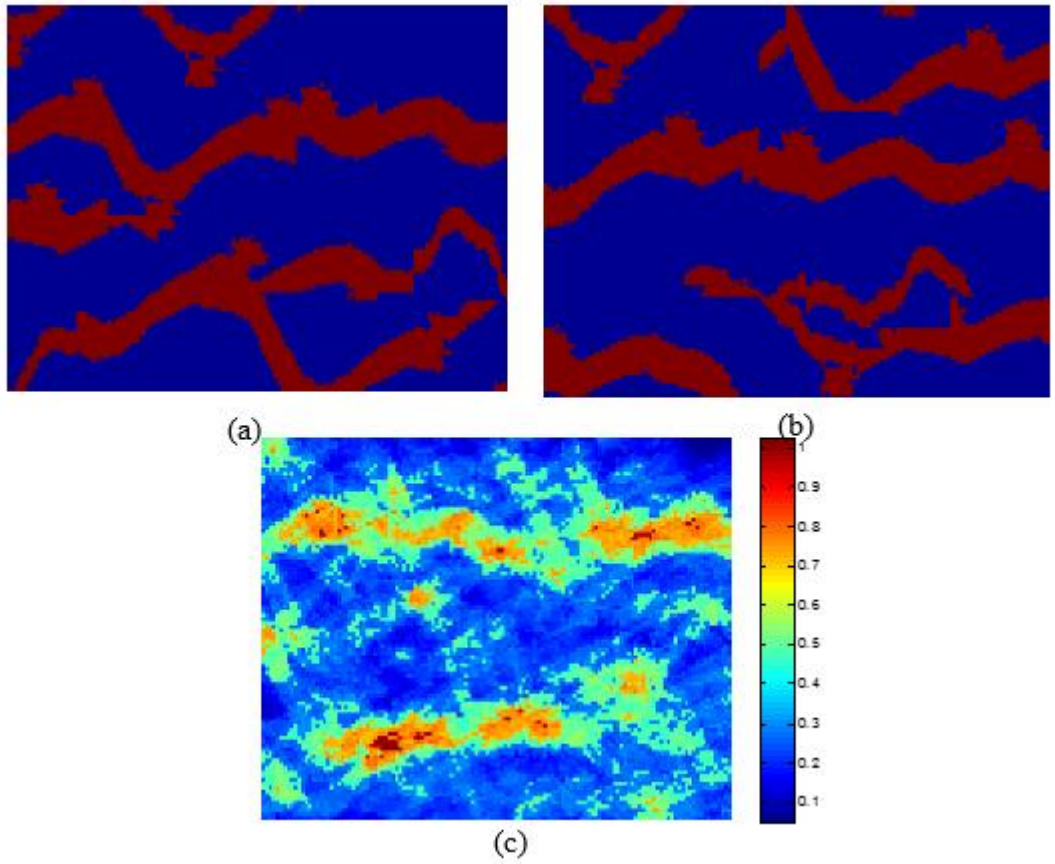


Figure 4-11.Two different realizations and the probability map: (a) Realization 1 (b) Realization 2 (c) the probability map calculated from 150 realizations

As can be seen in Figure 4-11, although each realization exhibits the curvilinear characteristic of the channels, the uncertainty of facies classification associated with final probability map is high (indicated by the number of locations with probability of pay facies close to 0.5) due to the sparsity of conditioning data and the final probability map is not reliable if used solely. The results motivate the calibration of information from the secondary source (seismic) and the integration of that information into the reservoir model. In the next section, we present the classification method based on seismic data.

4.8.2 Classification Based on Seismic Data

In order to stochastically calibrate the seismic attributes to lithology, we use a probabilistic neural network. Artificial neural networks utilize non-linear regression methods to create a non-linear calibration function that relates input data to matching outputs for a training dataset. The trained network can be used to predict outcomes corresponding to another set of input data. Neural networks can be created using differing mechanisms and functions, but, at their most basic, they can be thought of as consisting of three layers: input, transform, and output. The training inputs and outputs shape the learning transform (neurons) over multiple iterations repeating tries until a satisfying rule set is created. Applied to our case, the training phase develops a relationship between the seismic impedance (input) and the corresponding facies (output). The neural network is then applied to a set of seismic impedance as an input and the outcomes are the conditional probability of being in each lithofacies.

A four layer feed forward PNN using a general Gaussian kernel, or Parzen window, can implement exactly the general homoscedastic Gaussian mixtures used to approximate the optimum classifier. This is similar to the mixture of Gaussian modeling but with an important difference – instead of assuming a linear mixture model, the PNN actually calibrates a non-linear mixture model. Maximum likelihood method is used to train the PNN. The structure of the used PNN is shown in Figure 4-12 (Streit & Luginbuhl, 1994). The network shown in Figure 4-12 is called Generalized Fisher (GF) network, and is a generalized Specht's PNN (Specht, 1990). Specht's PNN is a three layer feed-forward NN that uses mixture of uncorrelated Gaussians to estimate the class conditional PDFs.

Specht's PNN is an excellent tool for initial exploration of new large training sets. Nonetheless, its usefulness in practice is limited by two factors. Firstly, because it is based on the

Parzen window estimator, the total number of Gaussian components must be equal to the number of samples in the training set, thus requiring large amount of data storage. Secondly, an intrinsic smoothing parameter must be estimated on the basis of classification performance. Since robust estimates of classification performance are difficult to establish for small sample size, estimates of smoothing parameters might be unreliable in practice.

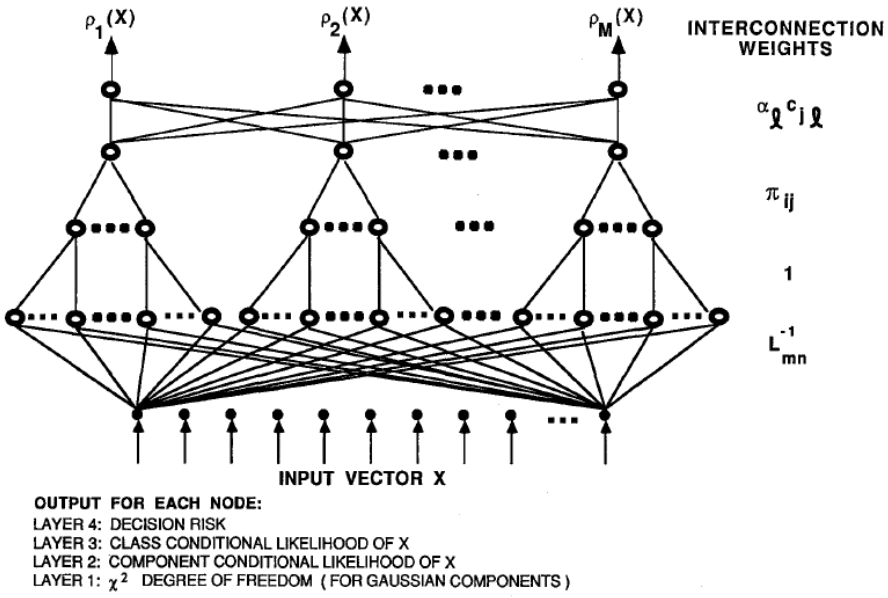


Figure 4-12. Four layer feed-forward probabilistic neural network (Streit and Luginbuhl, 1994)

The output layer of the shown network can be calculated by Equation (4-47).

$$P(A_i|X) = \frac{\alpha_i g_i(X|\lambda_i)}{\sum_{j=1}^{N} \alpha_j g_j(X|\lambda_j)} \quad , i = 1, \dots, N \tag{4-47}$$

Where

$$\sum_{j=1}^{N} \alpha_j = 1 \tag{4-48}$$

$$g_i(X|\lambda_i) = \sum_{j=1}^{G_i} \frac{\pi_{ij}}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left(\frac{-1}{2}(X-\mu_{ij})^t \Sigma^{-1}(X-\mu_{ij})\right) \tag{4-49}$$

$$\sum_{j=1}^{G_i} \pi_{ij} = 1 \tag{4-50}$$

where $\lambda_i = \{\mu_{ij}, \Sigma, \pi_{ij}\}$, G_i is the number of the sample within ith class, and all the parameter μ_{ij}, Σ, π_{ij} and α_j are determined during the training process. The training algorithm is based on Expectation- Maximization (EM) algorithm such that the likelihood of the training data is maximized. The training algorithm is discussed in Appendix.

We used hard conditioning data shown in Figure 4-2 and the corresponding collocated seismic data for training the network. After the training is done, the seismic impedances over the entire reservoir model are fed to the network to calculate the probability of each facies at a location given the collocated seismic impedance data. Figure 4-13 shows the probability map of pay zone given seismic impedance.

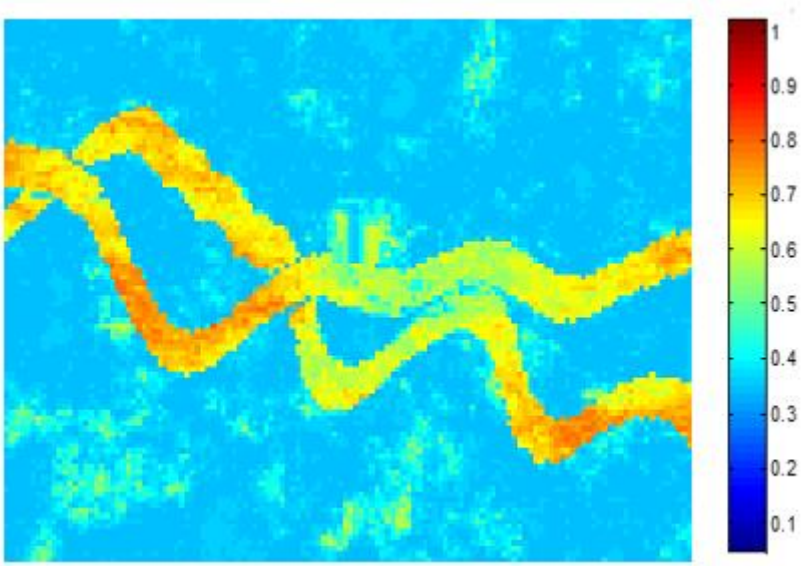


Figure 4-13. The probability of pay zone given seismic impedance data

4.8.3 Evaluating the redundancy factor and merging the probabilities

After the probability maps are generated, we calculate the average tau weight by Equations (4-37) to (4-44).

Since the facies at the well locations are determined by cluster analysis of well log measurements, instead of using facies indicator value for calculating the redundancy factor, we used the well log measurements (which are continuous variables) for calculating the redundancy factor between the primary and the secondary data. The probability maps are then merged to calculate the final *P(A|B,C)*. We also used Krishnan et al. (2004) equation for calculating tau weights. Figure 4-14 shows the scatter plot of porosity and seismic impedance at the well locations for pay and non-pay zones. The mixture of Gaussian behavior of the data can be seen in Figure 4-14.

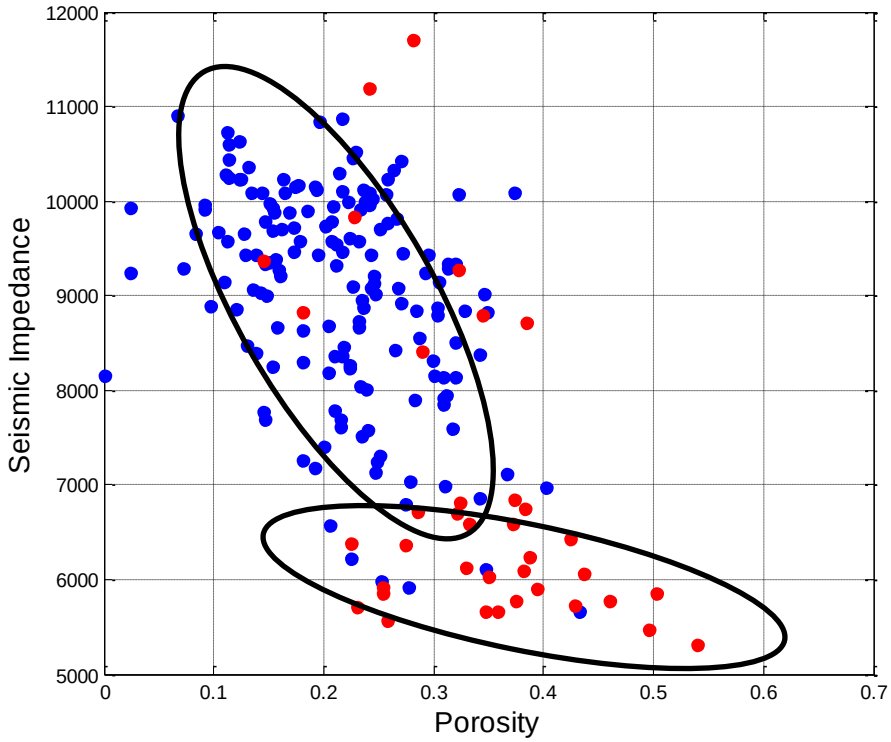


Figure 4-14. The scatter plot of porosity and seismic impedance. Colors red and blue correspond to pay and non-pay zones respectively.

Figure 4-15 shows the results of tau weight using Equation (4-37) as a function of porosity and seismic impedance. Using Equation (4-44) leads to the average tau of 1.73, whereas the Equation (4-36) leads to the tau weight of 0.7.

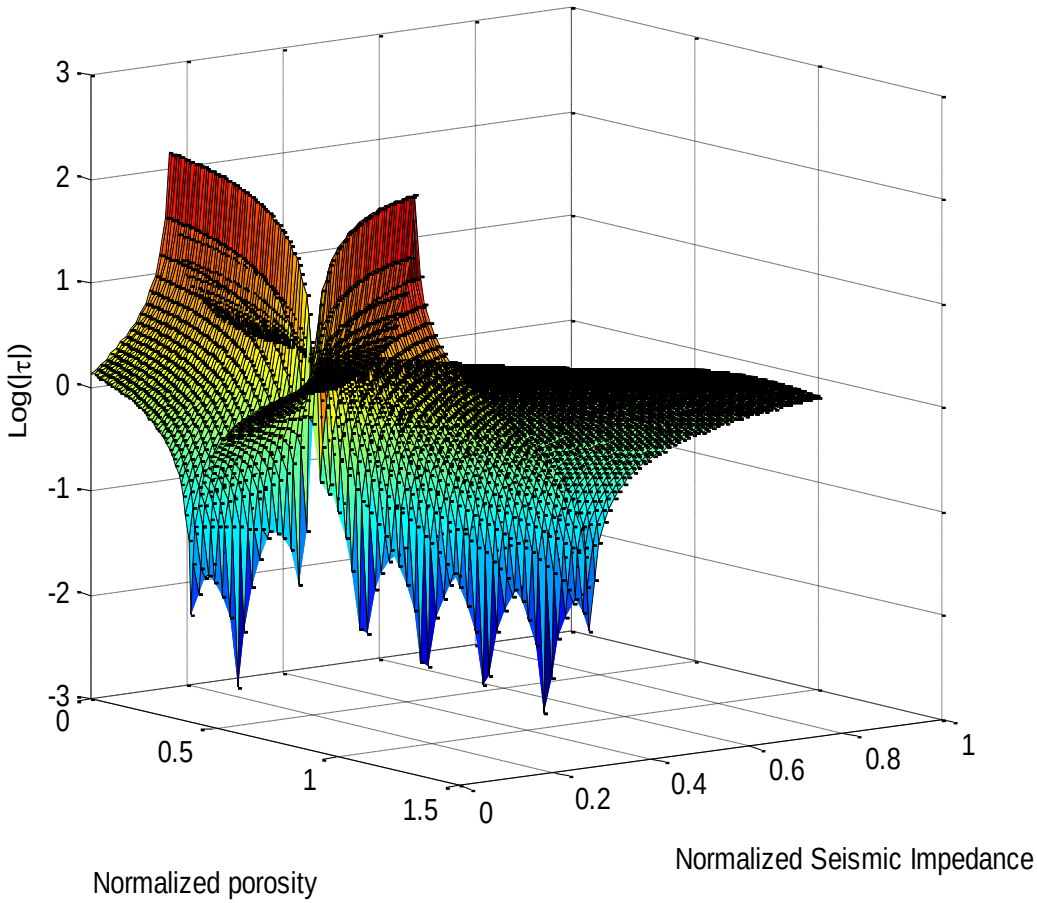


Figure 4-15 The variation in tau value with porosity and seismic impedance (This surface is used for calculating the average tau using Equation (4-44)).

The probabilities are then merged in three different scenarios:

1- Conditional independence ($\tau = 1$),

2- Krishnan weight ($\tau = 0.7$)

3- using Equation (4-44) ($\tau = 1.73$)

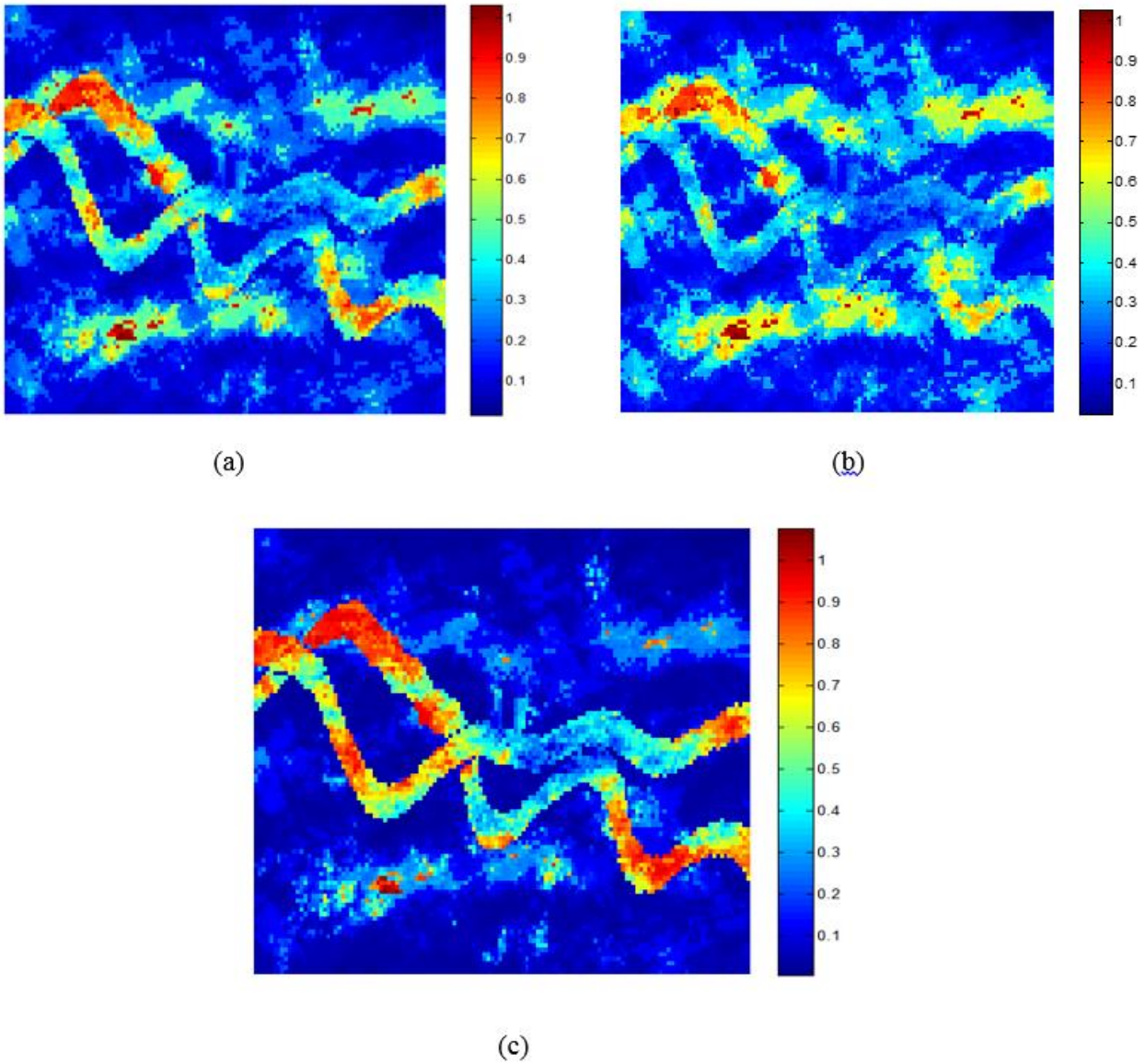


Figure 4-16. The merged probability of pay zone assuming (a) Conditional independence ($\tau = 1$), (b) Krishnan weight ($\tau = 0.7$) (c) using Equation (4-44) ($\tau = 1.73$)

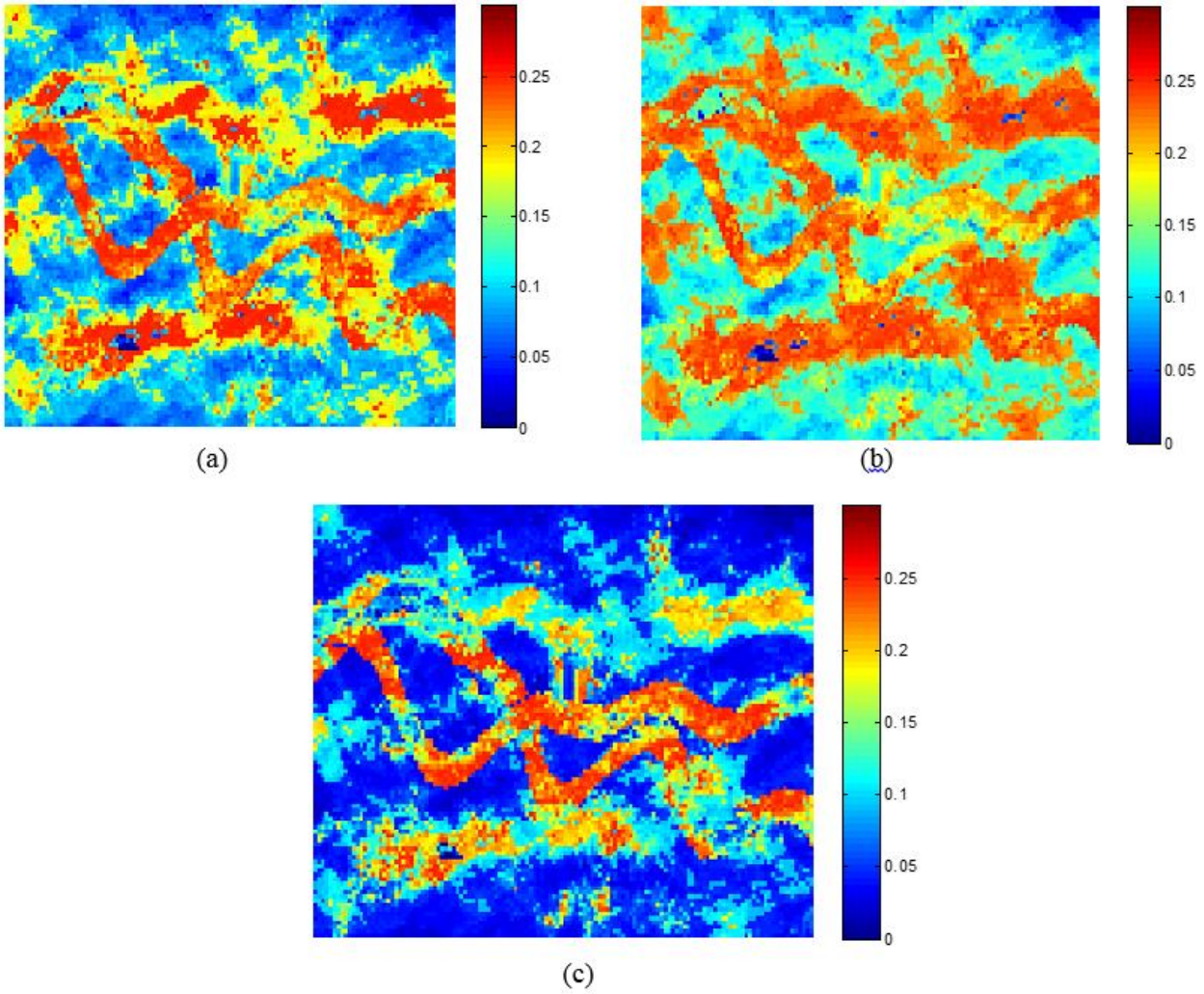


Figure 4-17. The variance map of pay zone assuming (a) Conditional independence ($\tau = 1$), (b) Krishnan weight ($\tau = 0.7$) (c) using Equation (4-44) ($\tau = 1.73$)

The results shown in the previous section demonstrate how the permanence of ratio hypothesis can be used to improve the accuracy of the simulation and decrease the uncertainty associated with prediction of reservoir facies.

As can be seen in Figure 4-16, the closest probability map to the reference map is the one calculated with our proposed method. This map also has less uncertainty than the other probability maps (indicated by the smaller area of regions with probability in the range of 0.5).

As was shown in the previous section, the redundancy factor calculated from our proposed method is greater than unity which means the influence of the seismic data should be increased.

4.9 CONCLUSION

In this chapter, we presented another method of data integration based on the permanence of ratio hypothesis. The permanence of ratio hypothesis is a probabilistic scheme to merge information from different sources and aims to calculate the conditional probability distribution $P(A|B,C)$ using the conditional probabilities $P(A|B)$ and $P(A|C)$ that are known or calibrated from the available information. In order to model the conditional probability, it would be convenient if the information from each data source can be assessed independently in order to find $P(A|B)$ and $P(A|C)$, and then these joint probabilities are merged to calculate $P(A|B,C)$ accounting for the redundancy between different data sources. This is precisely the premise of the permanence of ratio hypothesis.

In this chapter, the mathematical basis of the permanence of ratio hypothesis was presented. The advantage of this scheme over data integration assuming total independence between the data as well by assuming conditional independence between data has been demonstrated. It was shown that the permanence of ratio hypothesis is a tautology and the probability values calculated are guaranteed to be licit functions. On the other hand, integration employing the data independence assumption can lead to probabilities greater than 1.

We propose a methodology for calculating the redundancy between different sources of information. Our formulation is based on the information from each data modeled using a mixture of Gaussian assumption indicative of the multiple facies or categories of rock properties observed in the reservoir. We implemented our method for a synthetic reservoir example and

demonstrated that the proposed method leads to more accurate and less uncertain representation of rock facies compared to the existing approaches.

In the next chapter, we implement this method using real field data in order to find the potential sweet spots for future drilling and production.

Chapter 5: Characterization of Oolitic Bank in the Gulf of Mexico: A Case Study

In this chapter, we present a case study for finding the lithofacies map in a carbonate reservoir in the Gulf of Mexico using permanence of ratio hypothesis. In the next section, we present available data sources, which are core data, well log measurements, and seismic attributes. Then, in the advanced data analysis section, we show a deterministic neural network customized for facies identification based on only well measurements. The results indicate there is a mismatch in facies determination at some depths. Therefore, we evaluate and the seismic data and the well log data independently in order to calculate the probability of being in each facies given that source of information. Afterwards, we use the tau model to merge the information and find the final probability map. Then, finally the different realizations are generated in order to assess the uncertainty.

5.1 Geological Description of the Field

The field under study in this section is an offshore carbonate reservoir in the Gulf of Mexico. The field mainly consists of three rock types: mudstone, grainstone, and dolomitized rocks. Dolomitization is indicative of chemical alterations in the rocks. Dolomitization in general tends to render the rock more brittle and hence prone to fracturing. In the case of fractured brecchia, dolomitization can also cause permeable flow pathways to bet blocked and hence reduce the volumetric sweep efficiency of processes implemented to recover the stored oil and gas. Thus, the most productive rock type in the field is grainstone. The main purpose of this project was to calculate the proportion of these rock types at all locations and assess the uncertainty associated with it, so potential sweet spots for drilling are identified and the proper reservoir management decisions can be made.

5.2 EXPLORATORY DATA ANALYSIS OF DATA

5.2.1 Core Data

The core data is used for calibrating the lithofacies proportions observed at different depths to the log attributes. Figure 5-1 shows the proportion of different lithofacies observed along the cores. Facies 1, 2, and 3 are mudstone, grainstone, and dolomitized rocks respectively. Figure 5-2 shows the depth of the core data along each well.

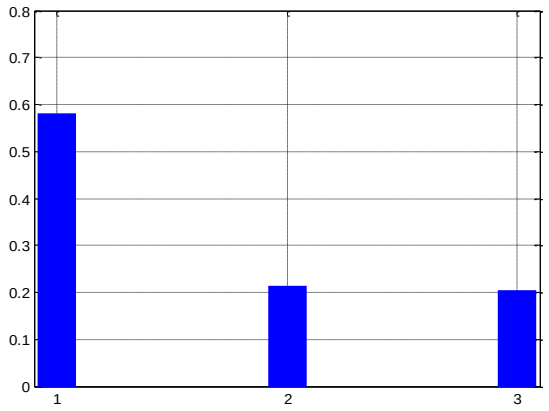


Figure 5-1. Proportion of facies inferred from cores used for calibrating the lithofacies to the log attributes. Facies 1, 2, and 3 are mudstone, grainstone, and dolomitized rocks respectively.

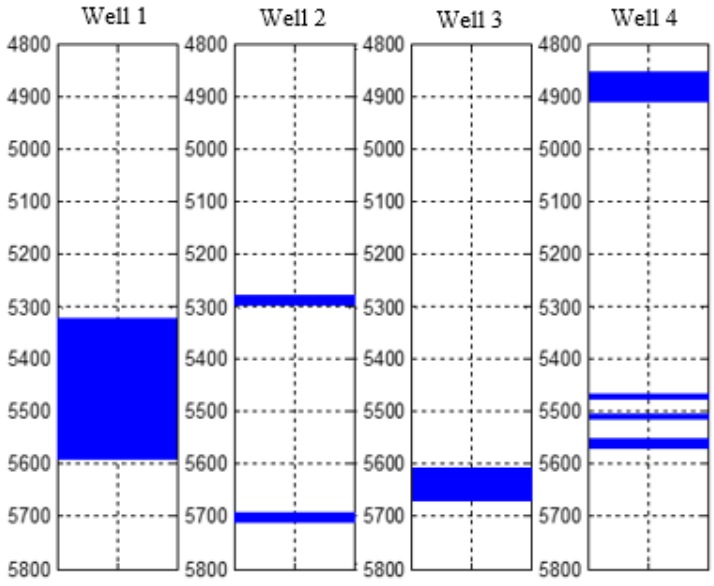


Figure 5-2. The depth along each well where core data is available.

5.2.2 Log Data

The well logs for 6 wells were provided. Figure 5-3 shows the log variables along different wells at the depth seismic data was available.

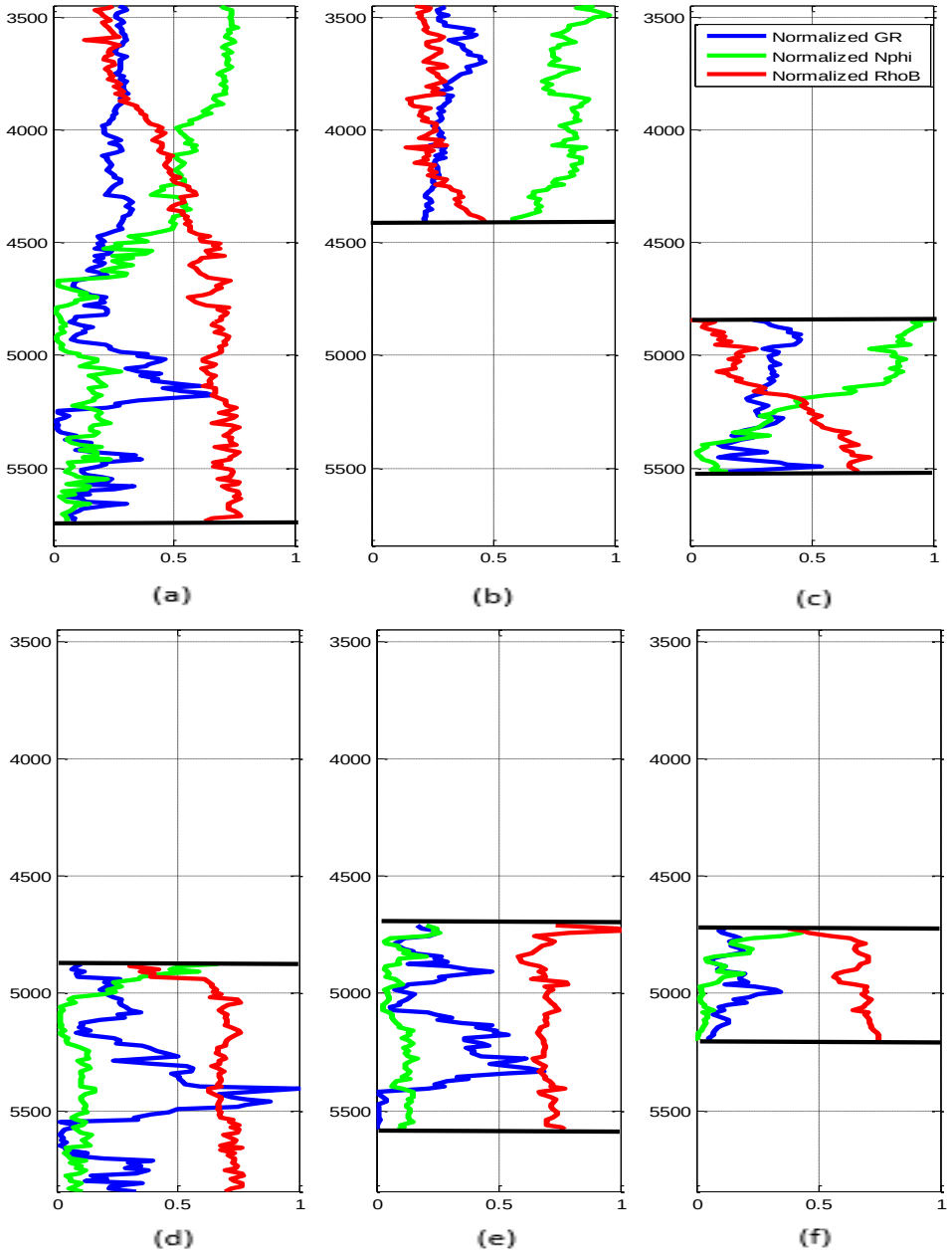
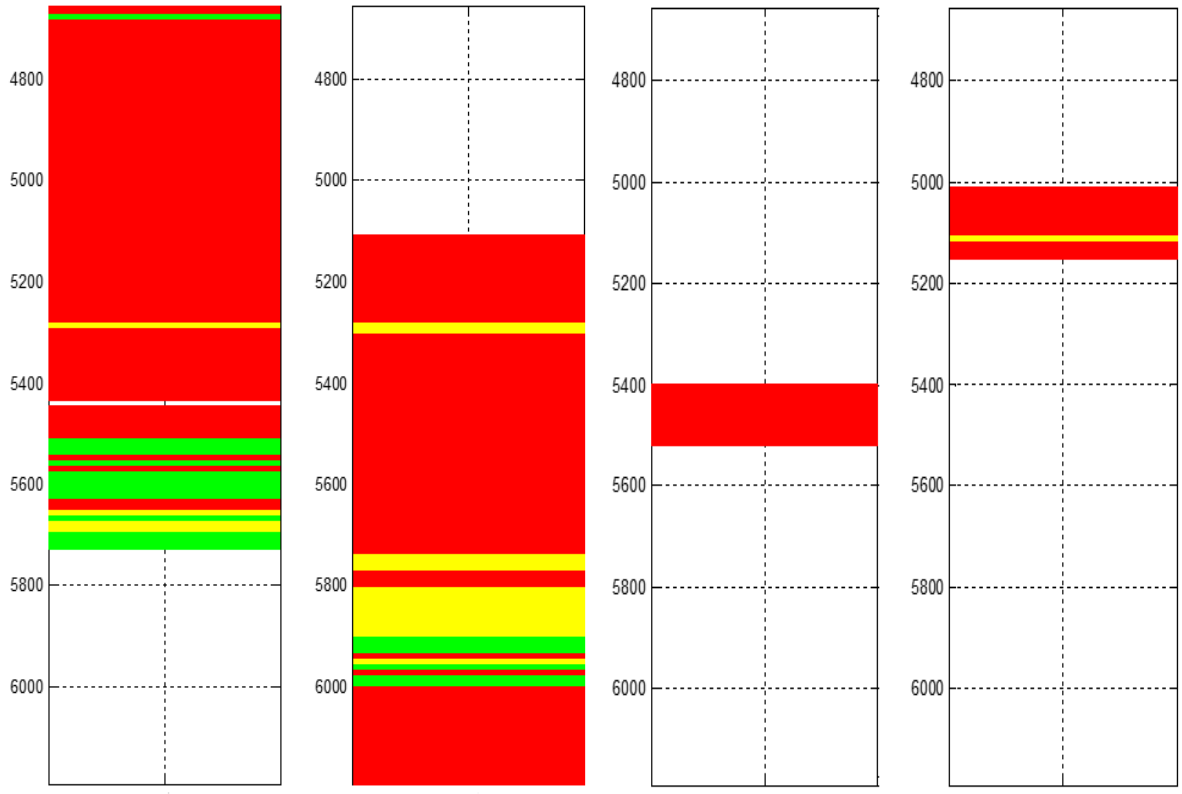


Figure 5-3.The Normalized gamma ray, neutron porosity and bulk density logs available for wells (a) Well 1 (b) Well 2 (c) Well 3 (d) Well 4 (e) Well 5 (f) Well 6.

5.2.3 Drill Cutting

Drilling cutting data is used for stochastically calibrating the lithofacies to seismic data. Figure 5-4 shows the drill cutting lithofacies along the wells. The colors red, green, and yellow are representative of mudstone, grainstone, and dolomitized rocks respectively.



(a) (b) (c) (d)

Figure 5-4 The available drilling cutting lithofacies at wells (a) Well 1 (b) Well 2 (c) Well 3 (d) Well 4.

5.2.4 Seismic Data

Post stack seismic inversion data is available for the following attributes:

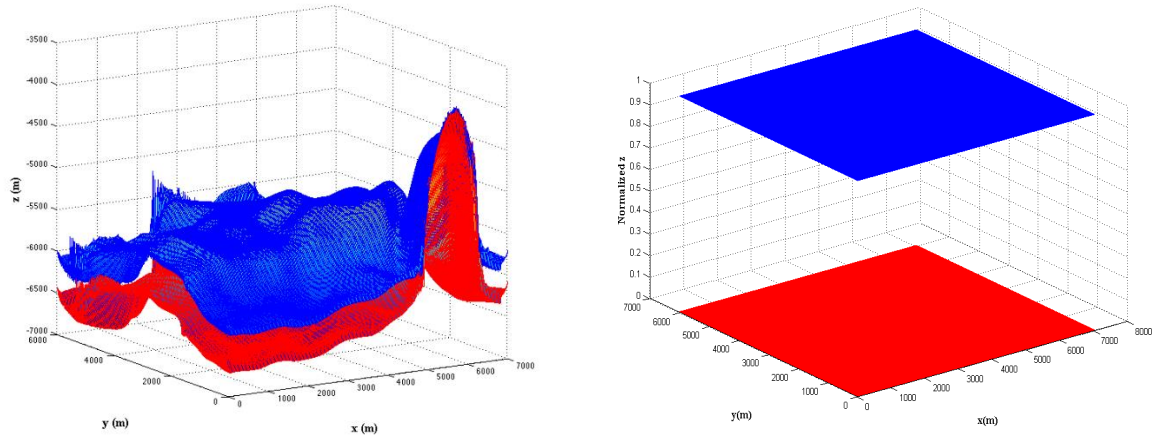
- volume of clay
- volume of dolomite
- volume of limestone

- acoustic impedance
- Compressional velocityVp
- bulk density (RhoB)

The top and base surfaces of the reservoir interpreted from seismic amplitude data vary with areal location (see Figure 5-5). In order to perform stochastic calculations, we map the vertical dimension (or resolution) of the reservoir layer into a uniform vertical grid (shown in Figure 5-5.b) by performing a stratigraphic transformation:

$$z_{trans} = \frac{z - z_{base}}{z_{top} - z_{base}} \qquad (5\text{-}1)$$

Where z_{top} and z_{base} vary horizontally. The number of grid blocks of the transformed region in x, y, and z direction is 281, 241, and 50 respectively. Figure 5-6 and 5-7 show a map of acoustic impedance and bulk density at the middle of the mapped box respectively.



(a) (b)

Figure 5-5. (a) Top and base surface of the reservoir : (a) before stratiographic transformation (b) after stratiographic transformation.

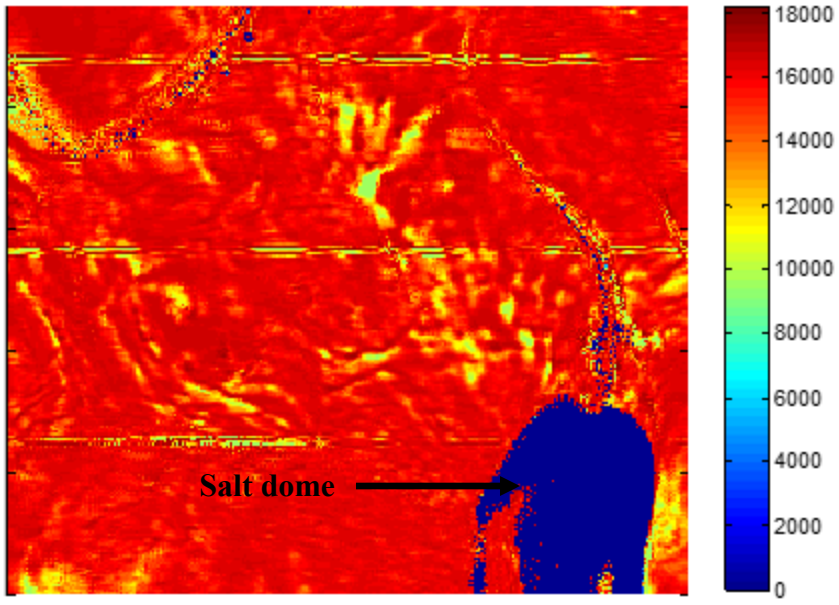


Figure 5-6. Acoustic impedance map at the middle z-section after performing the stratigraphic transformation.

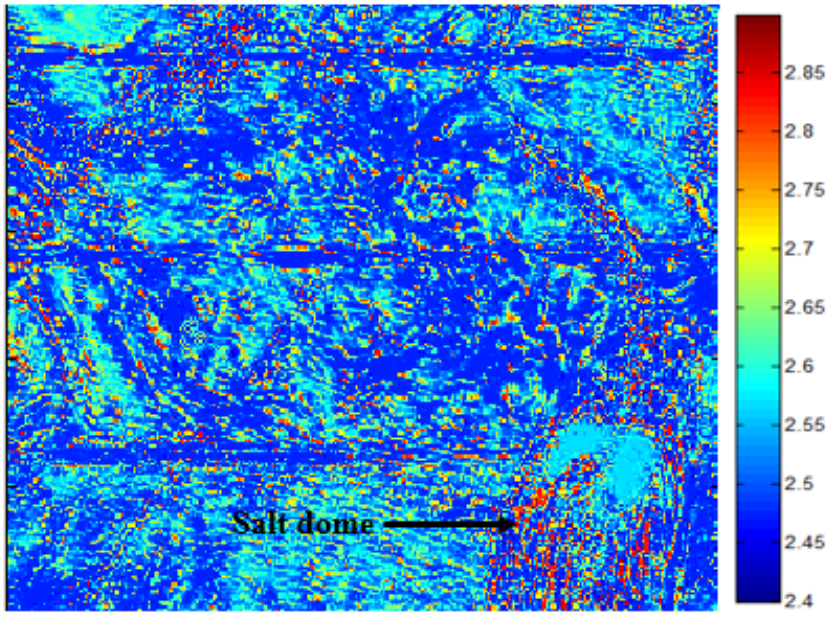


Figure 5-7. Bulk density map at the middle z-section.

It should be noted that the horizontal strikes seen in Figure 5-6 and Figure 5-7 are artifacts in seismic inversion process which is probably because the seismic attributes are calculated at smaller volumes and then fused together to construct the entire seismic volume.

5.2.4.1 Redundancy between seismic attribues

The covariance matrix between the different normalized seismic attributes is presented below. Please note that X1, X2, X3, X4, X5, X6 are Volume of Clay, Volume of dolomite, volume of limestone, Acoustic impedance, Vp, and bulk density (RhoB) respectively.

$$\mathrm{Cov} = \begin{bmatrix} 0.0075 & -0.0015 & -0.0025 & 0.0096 & 0.0080 & -0.0015 \\ -0.0015 & 0.0089 & -0.0053 & -0.0073 & -0.0059 & 0.0089 \\ -0.0025 & -0.0053 & 0.0088 & 0.0015 & 0.00079 & -0.0053 \\ 0.0096 & -0.0073 & 0.0015 & 0.1190 & 0.0829 & -0.0075 \\ 0.0080 & -0.0059 & 0.00079 & 0.0829 & 0.0593 & -0.0060 \\ -0.0015 & 0.0089 & -0.0053 & -0.0075 & -0.0060 & 0.0089 \end{bmatrix}$$

The correlation coefficient Matrix is:

$$\rho = \begin{bmatrix} 1 & -0.1794 & -0.3099 & 0.3203 & 0.3801 & -0.1812 \\ -0.1794 & 1 & -0.5996 & -0.2251 & -0.2547 & 1 \\ -0.3099 & -0.5996 & 1 & 0.0457 & 0.0345 & -0.5981 \\ 0.3203 & -0.2251 & 0.0457 & 1 & 0.9867 & -0.2299 \\ 0.3801 & -0.2547 & 0.0345 & 0.9867 & 1 & -0.2594 \\ -0.1812 & 1 & -0.5981 & -0.2299 & -0.2594 & 1 \end{bmatrix}$$

The Eigen Values of the covariance matrix is as follows:

Table 5-1. The Eigen values of the covariance matrix of seismic attributes

Eigen Value	% of total EV value
0.1793	0.8438
0.0212	0.9434
0.0087	0.9844
0.0025	0.9869
0.0008	1
0.0000	1

Correlation coefficient matrix and eigen values of the covariance matrix show a lot of dependence between the 6 attributes. We did the Principle Component Analysis (PCA), and used 2 most important Principle Components (PCs), which contain about 95% of the total information. Figure 5-8 and 5-9 show the Seismic PCs along the middle z-section.

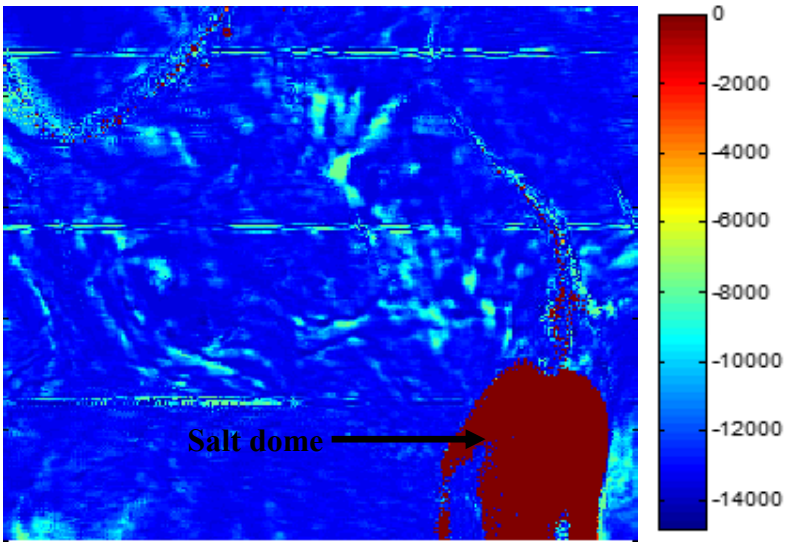


Figure 5-8. First principle component map computed using the seismic attributes along the middle z-section.

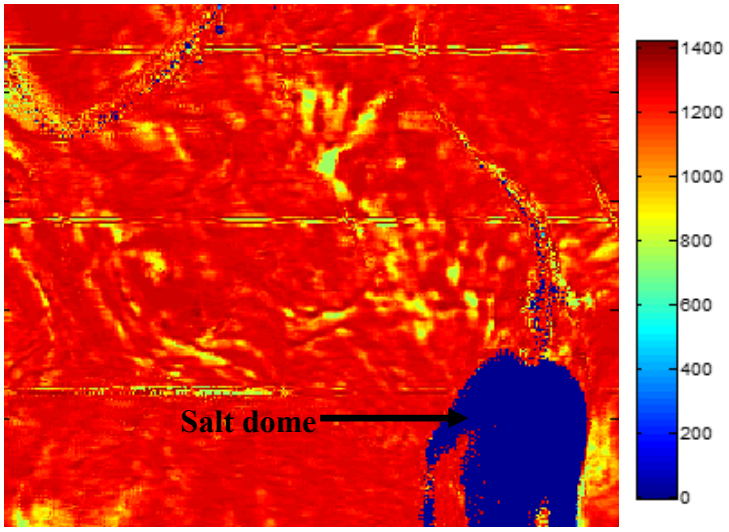


Figure 5-9. The second principle component map computed using the seiemic data along the middle z-section.

5.3 ADVANCED DATA ANALYSIS

5.3.1 Deterministic Neural Networks for Pattern Recognition

We employed a deterministic neural network to evaluate the rock types at all depths of well using the known core data at a few depths. The neural network calibrates the relationship between log attributes and the corresponding rock type recorded along cores. Neural network training was done using Well 1, Well 2, Well 3, and Well 4 core data. The trained network is then applied at all depths for all wells, to broadly categorize the rock types into dolomia, grainstone, mud/pack/wackestone and other.

In each classification problem, the data is divided to 3 categories: training data, validation data, and testing data. The training data is used to calibrate the output relationship with the input data. Then, the performance of the trained network is evaluated on validation data. If the error is acceptable, the training process will stop. Otherwise, the training is continued to decrease the error. Finally, the testing data is fed to the trained network to assess the accuracy of the network.

In order to evaluate the performance of the neural network, the confusion matrix is calculated. In classification and pattern recognition problems, a confusion matrix, also known as contingency matrix or an error matrix, is a table that allows visualization of the performance of the algorithm (neural network in our case). Each column of the matrix corresponds to the instances in the predicted class, whereas each row corresponds to the instances in the actual class. Therefore, the diagonal elements represent correctly classified cases, whereas the off-diagonal elements correspond to the incorrectly classified cases. For a perfect classification algorithm, the confusion matrix should be diagonal.

Overall, there were about 628 sample data extracted from the core samples that gave training error of about 15%. The training results and the confusion matrix are shown in the tables

below. Class 1, 2, 3 and 4 correspond to dolomia, grainstone, mudstone and other facies respectively.

Table 5-2. Training confusion matrix

Target Class / Output Class	1	2	3	4	Error
1	53 12.0%	0 0.0 %	0 0.0 %	6 1.4%	10.2 %
2	0 0.0 %	0 0.0 %	0 0.0 %	0 0.0 %	NaN
3	6 1.4 %	4 0.9 %	230 52.3 %	24 5.5 %	12.9 %
4	4 0.9 %	4 0.9 %	19 4.3 %	90 20.5 %	23.1 %
Error	15.9 %	100 %	7.6 %	25.0 %	15.2 %

Table 5-3. Validation confusion matrix

Target Class / Output Class	1	2	3	4	Error
1	11 11.7 %	0 0.0 %	0 0.0 %	1 1.1 %	8.3 %
2	0 0.0 %	0 0.0 %	0 0.0 %	0 0.0 %	NaN
3	0 0.0 %	0 0.0 %	45 47.9 %	4 4.3 %	8.2 %
4	1 1.1 %	1 1.1 %	4 4.3 %	27 28.7 %	18.2 %
Error	8.3 %	100 %	8.2 %	15.6 %	11.7 %

Table 5-4. Testing confusion matrix

Target Class / Output Class	1	2	3	4	Error
1	8 8.5 %	0 0.0 %	0 0.0 %	1 1.1 %	11.1 %
2	0 0.0 %	0 0.0 %	0 0.0 %	0 0.0 %	NaN
3	0 0.0 %	1 1.1 %	49 52.1 %	3 3.2 %	7.5 %
4	1 1.1 %	0 0.0 %	3 3.2 %	28 29.8 %	12.5 %
Error	11.1 %	100 %	5.8 %	12.5 %	9.6 %

As can be seen from the results on the confusion matrix, the confusion matrix value corresponding to Class 2 (grainstone) is not being reported. This is because there are very few observations of grainstone in the training data set and the number of observations (10) is equal to the number of hidden layers in the neural network. This causes the network to be overfitted for that class and the confusion matrix is not reported. According to the confusion matrix values for the other classes, the misclassification error of the network is very low.

As the results summarized in the confusion matrix reveal, there is some misclassification error associated with the neural network procedure and that error is not the same for all rock types. The classification error is affected by the number of samples available to train the neural network. There is a combination of error in predicting the rock type at a certain depth due to the inadequacy of samples together with the uncertainty in predicting the rock type due to inadequacy of well logs. The logs are affected by several other environmental variables and so they yield imprecise information about the rock type. In the subsequent sections we discuss two probabilistic classification schemes that can be used for this purpose.

5.3.2 Probabilistic Classification Using Mixture of Gaussian Method

Deterministic neural network has some error in determining lithofacies at some locations. In order to rigorously quantify the uncertainty in rock classification, we used probabilistic classification using mixture of Gaussian method for determining the probability of being in each facies at each location given the log attributes.

The mixture of Gaussian classification method assumes that all log attributes within each class follow normal distributions. The parameters of the normal distributions can be found by using the core data. Applied to our case, it is assumed that each log attribute shows a Gaussian behavior within different lithofacies. The mean corresponding to each Gaussian model and the covariance between the log attribute in different lithofacies is calculated, and then Bayes' rule is applied to calculate the conditional probability of being in each facies at well location given the attribute value:

$$P(A_i|\underline{B}) = \frac{f(\underline{B}|A_i)P(A_i)}{\sum_{j=1}^{N} f(\underline{B}|A_j)P(A_j)} \quad , i = 1, \dots, N \tag{5-2}$$

The likelihood of the attribute value given a rock type is calculated using the Gaussian density:

$$f(\underline{B}|A_i) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{\frac{-1}{2}} \exp\left(\frac{-1}{2}(\underline{B} - \mu)^t \Sigma^{-1} (\underline{B} - \mu)\right) \tag{5-3}$$

$\underline{B}$ is the vector of log attributes, μ and Σ are the mean and covariance matrix of log attributes, N is the number of facies, A_is are indices corresponding to different lithofacies, and $P(A_i)$ is the prior probability of class i and can be calculated using the prior proportion of each facies in the training data. Afterwards, we calculate the probability of observing the different lithofacies at depths where the drilling cutting is available. We obtain these probabilities by aggregating the drill cutting information available at a particular depth. We validate the mixture

of Gaussian approach for computing the probability of observing a facies at a particular depth by comparing the probability distribution with the drill cutting facies. For ease of comparison, we have extracted the most likely facies from the distribution obtained using the mixture of Gaussian hypothesis and that derived using the drill cutting information and show the comparison in Figure 5-10.

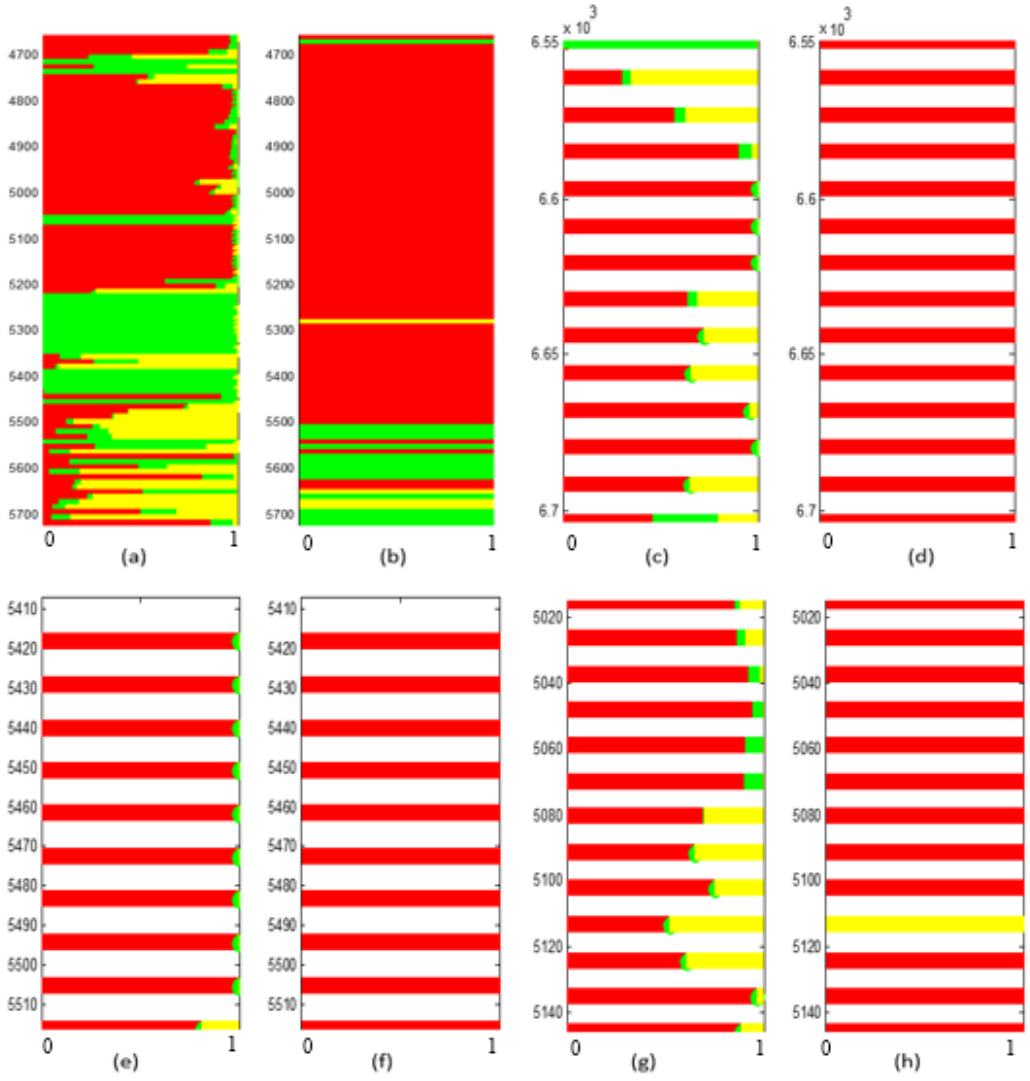


Figure 5-10 (a) Mixture of Gaussian Classification results for Well 1 (b) Drilling Cutting Facies for Well 1 (c) Mixture of Gaussian Classification results for Well 2 (d) Drilling Cutting Facies for Well 2 (e) Mixture of Gaussian Classification results for Well 3 (f) Drilling Cutting Facies for Well 3 (g) Mixture of Gaussian Classification results for Well 4 (h) Drilling Cutting Facies for Well 4

As the results indicate the classification is accurate for most wells except Well 1. This might be due the fact that the log measurements at that region are not capable of correctly distinguishing the rock types. In order to improve the classification algorithm, the seismic data should also be incorporated. In the next section, we present a probabilistic method for classification of rock types based on the seismic attributes. It will be shown the classification performance is improved for well 1 indicating that although the log attributes cannot correctly distinguish the rock types at that region, the seismic attributes are able to classify the rock types with smaller mismatch. In fact, this is the motivation of merging the information from these two sources.

5.3.3 Probabilistic neural networks for calibrating lithofacies information from seismic attributes

In order to stochastically calibrate the seismic attributes to lithology, we also used a probabilistic neural network. We used a four layer feed forward PNN (as explained in the previous chapter) for seismic classification. Applied to our case, the training phase develops a relationship between the multiple log attributes (input) and the corresponding facies (output). The neural network is then applied with a set of seismic attributes as input and the outcomes are the conditional probability of being in each lithofacies.

We used a subsample of drilling cutting data such that the proportion of lithofacies observed in the core data is reflected and used these to augment the data used for training the network. After the training is done, seismic derived PCs are fed to network to calculate the probability of being in each facies given seismic attributes. To validate the performance of the neural network we compare the probabilistic classification results of facies for Well1 to drilling cutting data. The reason for choosing Well1 is the abundant number of drilling cutting data

available. Figure 5-11 shows the comparison of probabilistic classification of lithofacies obtained using the neural network and that based on drill cutting data. As mentioned earlier, it can be seen that the classification performance based on the seismic attributes is more accurate than the one based on log measurements for Well 1. In fact, this is the motivation of merging the information from these two sources.

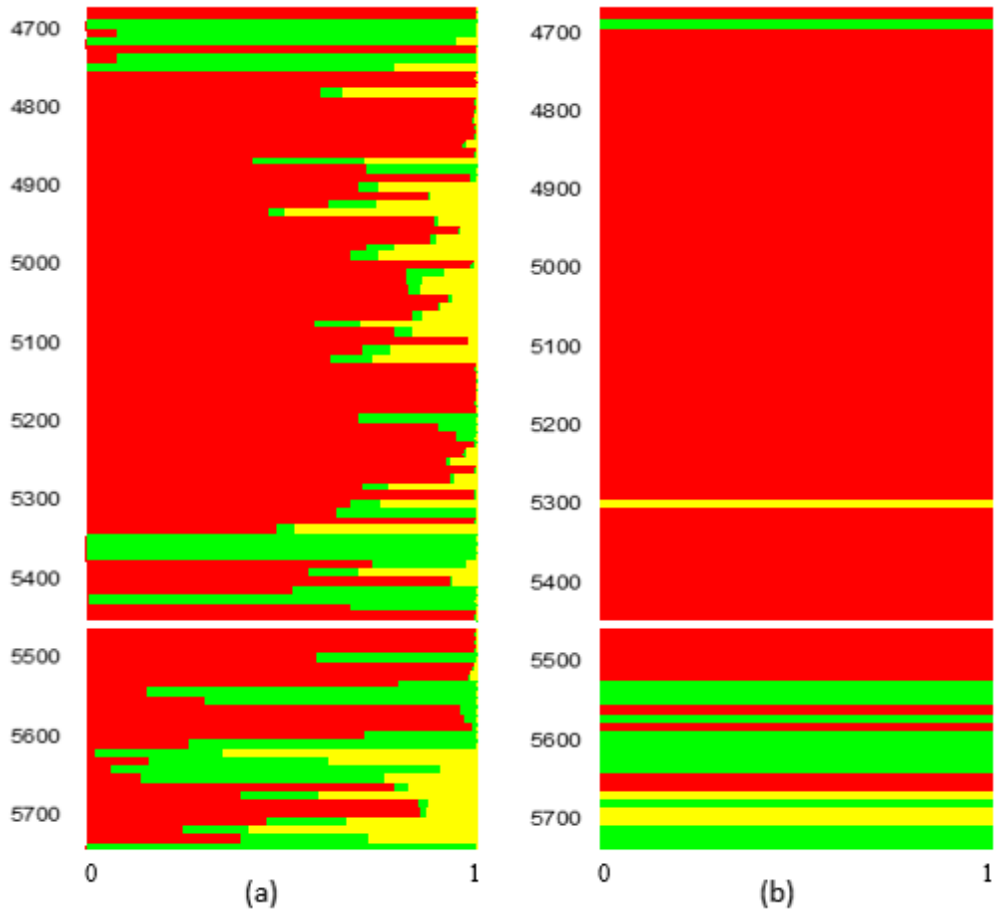
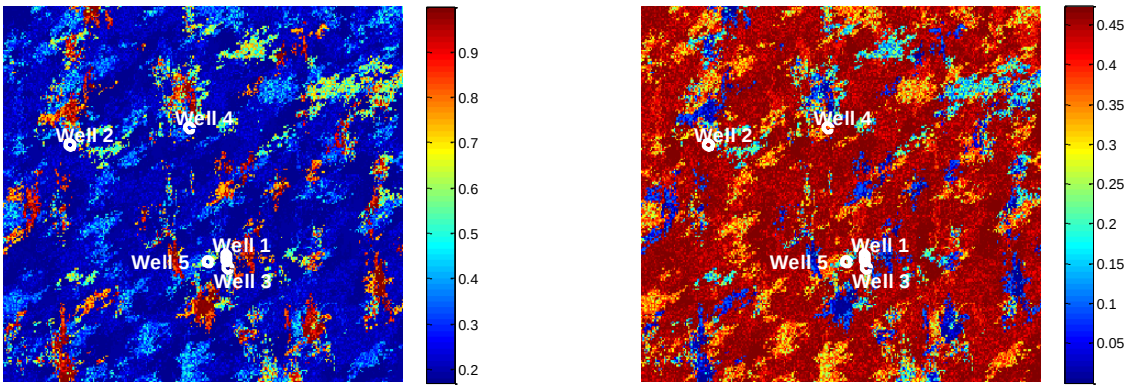


Figure 5-11. The comparison of probabilistic neural network classification with drilling cutting for Well 1 (a) probabilistic neural network results (b) Drill cutting data

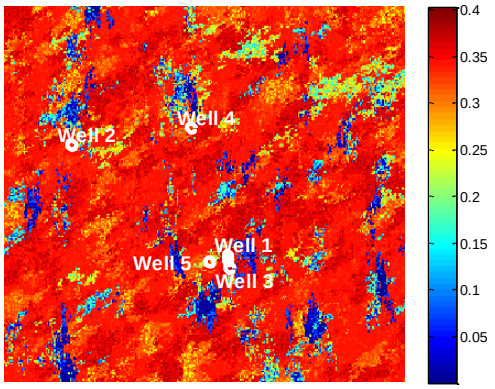
5.4 RESULTS

5.4.1 Probabilistic classification of facies at any location based on nearest well log Data

In order to find the probability of being in each lithofacies type at locations away from wells, a geostatistical simulation is performed conditioned to the well log data. . Multiple well log properties are combined by performing principal component analysis and then these principal components are interpolated to all locations in the reservoir using sequential indicator simulation. Afterwards, classification based on mixture of Gaussian assumption (using Equations (5-2) and (5-3)) is performed in order to find the proportion of different facies at each location. These maps for the probability of different rock types given only the well data, are based on an extrapolation of the principal component of log attribute values away from the well location. Figure 5-12 shows the map of probability of being in each facies along the middle z-section.



(a) (b)

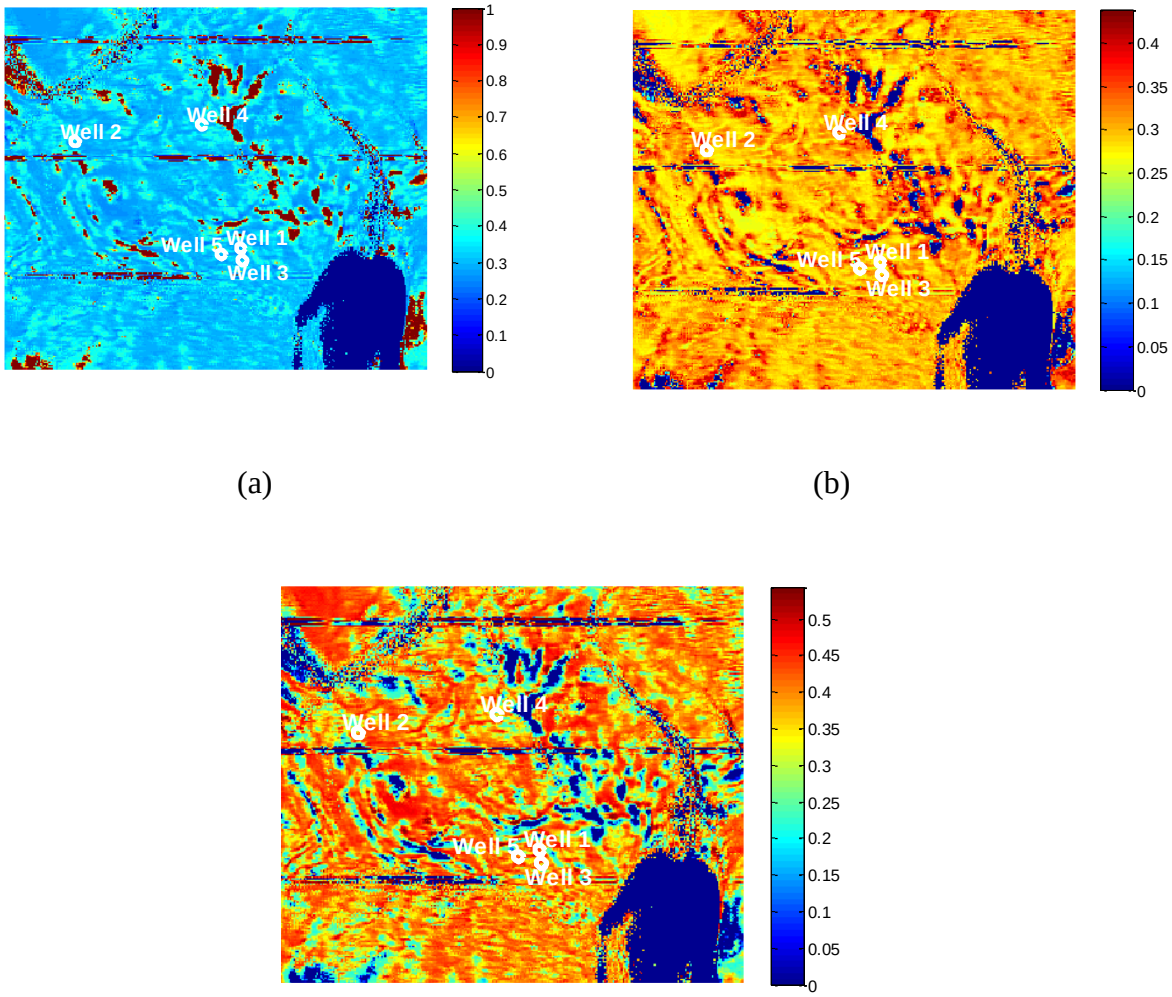


(c)

Figure 5-12. Probability map at the middle z-section obtained conditioned to the well log data. The map show the probability for (a) mudstone (b) grainstone (c) dolomitized rocks.

5.4.2 Probabilistic Classification of Facies Seismic attributes

The seismic attributes at any location are combined to obtain seismic PCs. These principal components are specified as input to the neural network in order to find the probability of each lithofacies at that location. This process was continued at other locations and Figure 5-13 a through c depict the probability map of each facies along the middle z-section.



(a) (b)

(c)

Figure 5-13. Probability map for each facies along the middle z-section given seismic data, (a) mudstone (b) grainstone (c) dolomitized rocks.

5.4.3 Merging Information

Subsequently, the probability of each lithofacies given log attributes is merged using permanence of ratio hypothesis. Since the value of the well log PCs were computed based on sequential indicator simulation at all locations, we used Equations (4-37) through (4-43) to

calculate the local tau at each location. Figure 5-14 shows the map of tau weight used for merging the probability maps along the middle z section.

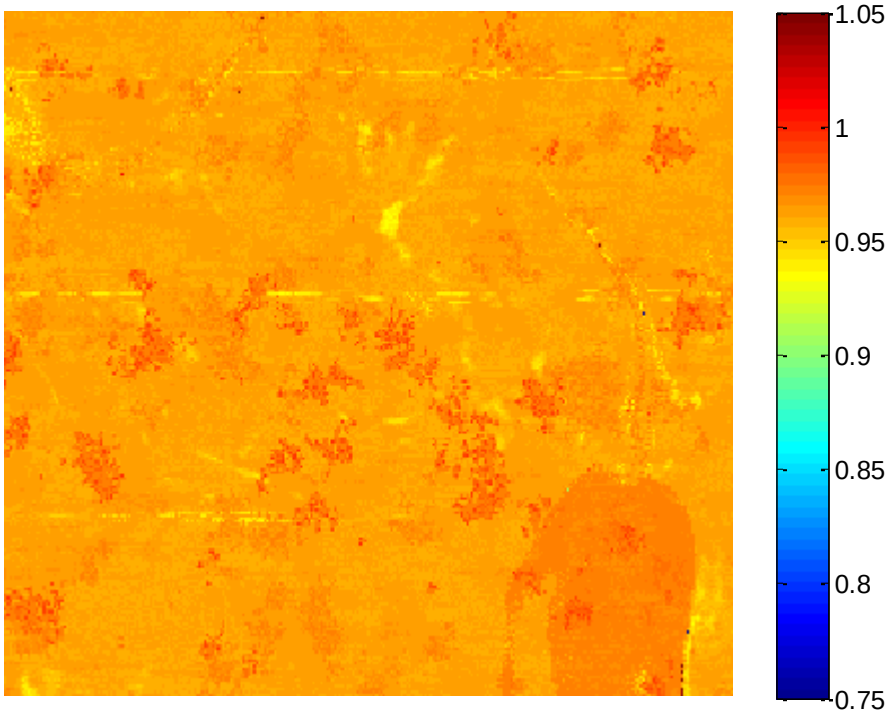


Figure 5-14. Local tau weight map used for merging the probability maps along the middle z section.

The tau map is used to merge the conditional probability distributions obtained from each source. Figure 5-15 through 5-17 show the merged probability map of each facies given all available data along the top, middle, and base z-sections respectively.

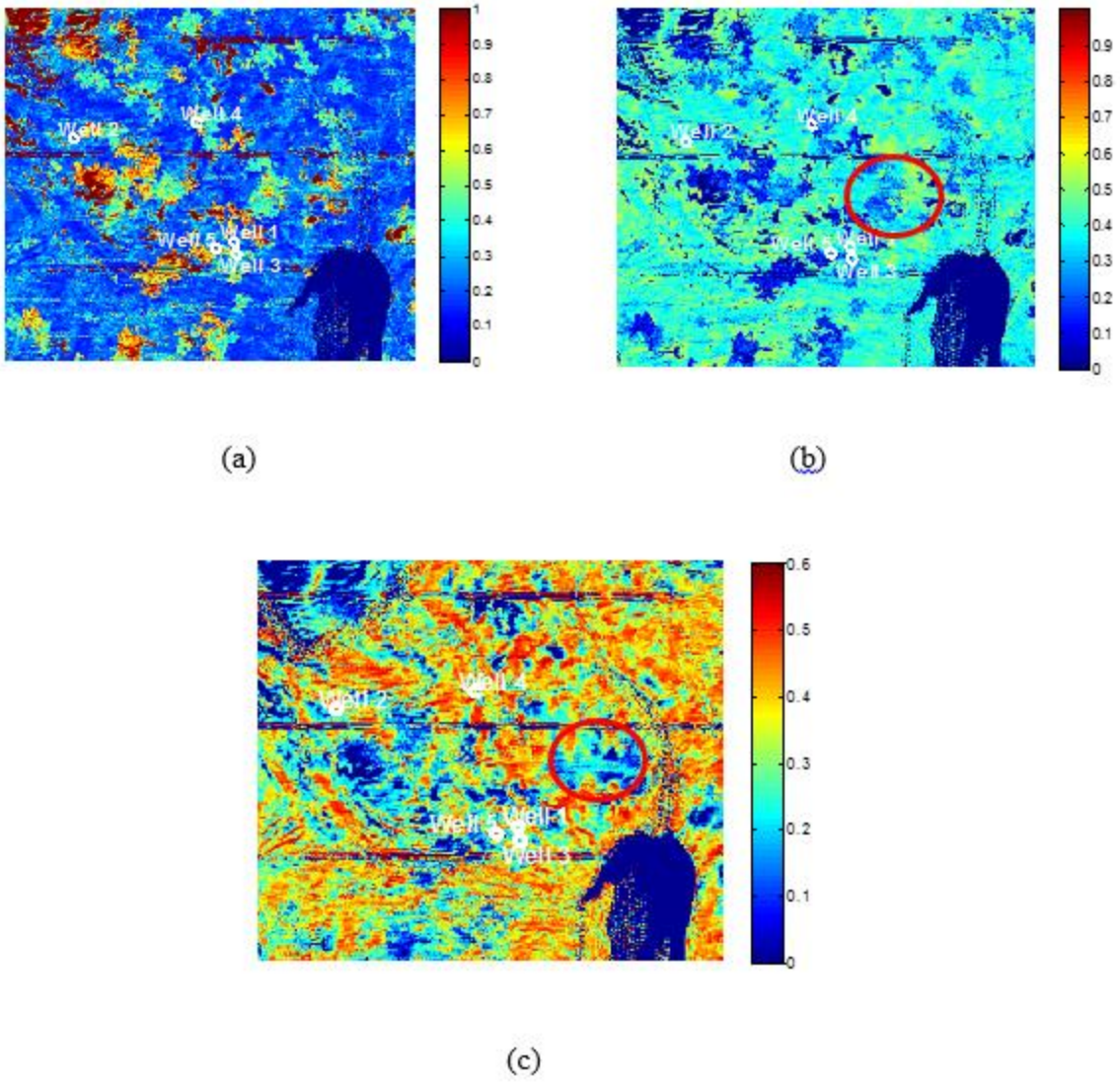


Figure 5-15. Merged probability map corresponding to the top z-section for (a) mudstone (b) grainstone (c) dolomitized rocks.

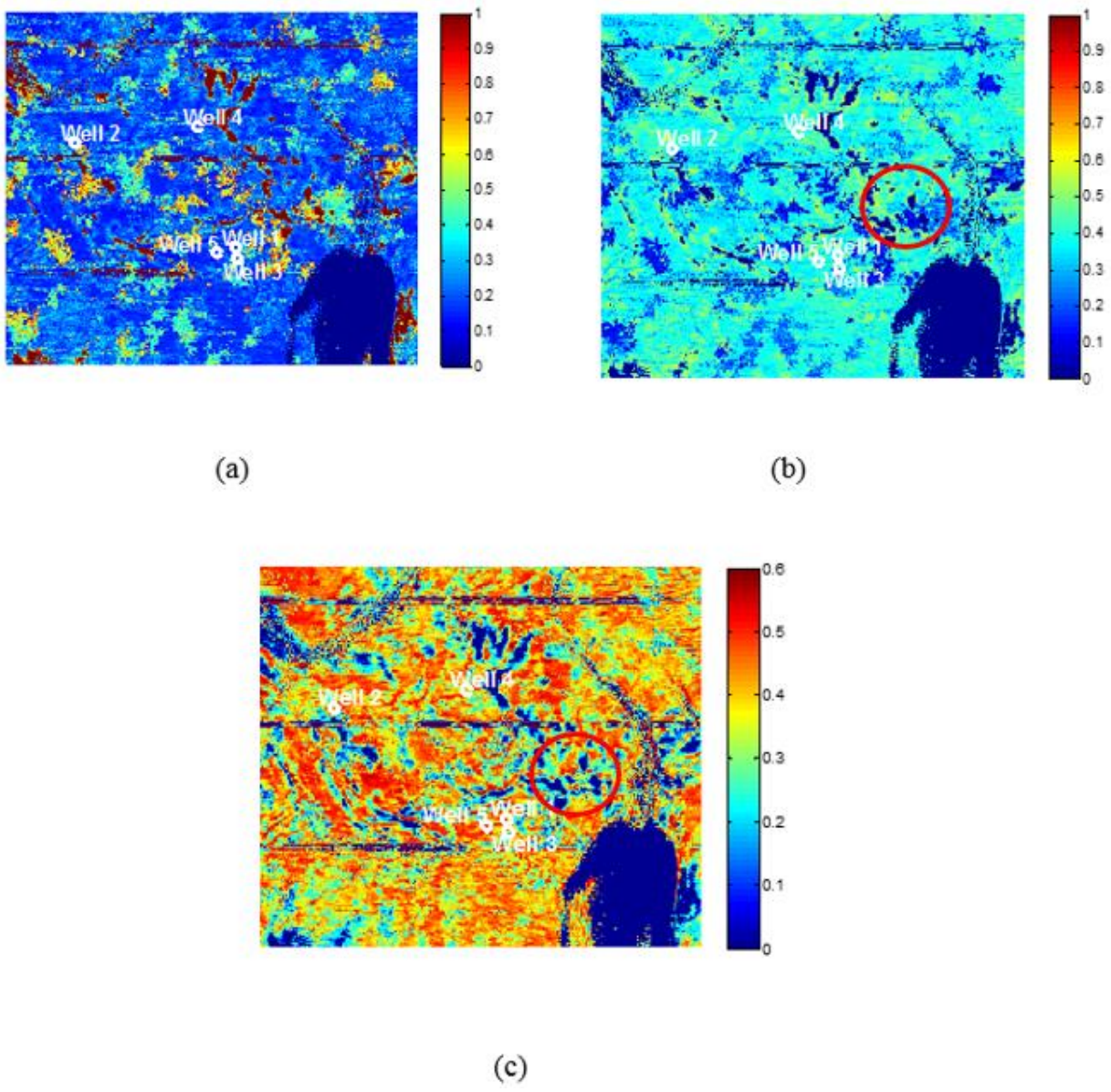


Figure 5-16. Merged probability map corresponding to the middle z-section for (a) mudstone (b) grainstone (c) dolomitized rocks.

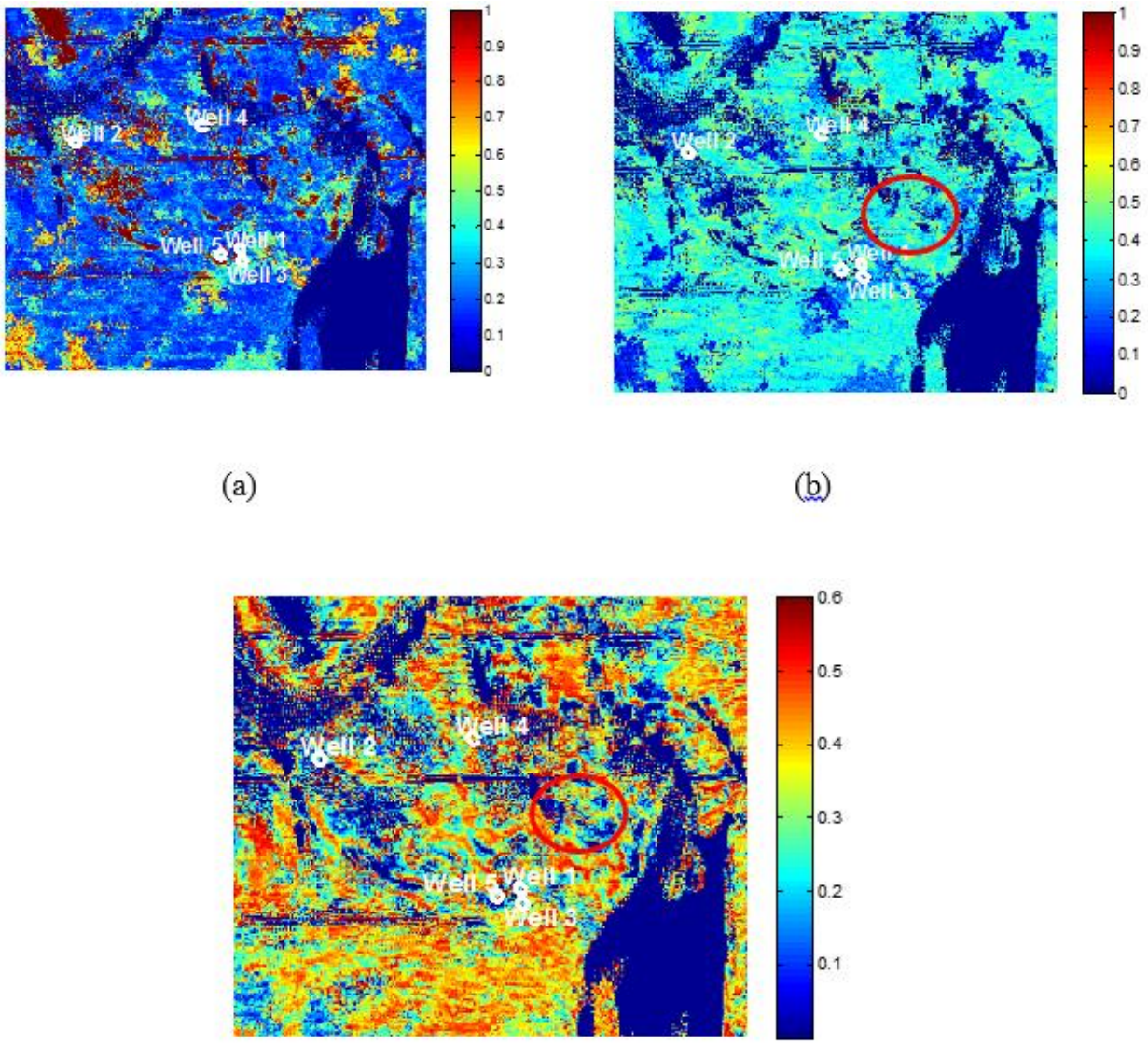


Figure 5-17. Merged probability map corresponding to the base z-section for (a) mudstone (b) grainstone (c) dolomitized rocks.

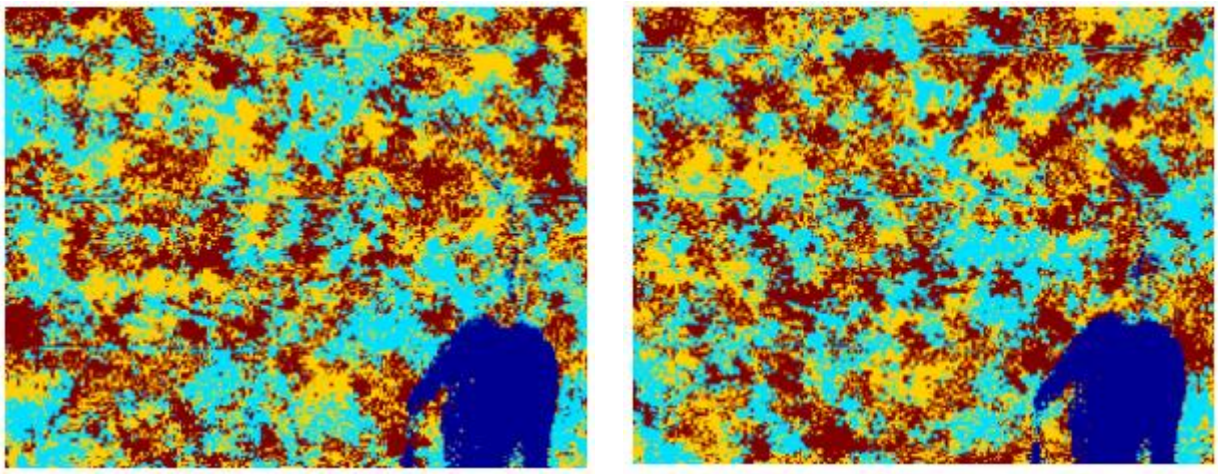
Dolomitization is indicative of chemical alterations in the rocks. Dolomitization in general tends to render the rock more brittle and hence prone to fracturing. In the case of fractured brecchia, dolomitization can also cause permeable flow pathways to bet blocked and hence reduce the volumetric sweep efficiency of processes implemented to recover the stored oil

and gas. As can be seen from the above Figures 5-15 through 5-17 show the proportion of different rock types along different z-sections. These maps can be used to decide the sweet spots for future drilling operations. Circled areas in the figures are examples of the location which have fairly high proportion of grainstone and small portion of dolomite.

5.4.4 Generating Multiple Realizations

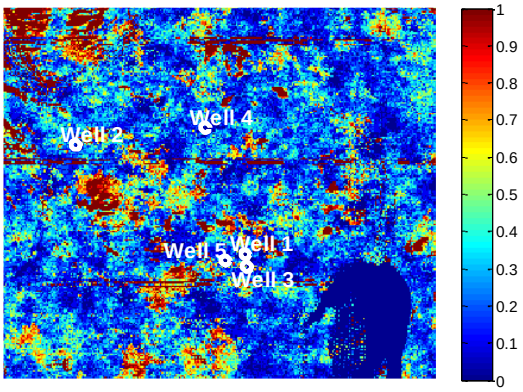
After the merged probabilities conditioned to both well log and seismic are available, multiple realizations are generated for assessing uncertainty. In order to perform the sampling from the merged conditional probability distribution, a correlated random number field was generated using the variogram model for the dominant lithofacies. The simulated random number is used to sample the facies at a location from the merged probability distribution. Figure 5-18 shows two different realizations along the middle z-section. The colors dark blue, light blue, yellow, and red correspond to salt, mudstone, grainstone, and dolomite respectively.

After generating 10 realizations, the probability of being in each facies can be calculated using the 10 realizations. The probability map generated in this manner is equivalent to the maximum aposteriori (MAP) obtained from the previous generated merged probability distributions. Figure 5-19 to 5-21 show the probability map obtained from ten realizations along three different z sections.

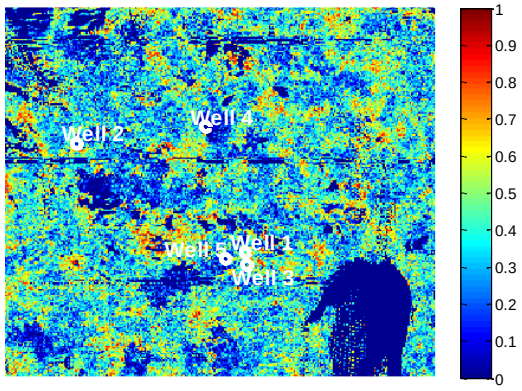


(a) (b)

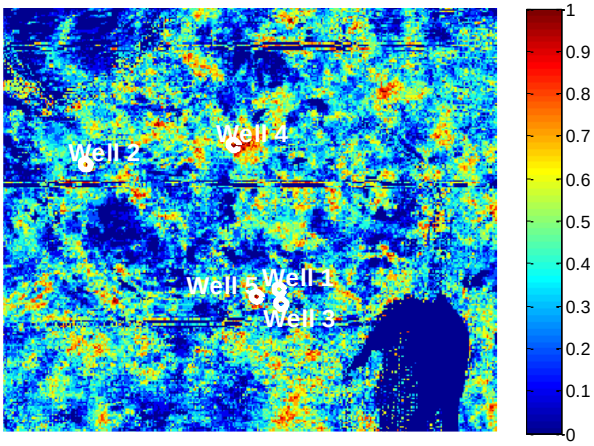
Figure 5-18. Two different realizations along the middle z-section. The colors dark blue, light blue, yellow, and red correspond to salt, mudstone, grainstone, and dolomite respectively.



(a)

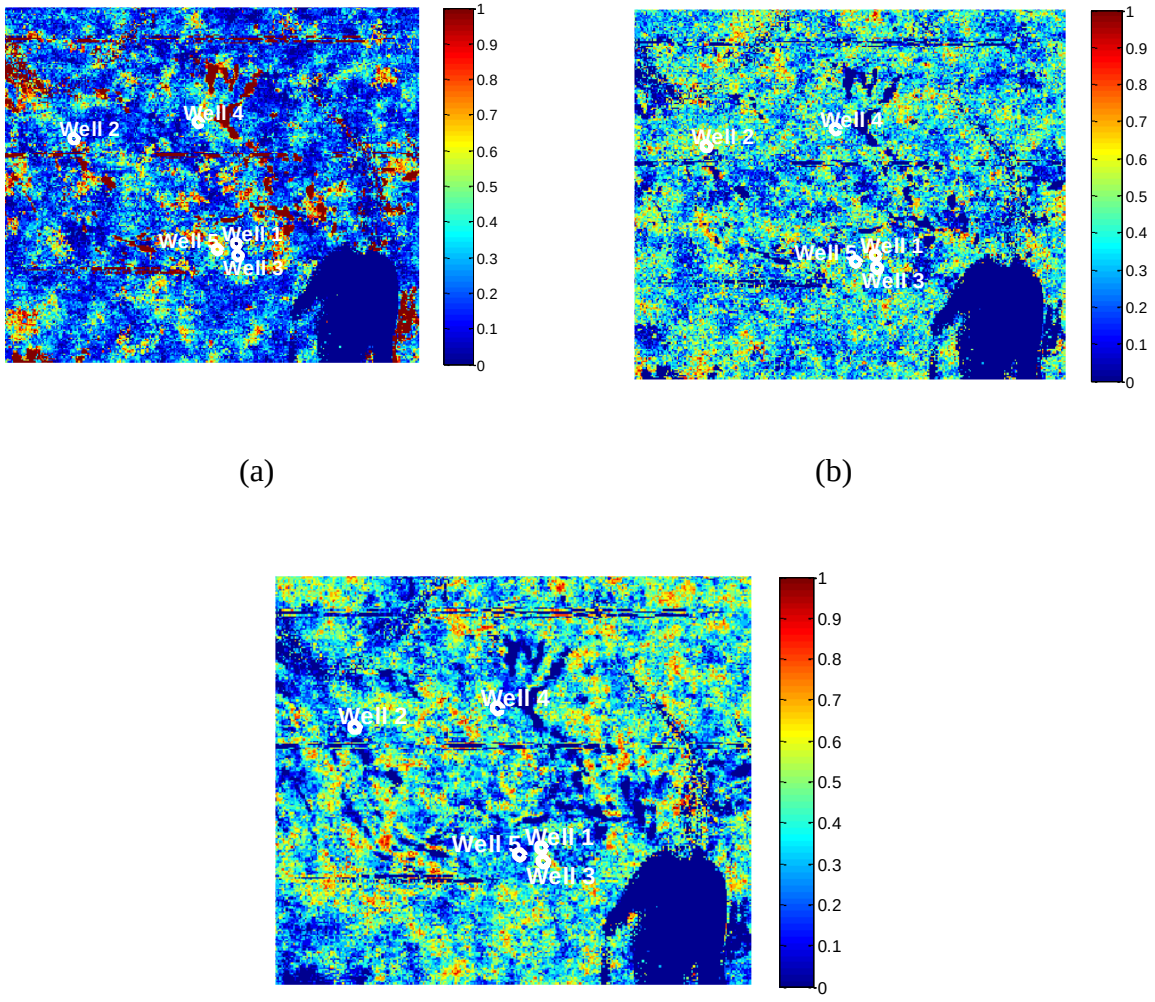


(b)



(c)

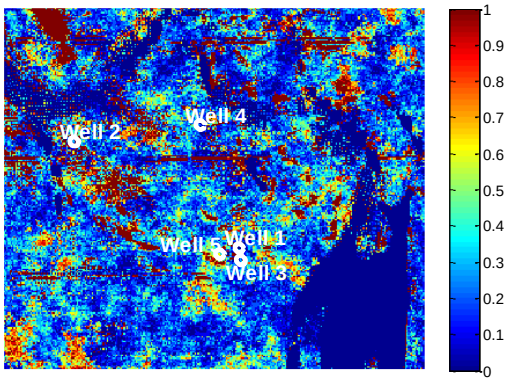
Figure 5-19. Probability map computed using 10 realizations corresponding to the top z-section for (a) mudstone (b) grainstone (c) dolomitized rocks.



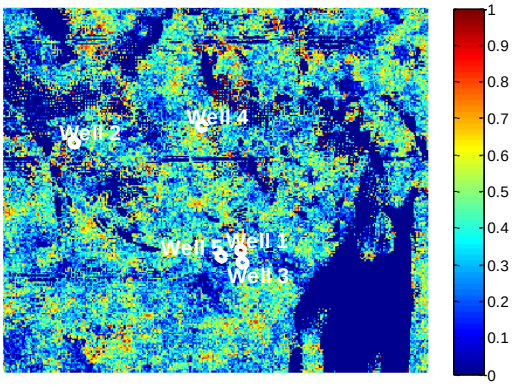
(a) (b)

(c)

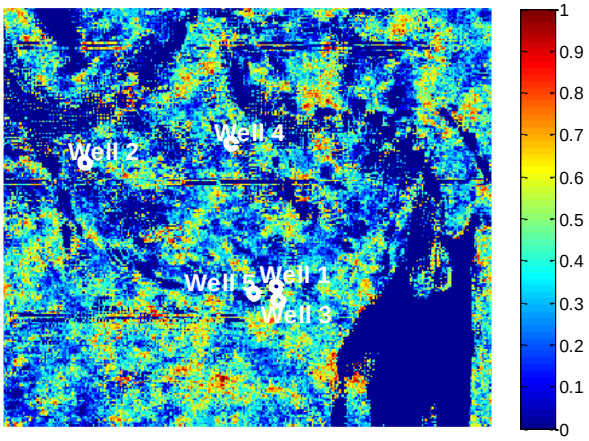
Figure 5-20. Probability map computed using 10 realizations corresponding to the middle z-section for (a) mudstone (b) grainstone (c) dolomitized rocks.



(a)



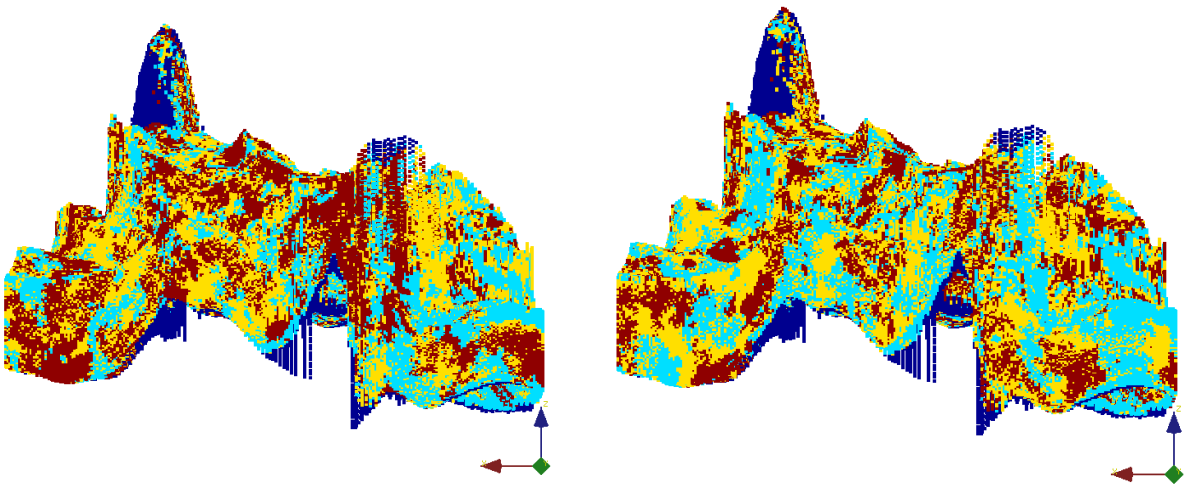
(b)



(c)

Figure 5-21. Probability map computed using 10 realizations corresponding to the base z-section for (a) mudstone (b) grainstone (c) dolomitized rocks.

After generating multiple realizations, we performed the inverse stratiographic transform to correctly account for the top and base horizons of the reservoir. Figure 5-22 shows the two realizations in Figure 5-18 after the inverse stratigraphic transform. Colors red, yellow, and light blue, and dark blue are representative of dolomitized rocks, grainstone, mudstone, and salt respectively.



(a) (b)

Figure 5-22. Two realizations sampled from the merged probability map using Monte Carlo sampling at a depth of 5331 m. These maps are shown after performing transformation into the original structure of the reservoir.

5.5 DISCUSSION OF RESULTS

There are two sets of probability maps for lithofacies, Figure 5-15 through 5-17 and Figure 5-19 through 5-21 that have been shown earlier. These exhibit slightly different characteristics. The maps obtained by direct plotting the probabilities obtained by the probability merging technique are smooth (similar to kriging maps) and it is easier to pinpoint locations with a significant proportion of a specific lithofacies such as dolomitized rock. However, the maps in Figure 5-19 to 5-21 correspond to proportions computed based on an ensemble of stochastic simulation realizations. They preserve the spatial structure as prescribed by variograms and they also preserve the variance or variability dictated by the data. It is harder to pinpoint candidate locations for future drilling based on the simulation maps but they do provide a good measure of uncertainty associated with any drilling decision.

In order to validate the results, it is necessary to see if the lithofacies maps generated have any bearing on the productivity of wells. Figure 5-23 shows the available production data

for a couple of wells in the reservoir. Based on comparing the well production and the proportion of lithofacies in the vicinity of wells, the following observations can be made.

Visually, looking at Figure 5-15, at a relative shallower depth, the Well 1 location lies in a region of smaller fraction of dolomitized rock and fairly high proportion of oolitic bank (grainstone) facies. At that depth, the Well 2 also lies in a region with similar characteristics. This explains the similar characteristics of both these wells in Interval I. Similar characteristics are also observed at the middle depth. However, at the deepest interval, the Well 2 is in presence of high proportion of dolomitized facies while Well 1 has much smaller fraction of dolomitized rock. This might explain the difference in production characteristics of the wells corresponding to Interval III.

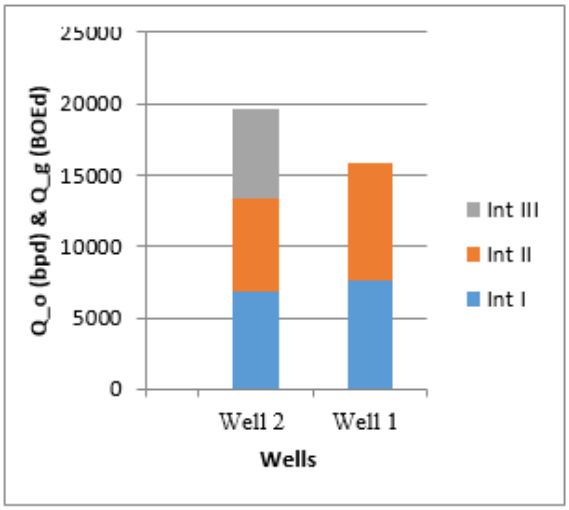
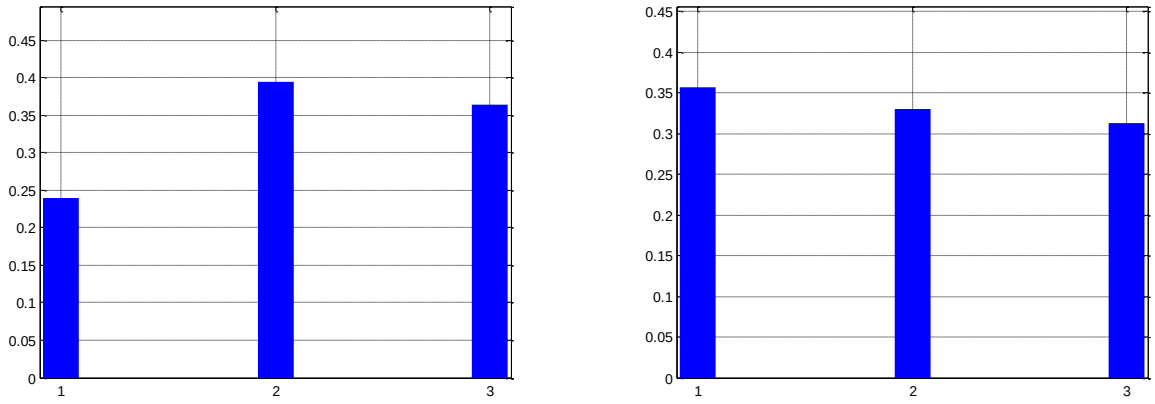


Figure 5-23. Oil and gas production per day at Well 1 and Well 2.

The variations in facies proportion in the vicinity of Well 1 and Well 2 are shown in Figure 5-24. It can be seen that both these wells are completed in regions with an abundance of oolitic bank facies (Facies code 2). Well 1 has a higher fraction of dolomitized facies (Facies

code 3) compared to Well 2. This might explain the slightly lower productivity of Well 1 compared to Well 2.



(a) (b)

Figure 5-24. Facies proportion in a small volume around (a) Well 1, (b) Well 2.

5.6 SUMMARY AND CONCLUSIONS

In this chapter, we implemented the permanence of ratio hypothesis to characterize a carbonate reservior in the Gulf of Mexico. The available data sets were drill cutting data, core data, well log measurements and 3D seismic volume. We used drilling cutting data to calibrate seismic attributes with lithofacies. We used core data to calibrate log measurements to lithofacies. Then, we merged the probability maps of lithofacies using permanence of ratio hypothesis. The final probability map is used for generating multiple realizations to evaluate the uncertainty associated with the prediction of lithofacies. The map obtained by sampling from the merged probability distributions exhibit more variability but also exhibit the spatial correlation expected in oolitic bank reservoirs. Finally, we explored the viability of the results by investigating the production data of two wells in the region and relating that to the proportion of lithofacies in the vicinity of the wells.

Chapter 6: Conclusions and Recommendations for Future Work

6.1 SUMMARY AND CONCLUSION

One of the most challenging issues in reservoir modeling is to integrate information/data obtained from different sources at different scales and precision. In general, the primary data are borehole measurements, but in general, these are too sparse to construct accurate reservoir models. Therefore, in most cases, the information from borehole measurements have to be supplemented with secondary data. The secondary data for reservoir modeling could be static data such as seismic data or dynamic data such as production history or well test data.

Most of the time, secondary data such as seismic data that provide indirect information about spatial variation of reservoir attributes are available exhaustively. A method by which knowledge/information obtained from these different sources can be combined is therefore required so that models for spatial variability of reservoir properties can be constructed reliably. It is shown in several studies that soft data integration can tremendously improve the accuracy of the reservoir model and decrease the uncertainty associated with reservoir performance predictions.

In this thesis, we present two different algorithms for integrating static secondary data such as seismic data. The first algorithm is an extension of the direct sampling-based multiple-point statistics method. We presented a methodology for integrating secondary soft data in that framework. The algorithm is based on direct pattern search over an ensemble of realizations. We showed that the proposed methodology is sutiable for modeling complex channelized reservoirs and reduces the uncertainty associated with production performance due to integration of secondary data.

The second algorithm presented in this study is based on the permanence of ratio hypothesis. The permanence of ratio hypothesis is a novel method for data integration which was proposed by Journel in 2002. The problem of integrating data from different sources can be posed as finding the conditional probability distribution of $P(A|D_1, D_2, \dots D_n)$ where A is the event being estimated and D_is are the data from different sources. In order to model the conditional probability, it would be convenient if the information from each data source can be assessed independently in order to find $P(A|D_i)$, and then merge these joint probabilities to calculate $P(A|D_1, D_2, \dots D_n)$ accounting for the redundancy between different data sources. This is precisely the premise of the permanence of ratio hypothesis. An odds-ratio measuring the relative distance to the occurrence of A given D_i is computed. The permanence of ratio hypothesis then states that the merging of the conditional probabilities from the individual sources of information can be accomplished using odd ratios computed for each source assuming that the ratio computed corresponding to one source of information remains invariant regardless of the occurrence of another data event. The redundancy between the information from different sources is subsequently accounted for using parameters (tau or nu parameters, Krishnan, 2004).

The tau model (Krishnan, 2008) can accommodate any complex description of data redundancy and an exact expression for the tau parameters can be derived from a tautology. However, techniques for practical calibration of tau weights from the statistics of data are lacking, which makes the theory difficult to apply. In this thesis, we derived a practical expression for the tau parameters and demonstrate the procedure for calibrating these parameters using the available data.The expression derived in this thesis is based on the mixture of Gaussian assumption i.e. we assumed that petrophysical properties of rocks with each rock type is normally distributed, and the overall properties have multi-modal normal distribution (each mode

corresponds to one rock type). Eventhough normal distribution has been proposed and implemented in this thesis, any other parametric distribution such as log-normal can also be used.

We presented a simple synthetic 1-D example to demonstrate how the proposed method can be used to integrate the secondary data. Then, we implemented the proposed methodology in a 2-D section of a synthetis 3-D reservoir (Stanford V) and showed that the proposed method has the highest accuracy and the lowest uncertainty among other integration methods.

We also presented a case study where permanence of ratio is used for modeling a carbonate reservoir in the Gulf of Mexico. We showed that the method has a better performance than when primary hard and secondary soft data are used within the traditional geostatistical framework.

6.2 FUTURE WORK

The following recommendations are suggested for future studies:

1- The algorithm presented in Chapter 3, is a pattern search algorithm and does not have a robust theoretical background underlying its implementation. Future studies to establish the theoretical basis for the algorithm and its relation to the permanence of ratio hypothesis is recommended.

2- The implementation of the permanence of ratio and the tau weight expression is only done for integrating static secondary data into reservoir models. Future studies to derive an expression for the redundancy factor in the case of dynamic secondary data, e.g. to perform history matching, is recommended. This would vastly expand the capabities of current popular schemes such as the probability perturbation method.

3- The tau expression derived in chapter 4, is based on a mixture of Gaussian

assumption, and might not be valid in some cases. Therefore, deriving a more robust methodology for calculating the tau weight, preferrably a non-parametric algorithm, is suggested.

Appendix: Training Algorithm of GF Network

The Generalized Fisher (GF) training algorithm is based on the Expectation Maximization (EM) method described in Streit & Luginbuhl (1994). The EM method consists of two steps: The first step is called the expectation step or E-step, and extends the likelihood function to the unobserved or missing data, and computes an expectation over the missing data to obtain an auxiliary function Q. The second step that is called Maximization step (M-step) maximizes the function Q with respect to the parameter set to be estimated.

Suppose independent samples of a random vector X with dimension N are observed, where X is described by a mixture of conditional PDF given by:

$$X \sim \sum_{i=1}^{M} \alpha_i g_i(X|\lambda_i) \tag{A.1}$$

Where

$$\sum_{j=1}^{N} \alpha_j = 1 \tag{A.2}$$

$$g_i(X|\lambda_i) = \sum_{j=1}^{G_i} \frac{\pi_{ij}}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left(\frac{-1}{2}(X-\mu_{ij})^t \Sigma^{-1}(X-\mu_{ij})\right) \tag{A.3}$$

$$\sum_{j=1}^{G_i} \pi_{ij} = 1 \tag{A.4}$$

Where G_i is the number of the sample within the i[th] class, and the parameter that need to be estimated are μ_{ij}, Σ, π_{ij} and α_j such that the following function is maximized (Streit & Luginbuhl, 1994).

$$Q = \sum_{j=1}^{M} G_j \log \alpha_j$$

$$+ \sum_{j=1}^{M} \sum_{i=1}^{G_j} \sum_{k=1}^{G_j} \acute{\omega}_{ij}(X_{kj}) \log \pi_{ij} \tag{A.5}$$

$$+ \sum_{j=1}^{M} \sum_{i=1}^{G_j} \sum_{k=1}^{G_j} \acute{\omega}_{ij}(X_{kj}) \log p_{ij}(X_{kj})$$

Where

$$\acute{\omega}_{ij}(X_{kj}) = \frac{\pi_{ij} \exp(\frac{-1}{2}(X_{kj} - \mu_{ij})^t \Sigma^{-1}(X_{kj} - \mu_{ij}))}{\sum_{l=1}^{G_j} \pi_{lj} \exp(\frac{-1}{2}(X_{kj} - \mu_{lj})^t \Sigma^{-1}(X_{kj} - \mu_{lj}))} \tag{A.6}$$

$$p_{ij}(X_{kj}) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp(\frac{-1}{2}(X_{kj} - \mu_{ij})^t \Sigma^{-1}(X_{kj} - \mu_{ij})) \tag{A.7}$$

Once the optimization problem is solved, the parameters are calculated as follows

$$\alpha_i = \frac{G_i}{G} \tag{A.8}$$

The other parameters are calculated through the following iterative algorithm. Now let $\lambda^{(n)}$ is the set of parameters at nth iteration, and then the weights are defined by

$$\acute{\omega}_{ij}{}^{(n)}(X_{kj}) = \frac{\pi_{ij}{}^{(n)} \exp(\frac{-1}{2}(X_{kj} - \mu_{ij}{}^{(n)})^t \Sigma^{-1}(X_{kj} - \mu_{ij}{}^{(n)}))}{\sum_{l=1}^{G_j} \pi_{lj}{}^{(n)} \exp(\frac{-1}{2}(X_{kj} - \mu_{lj}{}^{(n)})^t \Sigma^{-1}(X_{kj} - \mu_{lj}{}^{(n)}))} \tag{A.9}$$

The new intercomponent mixing properties, mean vector, and covariance matrix are updated as follows:

$$\pi_{ij}{}^{(n+1)} = \frac{1}{G_j} \sum_{k=1}^{G_j} \acute{\omega}_{ij}{}^{(n)}(X_{kj}) \tag{A.10}$$

$$\mu_{ij}{}^{(n+1)} = \frac{\sum_{k=1}^{G_j} \acute{\omega}_{ij}{}^{(n)}(X_{kj})X_{kj}}{\sum_{k=1}^{G_j} \acute{\omega}_{ij}{}^{(n)}(X_{kj})} \tag{A.11}$$

$$\Sigma^{(n+1)} = \frac{1}{G}\sum_{j=1}^{M}\sum_{i=1}^{G_j}\sum_{k=1}^{G_j} \acute{\omega}_{ij}{}^{(n)}(X_{kj})(X_{kj} - \mu_{ij}{}^{(n+1)})(X_{kj} - \mu_{ij}{}^{(n+1)})^t \tag{A.12}$$

The convergence of algorithm is tested by comparing $Q^{(n+1)}$ to $Q^{(n)}$ to determine if the parameters have stabilized. If the parameters have stabilized, then the algorithm is terminated.

References

Andersen, T., Zachariassen, E., Otterlei, C., Hatland, K., & Liestolt, F. (2006). Method for conditioning the reservoir model on 3D and 4D elastic inversion data applied to a fluvial reservoir in the North sea. *SPE 100190* .

Caers, J., & Srinivasan, S. (2003). Combining geological information with seismic and production data. *Developments in petroleum science , 51,* 499-525.

Caers, J., & Srinivasan, S. (2002). Statistical pattern recognition and geostatistical data integration. *Soft Computing for Reservoir Characterization and Modeling. Physica-Verlag HD ,* 355-386.

Caers, J., Hoffman, T., Strebelle, S., & Wen, X.-H. (2006). Probabilistic integration of geologic scenarios, seismic, and production data—a West Africa turbidite reservoir case study. *The Leading Edge , 25* (3), 240-244.

Castro, S., Caers, J., Otterlei, C., Andersen, T., Hoye, T., & Gomel, P. (2006). A Probabilistic Integration of Well Log Geological Information 3D/4D Seismic and Production Data: Application to the Oseberg Field. SPE Annual Technical Conference and Exhibition.

Daly, C. (2004). Higher order models using entropy, Markov random fields and sequential simulation. *Geostatistics Banff.* Banff, Alberta.

Doyen, P. M. (1988). Porosity from seismic data: A geostatistical approach. *Geophysics* , 1263-1275.

Doyen, P. (2007). Seismic reservoir characterization: An earth modelling perspective. *EAGE.*

Dubrule, O. (2003). Geostatistics for Seismic Data Integration in Earth Models. *Society of Exploration Geophysicists (SEG).*

Dubrule, O. (2003). *Geostatistics for seismic data integration in Earth models: 2003 Distinguished Instructor Short Course.* SEG Books.

Eidsvik, J., Avseth, P., Henning, O., Mukerji, T., & Mavko, G. (2004). Stochastic reservoir characterization using prestack seismic data. *Geophysics , 69* (4), 978-993.

Eskandari, K., & Srinivasan, S. (2007). Growthsim–A Multiple Point Framework for Pattern Simulation. EAGE Petroleum Geostatistics.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* Oxford University Press.

Haas, A., & Dubrule, O. (1994). Geostatistical inversion-a sequential method of stochastic reservoir modelling constrined by seismic data. *first break* .

Hampson, D. P., Schuelke, J. S., & Quirein, J. A. (2001). Use of multiattribute transforms to predict log properties from seismic data. *Geophysics , 66* (1), 220-236.

Huang, Y.-C., & Srinivasan, S. (2012). Efficient conditional simulation of spatial patterns using a pattern-growth algorithm. Geostatistics Oslo 2012.

John, A. K., Lake, L. W., Torres-Verdin, C., & Srinivasan, S. (2008). Seismic facies identification and classification using simple statistics. *SPE Reservoir Evaluation & Engineering* , 984-990.

Journel, A. (2002). Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses. *Mathematical geology , 34* (5), 573-596.

Journel, A. (1986). Constrained interpolation and qualitative information—the soft kriging approach. *Mathematical Geology , 18* (3), 269-286.

Kashib, T., & Srinivasan, S. (2006). A probabilistic approach to integrating dynamic data in reservoir models. *Journal of petroleum science and engineering , 50* (3), 241-257.

Krishnan, s. (2008). The Tau Model for Data Redundancy and Information Combination in Earth Sciences: Theory and Application. *Mathematical Geosciences , 40* (6), 705-727.

Krishnan, S., Boucher, A., & Journel, A. G. (2004). Evaluating information redundancy through the tau model. Geostatistics Banff.

Lindseth, R. O. (1979). Synthetic sonic logs- a process for stratigraphic interpretation. *Geophysics , 44* (1), 3-26.

Mao, S., & Journel, A. G. (1999). Generation of a reference petrophysical and seismic 3D data set: the Stanford V reservoir. Proceedings of the Stanford Center for Reservoir Forecasting Annual Meeting.

Mariethoz, G. (2010). A general parallelization strategy fro random path based geostatistical simulation methods. *Computers and Geoscience* , 953-958.

Mariethoz, G., Renard, P., & Straubhaar, J. (2010). The Direct Sampling method to perform multiple-point geostatistical simulations. *Water Resources Research , 46* (11).

Mariethoz, G., Renard, P., Cornaton, F., & Jaquet, O. (2009). Truncated plurigaussian simulations to characterize aquifer heterogeneity. *Groundwater , 47* (1), 13-24.

McEliece, R. (2002). *The Theory of information and coding: a mathematical framework for communication.* Cambridge University Press.

Riddiford, F. A., & Goupillot, M. (1994). Geotechnical Integration and its Impact on Field Management. The IRMA Methodology. *SPE Annual Technical Conference and Exhibition.*

Russell, B., Hampson, D., & Todorov, T. (2001). Combining geostatistics and multiattribute transforms–A channel sand case study. *71st Annual International Meeting, SEG*, (pp. 638-641).

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal , 27*, 379-423.

Specht, D. F. (1990). Probabilistic Neural Networks. *Neural Networks , 3*, 109-118.

Srinivasan, S., & Sen, M. K. (2010). Mapping of diagenesis in a carbonate reservoir in the Gulf of Mexico by a stochastic data integration technique. *2010 SEG Annual Meeting.* Society of Exploration Geophysicists.

Srinivasan, S., & Sen, M. K. (2010). Mapping of diagenesis in a carbonate reservoir in the Gulf of Mexico by a stochastic data integration technique. Denver: SEG Annual Meeting.

Srinivasan, S., & Sen, M. (2009). Stochastic modeling of facies distribution in a carbonate reservoir in the Gulf of Mexico. *Geohorizons* .

Straubhaar, J., Walgenwitz, A., & Renard, P. (2013). Parallel multiple-point statistics algorithm based on list and tree structures. *Mathematical Geosciences , 45* (2), 131-147.

Strebelle, S. (2002). Conditional Simulation of Complex geological structures using multiple-point statistics. *Mathematical Geology* , 1-22.

Strebelle, S., Payrazyan, K., & Caers, J. (2003). Modeling of a Deepwater Turbidite Reservoir Conditional to Seismic Data Using Principal Component Analysis and Multiple-Point Geostatistics. *SPE Journal , 8* (3), 227-235.

Streit, R. L., & Luginbuhl, T. E. (1994). Maximum likelihood training of probabilistic neural networks. *IEEE Transaction on Neural Networks , 5* (5), 764-783.

Zhang, R., Sen, M. K., & Srinivasan, S. (2012). A prestack basis pursuit seismic inversion. *Geophysics* , R1-R11.