CHAPTER 1

Introduction

Geostatistics has become increasingly popular for numerical modeling and uncertainty assessment in the earth sciences. The beginnings of geostatistics can be traced to innovation in the mining industry in the late 1950s and early 1960s. Application of geostatistics has expanded into other industries including reservoir characterization, meteorology, environmental, hydrogeology, and agriculture. The common features of problems addressed by geostatistics include variables regionalized in space, sparse data, and nonrandom patterns of heterogeneity.

The number of geostatistical theoreticians and practitioners is growing. A reasonable number of courses, books, and papers are available that document the techniques and principles of geostatistics. These include Armstrong (1998), Chilès and Delfiner (1999), Christakos (1992), Clark and Harper (2000), Cressie (1991), David (1977), Davis (1986), Deutsch (2002), Goovaerts (1997), Hohn (1988), Isaaks and Srivastava (1989), Journel (1989), Journel and Huijbregts (1978), Mallet (2002), Matern (1960), Matheron (1969, 1971), Yarus (1995), and Wackernagel (2003). Despite the resources available to the student of geostatistics, the discipline of geostatistics is relatively small and fast growing. There is no comprehensive textbook that explains theory, practice, and provides solved problems to guide the student through the learning process. This book provides a tour through important problems in geostatistics.

This book is aimed at a *student* of geostatistics. We are all students; some are just more formally recognized as such: senior undergraduate students and graduate students. Senior practicing geostatisticians interested in understanding the theoretical principles would be less formally recognized students. Young professionals who recently graduated in the geosciences or earth science related engineering are a special target of this book. Unfortunately, many universities have no requirement for classes in numerical geological modeling with

Solved Problems in Geostatistics. By O. Leuangthong, K.D. Khan, and C.V. Deutsch

geostatistics. The young professional faced with the task of geostatistical modeling without a formal class in the subject, needs a few key references. This book is one of those references.

Most people benefit from seeing how problems are solved. Looking through the problems provides an overview of the discipline. Reviewing the solution methodology provides an overview of the mathematical and numerical tools used in the discipline. Our goal is to present a collection of problems, some theoretical, some practical, that define the key subject areas of geostatistics. These problems point to specific textbook references and specialized papers that, collectively, more completely define the discipline.

Most of these problems have been used in some variant of undergraduate courses, graduate courses, or industry short courses. The authors attended and/or teach at the University of Alberta and Stanford University. The problems lean toward assignments encountered at these two schools and are considered important by the authors. The problems have been modified to focus on key principles and to target a technologically evolving audience. There are many more problems in geostatistics than we could possibly cover in this book.

There are three different types of problems: (1) analytical problems that can be solved by hand, (2) numerical problems that can be solved using any generic spreadsheet type of software, and (3) practical problems that often require some specialized geostatistical software. In the latter case, the intent is not to focus on any particular public domain or commercial geostatistical package; the reader should be able to use any geostatistical software.

For each problem, the objective is given at the beginning to highlight the key learning that is intended. This is followed by the background and assumptions required to proceed with the exercise. The problem is then described with any required figures and/or graphs. A solution plan is then presented that should walk the reader through to the correct solution. This plan is only provided to facilitate the learning process; readers should attempt to solve the problem without this solution plan if possible. In most cases, a partial solution is then presented to permit the reader to check his/her individual solution. Finally, some closing remarks are given with respect to some key ideas that are related, but may not have been required to perform the exercise.

1.1 PLAN OF THIS BOOK

There are nine core chapters. Each chapter is aimed at a particular topic and consists of three problems. We expect students to take from 1 to 20 hours to work through their own solution to each problem.

Chapter 2 addresses some basic probability concepts. Although the use of parametric distributions is diminishing in modern practice, the ability to calculate moments and quantiles of analytical distributions provides valuable insight into

the probabilistic paradigm underlying geostatistics. The first problem calls on basic probability theory. Weighted combinations of data are used commonly in geostatistics. The second one relates to calculating the variance of a linear combination. Finally, the transfer of data and statistics between standardized and nonstandardized units is considered in the last problem.

Chapter 3 focuses on the need for representative statistics. Sites are not sampled equally and considering data as equally representative independent samples from an underlying population is unrealistic. The first problem is a basic one that shows how unequal weights are assigned to data. The second one is a theoretical problem related to the practical notion that there are parts of the distribution that may not have been sampled. In this case, secondary data can be used to establish a representative distribution. The final problem relates to a comparison of practical tools for determining representative statistics.

Chapter 4 reveals the fundamental tool of Monte Carlo Simulation (MCS). The first exercise uses MCS to demonstrate the Central Limit Theorem, that is, the sum of identical equally distributed random variables tends to a Gaussian distribution regardless of the starting distribution. The bootstrap and spatial bootstrap are introduced in the second problem as methods for assessing parameter uncertainty. The third problem is a practical one related to the transfer of uncertainty from input random variables to some nonlinear response variable of interest.

Chapter 5 presents the venerable variogram. The notion of geometric anisotropy is remarkably simple until faced with a practically difficult three-dimensional problem, and then the intricacies of dealing with anisotropy become far more challenging. The first problem explores the details of geometric anisotropy used throughout geostatistics. The second one requires the student to calculate a variogram by hand from a small dataset. Finally, the third problem calls for variogram modeling and the use of the variogram in understanding how variance decreases as scale increases.

Chapter 6 exposes kriging as an optimal estimator. Kriging was independently developed by many experts in many different areas of application. The first problem asks the student to derive the essential equations underlying the kriging estimator. Many variants of kriging are based on constraints to achieve unbiasedness under different models for the mean. The second one asks the student to derive the basis of ordinary, universal and external drift kriging. Kriging is a unique mathematical technique that accounts for proximity of the data to the unsampled value and redundancy between the data. The final problem explores some of the properties of the kriging estimator.

Chapter 7 focuses on Gaussian simulation and its unquestioned importance in modeling uncertainty in continuous variables. The first problem explores the bivariate Gaussian distribution. This setting is good for learning about the properties of the multivariate Gaussian distribution since it is difficult to visualize higher order distributions. The second problem brings out the importance of conditioning in geostatistics; that is, enforcing the reproduction of

local data. The use of kriging to condition unconditional realizations is revealed. Finally, a more complete problem set is presented for simulation of practical-sized realizations.

Chapter 8 is devoted to problems of indicator geostatistics. The rich depths of categorical variable modeling are touched by indicators. The first problem relates to the theoretical link between indicator variograms and object sizes and shapes. The second one demonstrates the use of indicator variograms in the context of checking a multiGaussian assumption. The third problem asks for a hand calculation to perform indicator kriging for the construction of a conditional probability distribution of a categorical variable.

Chapter 9 explores the details of modeling more than one variable simultaneously. The first problem demonstrates the simultaneous fitting of direct and cross variograms with the only practical "full" model of coregionalization: the linear model of coregionalization. The second problem applies the well established paradigm of cosimulation in a multivariate Gaussian framework. In the last problem of this chapter the challenging problem of modeling multiple variables at different scales is introduced.

Chapter 10 touches on a few newer topics. The use of utility theory and loss functions combined with geostatistically-derived distributions of uncertainty is applied in the first problem. The second one is related to the use of probability combination schemes, such as permanence of ratios. These have gained in popularity and are applicable to modeling trends of categorical variables. Finally, the last exercise relates to multiple point geostatistics, which has become increasingly popular in recent years.

These 27 problems highlight geostatistical methods that are either well known or form the basis for techniques that have emerged as particularly promising for future research and application. Some of the problems require data or specific programs. These files and other supplementary material are available from the website **www.solvedproblems.com**. Corrections to the problems and solutions together with additional problems are also available from this website.

1.2 THE PREMISE OF GEOSTATISTICS

At the time of writing this book, the philosophical framework and toolset provided by geostatistics provides the best approach to predict spatially distributed variables. The challenge of spatial inference is overwhelming. Less than one-trillionth of most geological sites are sampled before we are asked to provide best estimates, quantify uncertainty, and assess the impact of geological variability on engineering design.

Matheron formalized the theory of geostatistics in the early 1960's (Matheron, 1971). Geostatistics was not developed as a theory in search of practical problems. On the contrary, development was driven by engineers and

geologists faced with real problems. They were searching for a consistent set of numerical tools that would help them address real problems, such as ore reserve estimation, reservoir performance forecasting, and environmental site characterization. Reasons for seeking such comprehensive technology included (1) an increasing number of data to deal with, (2) a greater diversity of available data at different scales and levels of precision, (3) a need to address problems with consistent and reproducible methods, (4) a belief that improved numerical models should be possible by exploiting computational and mathematical developments in related scientific disciplines, and (5) a belief that more responsible decisions would be made with improved numerical models. These reasons explain the continued expansion of the theory and practice of geostatistics. Problems in mining, such as unbiased estimation of recoverable reserves, initially drove the development of geostatistics (Sichel, 1952; Krige, 1951). Problems in petroleum, such as realistic heterogeneity models for unbiased flow predictions, were dominant from the mid-1980s through the late-1990s. Geostatistics is extensively applied in these two areas and is increasingly applied to problems of spatial modeling and uncertainty in environmental studies, hydrogeology, and agriculture.

The uncertainty about an unsampled value is modeled through a probability distribution. These probability distributions are location dependent. The set of probability distributions over the domain of interest defines a random function (RF), which is the central aim of geostatistics. Inference of statistics related to a RF requires a choice of how to pool data together for common analysis. Furthermore, we may have to model large-scale trends because most regionalized variables exhibit large-scale variations. The decision of *stationarity* is this combination of (1) a choice of how data are pooled together, and (2) the location dependence of spatial statistics. The suitability of geostatistical modeling for its intended purpose requires a reasonable decision of stationarity. It is difficult to conceive of a problem set that would reveal just how important stationarity is to geostatistics.

Geostatistics is concerned with constructing high-resolution models of categorical variables, such as rock type or facies, and continuous variables, such as mineral grade, porosity, or contaminant concentration. It is necessary to have *hard* truth measurements at some volumetric scale. All other data types including remotely sensed data are called *soft* data and must be calibrated to the hard data. It is neither possible nor optimal to construct models at the resolution of the hard data. Models are generated at some intermediate geological scale and then upscaled for computationally intensive process performance. A common goal of geostatistics is to construct detailed numerical models of geological heterogeneity that simultaneously account for a wide range of relevant data of varying resolution, quality, and certainty, so much of geostatistics relates to data calibration and reconciling data types at different scales.

1.3 NOMENCLATURE

The following list consists of fairly standard nomenclature used in a wide variety of geostatistical literature.

$C(\mathbf{h})$: Covariance between two random variables separated by vector $\mathbf{h}$
$Circ_a(\mathbf{h})$: Circular variogram model of parameter a for points separated by $\mathbf{h}$
$\mathrm{Cov}\{X,Y\}$: Covariance between X and Y
$D^2(v,V)$: Dispersion variance of samples of scale v within volumes V
$E\{\bullet\}$: Expected value of $\bullet$
f(z): Probability density function of a random variable Z
F(z): Cumulative distribution function of a random variable Z
$\gamma(\mathbf{h})$: Semivariogram between two random variables separated by vector $\mathbf{h}$
$\gamma(v,V)$: Average semivariogram between volumes v and V
$\Gamma(\mathbf{h})$: Standardized semivariogram (sill of 1.0)
G(y): Standard normal or Gaussian cumulative distribution function
$\mathbf{h}$: Vector separation distance between two points
h: Scalar distance between two points
$i(\mathbf{u};k)$: Indicator value for category k at location $\mathbf{u}$
k: Index of a particular rock type or category
$L(z^*\text{-}z)$: Loss function for error $z^*\text{-}z$ (estimate minus truth)
λ_α: Kriging weight applied to datum α
μ: Mean of a statistical population
p_k: Proportion of rock type or category k
$\mathrm{Prob}\{\bullet\}$: Probability (proportion over similar circumstances) of $\bullet$
ρ: Correlation coefficient (standardized covariance) between two variables
σ^2: Variance
σ_K^2: Kriging variance associated with estimating a particular locaiton
$\mathrm{Sph}(\mathbf{h})$: Spherical semivariogram function for vector $\mathbf{h}$
$\mathbf{u}$: A location in space (1, 2, or 3D)
$\mathrm{Var}\{\bullet\}$: Variance of $\bullet$, that is, the expected squared difference from the mean
w: Weight assigned to a data (often from declustering)
Z: Generic random variable
z: A particular outcome of the random variable Z
z_k: A particular threshold (the k^{th} one) of the random variable Z
$Z(\mathbf{u})$: Generic random variable at location $\mathbf{u}$

Additional notation will be defined in the text as needed. Readers may refer to Olea (1991) for a more extensive and general glossary of geostatistical terminology.