

Srikanta Mishra | Akhil Datta-Gupta

Applied Statistical Modeling and Data Analytics

A Practical Guide for the Petroleum Geosciences

Applied Statistical Modeling and Data Analytics

A Practical Guide for the Petroleum Geosciences

The essential reference for geoscientists and engineers who apply statistical modeling and data analytics techniques to solve petroleum industry challenges

- Authored by internationally renowned experts in developing and applying statistical methods for oil and gas and other subsurface problem domains
- Written by practitioners for practitioners
- Presents an easy to follow narrative that progresses from simple concepts to more challenging ones
- Includes online resources with software applications and practical examples for the most relevant and popular statistical methods, using data sets from the petroleum geosciences
- Addresses the theory and practice of statistical modeling and data analytics from the perspective of petroleum geoscience applications

Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geosciences

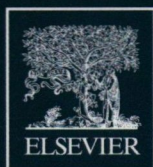
provides a practical guide to many of the classical and modern statistical techniques that have become established for oil and gas professionals in recent years. It serves as a "how-to" reference volume for the practicing petroleum engineer or geoscientist interested in applying statistical methods in formation evaluation, reservoir characterization, reservoir modeling and management, and uncertainty quantification.

Beginning with a foundational discussion of exploratory data analysis, probability distributions, and linear regression modeling, the book focuses on fundamentals and practical examples of such key topics as multivariate analysis, uncertainty quantification, data-driven modeling, and experimental design and response surface analysis. Datasets from the petroleum geosciences are extensively used to demonstrate the applicability of these techniques. The book will also be useful for professionals dealing with subsurface flow problems in hydrogeology, geologic carbon sequestration, and nuclear waste disposal.

Dr. Srikanta Mishra is Institute Fellow and Chief Scientist (Energy) at Battelle Memorial Institute, where he leads computational modeling and data analytics activities for geologic carbon storage, shale gas and oil development, and improved oil recovery projects. He joined Battelle in 2010 after a distinguished 20+ year career in geosystems consulting and applied research, including an appointment as adjunct professor of Petroleum Engineering at the University of Texas at Austin. He holds a PhD degree from Stanford University, an MS degree from University of Texas and a BTech degree from Indian School of Mines – all in petroleum engineering.

Dr. Akhil Datta-Gupta is Regents Professor, University Distinguished Professor, and Peterson '36 Chair in petroleum engineering at Texas A&M University. He directs the "Model Calibration and Efficient Reservoir Imaging" Joint Industry Project carrying out research on statistical modeling, multiphase flow simulation and inverse modeling of reservoir data. Dr. Datta-Gupta joined Texas A&M in 1994 after a brief industry career and was elected to the US National Academy of Engineering in 2012. He holds PhD and MS degrees from University of Texas and a BTech degree from Indian School of Mines – all in petroleum engineering.

Drs. Mishra and Datta-Gupta have published extensively on the topics covered in this book and have taught short courses at professional society meetings and client locations all over the world.



elsevier.com/books-and-journals

ISBN 978-0-12-803279-4



9 780128 032794

Mishra Gupta
Applied Statistical Modeling and Data Analysis
HVS

APPLIED STATISTICAL MODELING AND DATA ANALYTICS

A Practical Guide for the
Petroleum Geosciences

SRIKANTA MISHRA

Battelle Memorial Institute, Columbus, Ohio, USA

AKHIL DATTA-GUPTA

Texas A&M University, College Station, Texas, USA



Elsevier

Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands

The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

© 2018 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN: 978-0-12-803279-4

For information on all Elsevier publications
visit our website at <https://www.elsevier.com/books-and-journals>



Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Publisher: Candice Janco

Acquisition Editor: Amy Shapiro

Editorial Project Manager: Tasha Frank

Production Project Manager: Anitha Sivaraj

Cover Designer: Christian J. Billow

Typeset by SPi Global, India

APPLIED STATISTICAL MODELING AND DATA ANALYTICS

Dedication

This book is dedicated to

*our **parents**, for teaching us the letters and giving us the spirit;
our **spouses**, for sustaining the present with love and understanding;
and our **children**, for continuing the quest of knowledge.*

न हि ज्ञानेन सदृशं पवित्रमिह विद्यते ।

na hi jnaanena sadrsham pavitram iha vidyate

There is nothing as pure as knowledge.

Bhagavad Gita, IV.38

Preface

As practicing petroleum engineers and geoscientists, we have long been fascinated by the myriad applications of statistical methods in characterizing, monitoring, and forecasting the behavior of subsurface fluid-bearing geosystems. Statistics has been a quintessential tool in our arsenal for fundamental tasks such as describing and modeling data from well logs, core samples, and injection/production tests. It has improved our designs of laboratory, field, and numerical experiments to understand the relationship between geologic inputs (e.g., porosity, permeability, and well-log attributes) and state variables (e.g., pressure and production rates). Statistical techniques have also helped improve estimates of uncertainty in static and dynamic reservoir model predictions arising from the underlying model input uncertainty. Our journey has taken us from basic exploratory data analysis and regression modeling to more advanced multivariate analysis, nonlinear and nonparametric regression modeling, experimental design and response surface analysis, and uncertainty quantification methods. Lately, we have been intrigued by the promise of big data analytics for oil and gas projects based on its success in other problem domains and have been exploring machine-learning techniques to develop data-driven insights for understanding and optimizing the performance of petroleum reservoirs.

During this “random walk” along the road of applied statistics, we have studied many research papers on the subject from fellow petroleum engineers and geoscientists and have also contributed to the growing literature. We have consulted several books on statistical modeling and data analytics that target both specialized and broad audiences. Through serendipity, we have come to realize that there is no single textbook or reference volume currently in the market that addresses the theory and practice of these topics from the perspective of petroleum engineering or geoscience applications.

This book is our humble attempt to fill the void. It seeks to provide a practical guide, via theoretical background and practical examples, to many of the classical and modern statistical techniques that have become, or are becoming, mainstream for oil and gas professionals. It is intended to serve as a “how to” reference for the practicing petroleum engineer or geoscientist interested in applying statistical modeling and data analytics techniques in formation evaluation, reservoir characterization, reservoir modeling and management, and production operations.

This is a book on the application of statistics, written by practitioners, for practitioners. As such, we have tried to strike a judicious balance between statistical rigor and formalism and practical considerations regarding the fundamentals and applicability of various relevant concepts. Beginning with a foundational discussion of exploratory data analysis (Chapter 2), probability distributions (Chapter 3), and linear regression modeling (Chapter 4), the book focuses on fundamentals and practical examples of such key topics as multivariate analysis (Chapter 5), uncertainty quantification (Chapter 6), experimental design and response surface

analysis (Chapter 7), and data-driven modeling (Chapter 8). Datasets from the petroleum geosciences are extensively used to demonstrate the applicability of these techniques. We have chosen not to discuss topics related to geostatistics or time series analysis, as there are several excellent practical references available on the subject.

Although this book is primarily organized in the form of a ready reference guide for practitioners in the petroleum geosciences, it can also be used as a textbook for an upper division or graduate-level course on the subject. To that end, we have added several pedagogical examples and exercise problems to each chapter. The book will also be useful for professionals dealing with subsurface flow problems in hydrogeology, geologic carbon sequestration, and nuclear waste disposal. The material in the book has been collated from class notes of graduate and undergraduate courses that we have taught and short courses and workshops that we have offered at professional society meetings and client locations.

We view statistical modeling and data analytics as both an established and an emerging field, where basic concepts from classical statistics provide the building blocks for applying newly developed algorithms from the computer and data science domains. We hope this book will empower petroleum engineers and geoscientists with a greater appreciation of relevant principles and tools for converting data into information—particularly the actionable kind that lead to better decisions.

Acknowledgments

We wish to express our gratitude and appreciation to

- our professors at Stanford University, The University of Texas at Austin, and Indian School of Mines—who taught us how to formulate and solve problems of practical significance; shared with us their knowledge, experience, and insights; and inspired us with their scholarship and professionalism;
- our students in university classes and professional short courses on statistical modeling and uncertainty analysis—who beta tested the class notes and slides that morphed into this book and frequently challenged us to refine our understanding of the subject matter;
- our employers Battelle Memorial Institute and Texas A&M University and multiple governmental and industrial sponsors—who provided us with the resources for carrying out much of the research that culminated in this book;
- our colleague Jared Schuetter—who contributed significantly to Chapters 7 and 8; and helped improve our understanding of big data analytics;
- our colleagues Banda RamaRao, Neil Deeds, David Sevougian, and Mike King—who collaborated in many enriching topical discussions and contributed to this book through concepts, text, and figures;
- our colleagues Shuvajit Bhattacharya and Isis Fukai and advisees Atsushi Iino, Changdong Yang and Aditya Vyas—who reviewed the draft manuscript from the perspective of practicing petroleum engineers and geoscientists and provided many helpful comments;
- our Elsevier editors Tasha Frank and Louisa Hutchins and Production Manager Anitha Sivaraj—who guided us through the proposal development, manuscript submission, and book production processes with patience; and last, but not the least,
- the Mishra family (Snigdha, Ananya, and Anshuman) and the Datta-Gupta family (Mausumi, Alina, and Antara) who indulged this passion of ours by willingly giving us the gift of time.

Contents

Preface ix

Acknowledgments xi

1. Basic Concepts

- 1.1 Background and Scope 1
- 1.2 Data, Statistics, and Probability 6
- 1.3 Random Variables 10
- 1.4 Summary 12
- References 13

2. Exploratory Data Analysis

- 2.1 Univariate Data 15
- 2.2 Bivariate Data 22
- 2.3 Multivariate Data 26
- 2.4 Summary 28
- References 28

3. Distributions and Models Thereof

- 3.1 Empirical Distributions 31
- 3.2 Parametric Models 33
- 3.3 Working With Normal and Log-Normal Distributions 46
- 3.4 Fitting Distributions to Data 50
- 3.5 Other Properties of Distributions and Their Evaluation 56
- 3.6 Summary 65
- References 67

4. Regression Modeling and Analysis

- 4.1 Introduction 70
- 4.2 Simple Linear Regression 70
- 4.3 Multiple Regression 78
- 4.4 Nonparametric Transformation and Regression 84
- 4.5 Field Application for Nonparametric Regression: The Salt Creek Data Set 90
- 4.6 Summary 93
- References 96

5. Multivariate Data Analysis

- 5.1 Introduction 97
- 5.2 Principal Component Analysis 98
- 5.3 Cluster Analysis 102
- 5.4 Discriminant Analysis 108
- 5.5 Field Application: The Salt Creek Data Set 110
- 5.6 Summary 116
- References 118
- Further Reading 118

6. Uncertainty Quantification

- 6.1 Introduction 120
- 6.2 Uncertainty Characterization 124
- 6.3 Uncertainty Propagation 132
- 6.4 Uncertainty Importance Assessment 142
- 6.5 Moving Beyond Monte Carlo Simulation 151
- 6.6 Treatment of Model Uncertainty 160
- 6.7 Elements of a Good Uncertainty Analysis Study 163
- 6.8 Summary 164
- References 165

7. Experimental Design and Response Surface Analysis

- 7.1 General Concepts 170
- 7.2 Experimental Design 170
- 7.3 Metamodeling Techniques 178
- 7.4 An Illustration of Experimental Design and Response Surface Modeling 182
- 7.5 Field Application of Experimental Design and Response Surface Modeling 187
- 7.6 Summary 191
- References 192
- Further Reading 193

8. Data-Driven Modeling	9. Concluding Remarks
8.1 Introduction 195	9.1 The Path We Have Taken 225
8.2 Modeling Approaches 197	9.2 Key Takeaways 227
8.3 Computational Considerations 209	9.3 Final Thoughts 231
8.4 Field Example 214	References 231
8.5 Summary 221	
References 223	Index 233

Companion site for this book is available at <https://www.elsevier.com/books-and-journals/book-companion/9780128032794>

Basic Concepts

OUTLINE

1.1 Background and Scope	1	1.3 Random Variables	10
1.1.1 What Is Statistics?	1	1.3.1 Discrete Case	10
1.1.2 What Is Big Data Analytics?	3	1.3.2 Continuous Case	11
1.1.3 Data Analysis Cycle	4	1.3.3 Indicator Transform	12
1.1.4 Some Applications in the Petroleum Geosciences	5	1.4 Summary	12
1.2 Data, Statistics, and Probability	6	Exercises	12
1.2.1 Outcomes and Events	6	References	13
1.2.2 Probability	7		
1.2.3 Conditional Probability and Bayes Rule	8		

1.1 BACKGROUND AND SCOPE

We introduce the reader to some fundamental concepts of classical statistics such as probability and random variables, along with basic concepts from the emerging field of data analytics and big-data technologies. We also list some typical applications of the relevant techniques for data analysis in the petroleum geosciences.

1.1.1 What Is Statistics?

Statistics is the science of acquiring and utilizing data. It provides us with the tools for data collection, summarization, and interpretation—with the goal of identifying the underlying structure, trends, and relationships inherent in the data. This is how we convert data into information.

Fundamental to statistics are the concepts of *population* and *sample*. A population is the universe of all possible outcomes and events, whereas a sample is a finite subset extracted

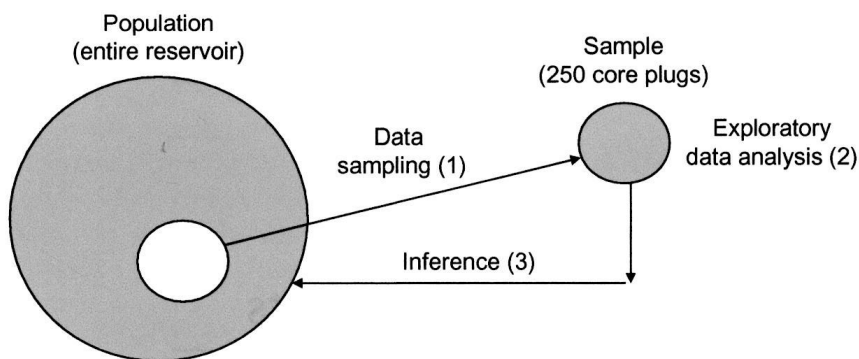


FIG. 1.1 Schematic showing population—sample relationship.

from the population. Statistical analyses are performed on the sampled data to draw *inference* about the characteristics of the population, without having to study the entire population. The population is exhaustive and is characterized by its *parameters*. The sample is limited and is characterized by the *statistic* that is related to the population parameters.

Fig. 1.1 shows a schematic of the relationship between population and sample. Here, the population represents permeability values for an entire oil reservoir at the scale of a small core plug. To learn more about the distribution of permeability values, in step (1), we randomly sample this population using a finite number of core plugs (e.g., 250). In step (2), we analyze these permeability values to determine the proportion of plugs with permeability greater than 10 mD (e.g., 65%). Finally, in step (3), we determine the representativeness of this result for the entire population (e.g., 95% certain that margin of error is $\pm 6\%$).

Application of statistics to any dataset generally begins with exploratory data analysis. Here, the goal is to quantify and visualize the range of values a given variable can take, summary attributes such as averages and spread, and the nature and strength of correlation between two or more variables (Chapter 2). In the next step, the distribution of the variable is examined to understand the relative likelihood of various values within the observed range and the possibility of describing the distribution using a compact mathematical form (Chapter 3). Another common task involves exploring how the relationship between two variables can be described using a linear regression model or variants thereof (Chapter 4). When multiple variables are included in the dataset, it is useful to identify the degree of redundancy among different variables and if the dataset can be partitioned into any statistically homogeneous subpopulations (i.e., clusters). This is the scope of multivariate analysis (Chapter 5).

The broad classes of techniques described above fall within the realm of classical statistics and have been employed by petroleum engineers and geoscientists for many years (see Stanley, 1973 and references therein). Recent contributions (e.g., Davis, 2002; Jensen et al., 2000) discuss the geoscience-oriented application of these techniques in greater detail, including other topics not covered in this book such as geostatistics and time series analysis.

Statistical methods are also relevant in the context of uncertainty analysis, where the goal is to translate the uncertainty in the inputs of a model into uncertainty in corresponding model predictions (Chapter 6). Here, the concepts mentioned in the previous paragraph are fundamental to characterizing the uncertainty both in the model inputs and the model results and