1
Introduction

Figure 1.1 data crunching times
numbers pulsing 'round the globe:
how much is enough?

Carl Sagan famously wrote about the "pale blue dot" that all of us share: "That's home. That's us … there is nowhere else, at least in the near future, to which our species could migrate. Like it or not, for the moment the Earth is where we make our stand" (Sagan 1994, p. 8). More than 20 years after Sagan's famous book, there are plans for human settlement on Mars. But Earth is still home, and on this rock revolving around a golden sun, people everywhere constantly use information to make decisions about utilizing, managing, and sustaining our valuable natural resources. How can we quantitatively analyze and evaluate different information sources for supporting decisions in the Earth sciences?

1.1 What is the value of information?

Making decisions in the Earth sciences can be challenging. There is often significant uncertainty pertinent to the decision – for instance, the availability and spatial distribution of the resources under consideration. Moreover, there could be a lot at stake, as investments may be considerable and there may be huge financial losses or adverse environmental consequences. Petroleum exploration and production, mining, agriculture, and forestry are examples of domains where large-scale efforts are typical. Similarly, endeavors in domains such as conservation biology, ecology, groundwater management, and climate sciences strive to sustain and better manage natural resources and affect several stakeholders.

A unique aspect of the complexity of decisions in the Earth sciences is the inherent spatial variability. The subsurface formations have been forged through millions of years of coupled geological, physical, chemical, and biological processes, resulting in spatial trends and dependence between rock properties at different spatial locations. The subsurface properties have spatial dependence, but at the same time they are very heterogeneous, variable, and uncertain. The decision maker's characterization of uncertainties is then best represented by spatial statistics, because properties at a particular geographic location cannot be treated independently of those at other locations. Reliable information that resolves some of the uncertain properties at one location could therefore go a long way toward improving the overall quality of decisions.

The past few years have seen a tremendous surge of interest in "big data." This has largely been driven by the development in electronics, telecommunications, computer science, online commerce, social media, and our ability to automatically acquire and store data. While this may be a subject of great current interest in popular culture, the Earth sciences have arguably been dealing with big data for a while now. The aspect of volume – the first of the five "Vs" of big data – has certainly been well represented; examples include large geophysical surveys, especially in exploration reflection seismology, and weather and atmospheric data from remote sensing satellites. The utilization and sustenance of the Earth's resources involve multidisciplinary work that can entail acquiring, processing, modeling, and interpreting copious amounts of a variety of data types – the second of the five "Vs." As an example, typical analysis of basin and petroleum systems includes geophysical seismic data, well logs, geochemical analysis, information from biostratigraphy and paleoclimate studies, structural geology, the study of depositional environments, and core analysis. The other "Vs" include velocity – data acquired at a rapid rate (e.g., continuous streams of data from remote sensing sensors) – and veracity – whether the data are accurate and trustworthy. Eventually, the goal is to make better decisions. This is where the last of the five "Vs" of big data becomes important: **value**.

What is the value of the data and how much data are enough? Information almost always comes at a price, so when is the information worth its price? At the very core of this book lies the decision theoretic notion of **value of information** (henceforth referred to as VOI), which we use to evaluate and analyze various sources of data. The power of analyzing information sources using VOI is that: (i) it allows the decision maker to perform a reasonable evaluation before the information is purchased and therefore revealed and (ii) if the decision maker can model value using monetary units, then VOI is also in monetary units. These capabilities make VOI an extremely practical tool that addresses real problems in the real world.

Figure 1.2 demonstrates what we refer to as the "pyramid of conditions" that makes information valuable. Although all technical details are postponed until later in the book,

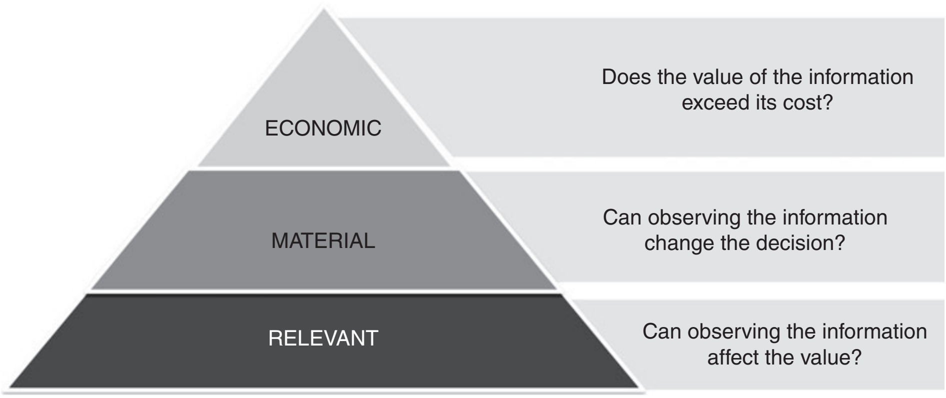

Figure 1.2 The pyramid of conditions specifies that information-gathering schemes should be economic, material, and relevant.

we feel that the figure captures the essence of VOI. The condition at the bottom of the pyramid specifies that information must be **relevant** to the value from the decision, so observing the information can impact the decision maker's beliefs about the key uncertainties. As an (extreme) example, while making a decision about whether you should take your umbrella to work, information about what your friend ate for breakfast hardly seems relevant to whether it will rain! The condition at the middle level of the pyramid specifies that information must be **material**, in the sense that observing it should have the capability to change your decision. If you enjoy both the rain and the sun and would rather not carry an umbrella, then listening to a weather forecast hardly seems worthwhile – you will choose to leave the umbrella at home, regardless of what the forecast has to say. Information must have the potential to allow the decision maker to flexibly adapt and differ from what he or she would have otherwise done. Finally, the condition at the very top of the pyramid is that information must be **economic** – the price of the information must be less than its value. The three requirements are shown as a pyramid because higher conditions cannot be satisfied unless those lower in the pyramid are satisfied. If an information source is not relevant, it cannot be material; if it is not material, it cannot be economic.

The reader may well ask: what is the catch? What do I need to do to harness the wonderful capabilities of VOI? The only catch is that the power of such a practical tool requires some modeling sophistication – it requires understanding and characterizing how the various pieces of the puzzle fit together, and for applications in the Earth sciences, this can often require an interdisciplinary effort. In this book, we recommend a four-stage workflow for using VOI to support information-gathering decisions in the Earth Sciences, as indicated in Figure 1.3:

1. To start with, the decision maker should **frame the underlying decision situation** to understand how the potential information would be used. What are the questions the decision maker is trying to address?

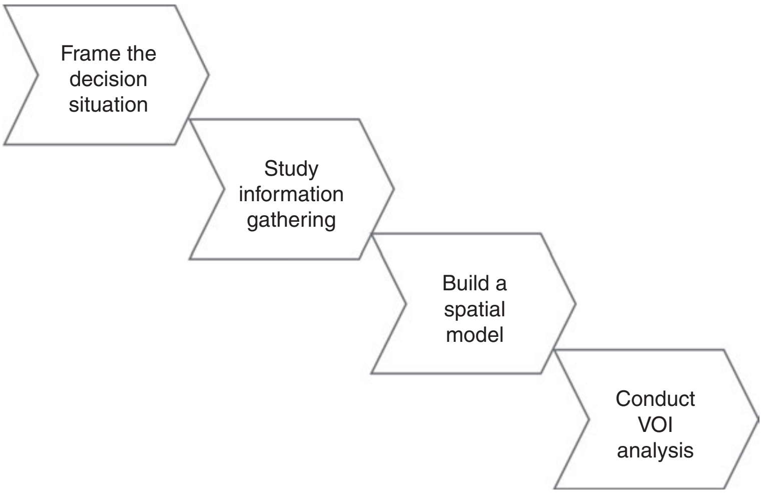

Figure 1.3 A workflow for value of information analysis consists of four steps.

2. Next, the decision maker should identify and **study the potential information-gathering schemes**. How is the information relevant to the underlying decision situation? How might the information affect the decision?
3. In the Earth sciences, spatial dependence is an important aspect of most problems. Therefore, it is often essential to **build a spatial model** that is a reasonable approximation to the real-world domain. This model captures how the various uncertainties, such as geological properties, are connected to each other and to the relevant data.
4. Finally, now that all the pieces are in place, **VOI analysis can be conducted** to address various issues of interest. Information sources of varying reliability and price can be compared, and the decision maker can proactively identify schemes that will increase their value for the decision situation.

As we will later show, the VOI for a data-gathering scheme is computed from comparing the values with and without the data. It is important to note that the underlying decision situation informed by the data plays an explicit part in the calculations. VOI analysis is useful for comparing different schemes. Since the VOI is computed before the data are actually revealed, the value with data must include some kind of averaging over the possible data sets. We highlight how such VOI analysis can be conducted in spatial contexts in the following section.

1.2 Motivating examples from the Earth sciences

Figure 1.4 shows a map view of data from a proposed mining project. The spatial distribution of the oxide grade is highly variable and uncertain. The map displays locations where oxide grade data have been acquired (black). Two types of data have been collected here, at different levels of accuracy. The lower-accuracy data set is obtained with a handheld

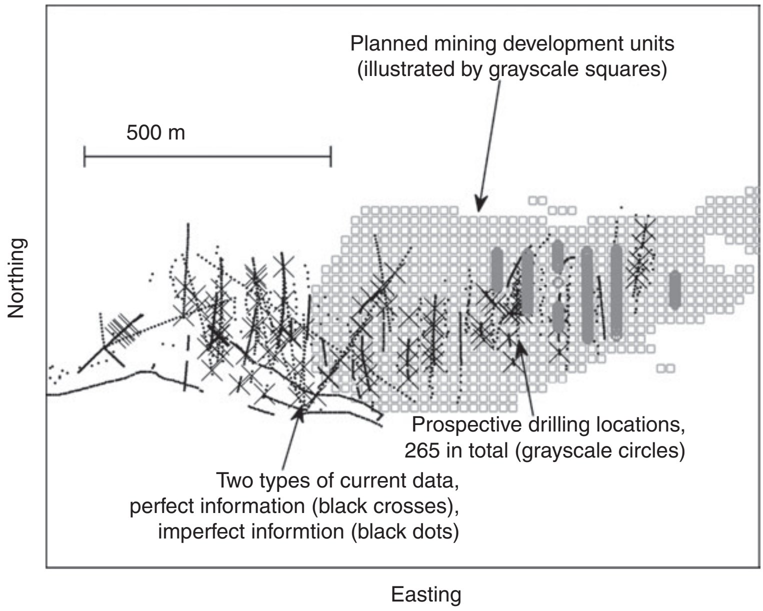

Figure 1.4 Map view of borehole measurement locations and potential mining locations for an oxide resource. The mining company can drill more boreholes before making the mining decision.

meter and requires almost no time for processing (marked with black dots in the display), while the higher-accuracy data set is acquired by taking a core plug from the borehole to the laboratory for careful scanning (marked with black crosses in the display). Based on these observations and expert geological knowledge about the ore, the question is: would mining be profitable? The company has made a careful plan for how the mining would be done (illustrated by gray squared units in the display), but it is difficult to make a decision under uncertainty due to the spatial distribution of oxide. Before making the decision, would mining company can collect more data. Potential locations for additional measurements are defined (marked with gray circles in the display). The data, of course, come at a price for drilling and processing. Are these data worth gathering? And, if yes, which type of data should we acquire – the low-accuracy data or high-accuracy data, or a combination of both? VOI analysis can be used to answer these questions by embedding the decision situation in a spatial modeling framework.

Figure 1.5 shows a network consisting of 38 nodes. The 13 nodes numbered with prefix “P” represent petroleum prospects, and the 25 bottom nodes of the network (illustrated by grayscale circles in the display) are segments of the prospects. This graph, a Bayesian network, is constructed to capture the relationships arising from the geological mechanisms within basin and petroleum systems, and the edges indicate physical connections between geological attributes at the prospects and segments. There is uncertainty about the presence of oil and gas, and the network has an associated probabilistic model that describes these

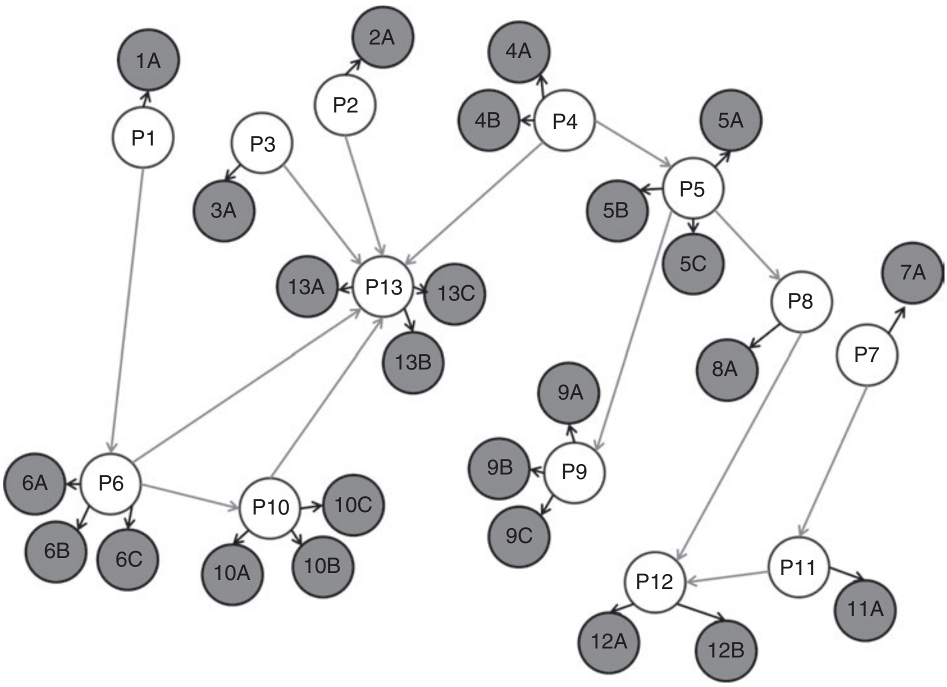

Figure 1.5 Network illustrating 13 oil field prospects and 25 segments where a petroleum company can drill exploration wells before making decisions about prospect development.

uncertainties. A petroleum company can use this model to evaluate their decisions about petroleum field development. Before pursuing expensive development decisions, would it be worthwhile to invest in some exploration wells? An exploration well landed at one segment would reveal the presence or absence of oil at that segment, resolving the uncertainty at that node, and because of the common geological mechanisms indicated by the network edges, the information at the exploration segment would also provide some information about the other segments. If exploration wells will be drilled, where should the company place them? VOI analysis is useful for evaluating such data-gathering schemes.

For the petroleum application, let us zoom in to a finer granularity. A common decision situation is whether to drill production wells at specified reservoir units or to avoid drilling. Before going through with the expensive drilling operations, it could be useful to do a careful subsurface characterization using geophysical data. Seismic and electromagnetic data can be useful in reservoir characterization, improving the prediction of reservoir variables such as lithology, porosity, and saturation. The seismic data undergo processing, which typically provides seismic amplitude information merged in a stack over all angles of the data into one entity (post-stack amplitudes). Would it be worthwhile to invert and interpret the pre-stack data to additionally obtain seismic amplitudes as a

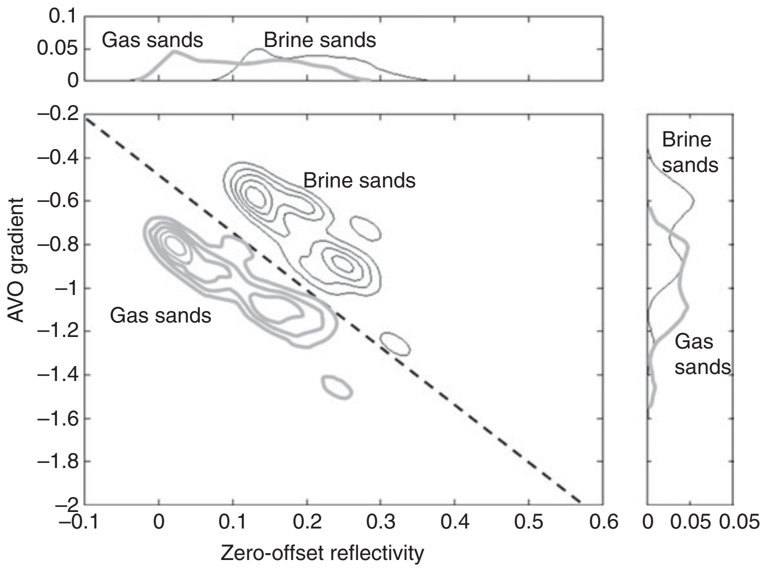

Figure 1.6 Seismic amplitude data are informative of the reservoir rock type and saturation. The contours indicate non-parametric probability density functions (pdfs) of the likely response of the seismic amplitude data for brine sands (black) and gas sands (thick, light gray curves). The marginal pdfs are shown on the right and at the top. The dashed black lines indicate classification boundaries for the two classes.

function of the incidence angle? And what processing accuracy is required for significantly improving the prediction of reservoir rocks and saturations?

Figure 1.6 shows contours of the modeled seismic amplitude responses for two reservoir facies classes (gas sands and brine sands). The contours are representative of bivariate distribution of the expected response for the seismic zero-offset attribute (first axis) and the amplitude-versus-angle attribute (second axis), given any of the two facies classes. These distributions can be assessed from well logs and rock physics and seismic models, but the uncertainty (or the spread and overlap of the distributions) depends on the underlying rock and fluid properties, their natural variability, and the accuracy of the seismic processing scheme. As shown by the marginal distribution on the top panel, the gas sands in this particular reservoir have generally lower zero-offset amplitudes, while the brine sands have somewhat higher amplitudes. However, there is a lot of overlap, and the classification based only on zero-offset amplitudes has a high misclassification error (~25%). If, on the other hand, the classification is based on both the zero-offset amplitude as well as the amplitude-versus-angle attribute, the bivariate classification of the two classes is much better, with a very small misclassification error (~1%). Is it worth purchasing the pre-stack attribute? How would a better classification impact the decision? Perhaps it may be better to get more accurate post-stack amplitudes and reduce the overlap in the marginal distributions of the post-stack amplitudes – or maybe purchase

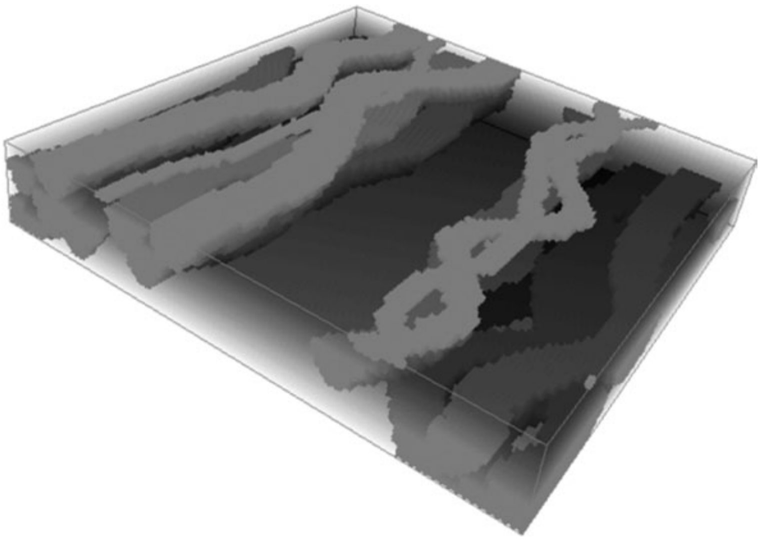

Figure 1.7 Three-dimensional view of one geostatistical realization of uncertain sand channel facies in a groundwater reservoir.

electromagnetic data. Which of the different information-gathering options is most valuable for the decision maker?

VOI analysis can play an important role in better managing groundwater, a valuable natural resource. Will purchasing geophysical electromagnetic measurements help a groundwater manager make better decisions about recharge? As an example, consider the case where groundwater is pumped out from a coastal fluvial aquifer and is used for agriculture. Seawater from the coast starts intruding due to excessive groundwater pumping, leading to increased water salinity and a decrease in usability for crops. Artificial groundwater recharge by pumping freshwater into the subsurface is considered as a way to mitigate salt water intrusion. Where should the recharge site be located? The subsurface distributions of the high-permeability sand channels and low-permeability shales, which impact the groundwater flow and the effectiveness of recharge in the aquifer, are uncertain. Geophysical measurements may be valuable for better characterizing the subsurface and thus help make informed decisions about the selection of possible recharge locations. Assessing the value of the data before actually acquiring them requires modeling the uncertain subsurface channel geometries, accounting for multiple possible scenarios, and conducting flow simulations and geophysical simulations using Monte Carlo computations.

Figure 1.7 shows one realization of subsurface channel geometries generated using multiple-point geostatistical methods. Many such realizations need to be generated to model the subsurface facies uncertainties, and one must simulate their geophysical signatures, as well as the effects of flow and recharge under different recharge alternatives available to the decision maker. The important question is whether the value of the geophysical electromagnetic data is more than the price.

We will return to these examples as well as others in the following chapters. They illustrate current-day applications in the Earth sciences where VOI analysis can be useful.

1.3 Contributions of this book

The main contribution of this book is the blending of decision analysis and spatial statistical concepts to support decisions around information gathering in the Earth sciences. Compared with the traditional use of VOI, we stress the spatial aspects of the statistical model, the alternatives, and the potential information-gathering schemes; many applications of VOI analysis in other disciplines do not need to contend with complexities arising from spatial dependence – a feature that is pervasive in the Earth sciences. Furthermore, we go well beyond the traditional and simplistic high-level models for VOI analysis that have been common for decades and work toward building more realistic models at a finer granularity. We feel that decision makers in the Earth sciences can benefit greatly from our models and methods by deploying them to analyze real-world information schemes. With burgeoning computational capabilities, efficient algorithms, and the ever-increasing availability of data, we sense that more sophisticated models will become increasingly popular and will greatly improve the quality of decisions.

In this book, we refer to decision situations that are typical in the Earth sciences as **spatial decision situations**, building a taxonomy of models based on the decision flexibility – i.e., whether there are a large or small number of alternatives and whether the decision maker's value function is coupled or decoupled, i.e. how complex it is to reasonably compute the decision maker's value from his or her decision situation. We compare various sorts of information-gathering schemes, categorizing them based on whether they provide perfect or imperfect information and whether they are partial or total schemes. We also categorize models based on assumptions around whether decision situations and information-gathering schemes are static or sequential.

We advocate the use of VOI analysis for evaluating data for spatial decision situations. There are several information measures, such as entropy, that are more popular in spatial applications. Unlike VOI, however, these measures only address aspects of the relevancy of the experiment without addressing the monetary gain in value and the ability to make better decisions. VOI is most useful in situations where data are rather expensive. When data come for free, there is not much point in evaluating them – but keep in mind that there is often much effort spent in processing vast amounts of inexpensive data.

Every day, people make important decisions about the development of the Earth's resources. At least as often, decision makers contemplate whether gathering more information will assist them in their difficult decision-making processes. In the future, our demands may change; there may be a drive for recovering unconventional resources in petroleum, we may explore subsea mining, or we may head toward renewable sources such as solar or wind energy. Similarly, environmental challenges will change, and new sustainability questions will arise. What will not change is the advantage of being able to frame decision situations, build useful models for the spatial variables, and evaluate possible information-gathering schemes for improved decision making.

1.4 Organization

This book is organized into seven chapters, including this introduction, as well as an appendix. Chapter 2 introduces basic probability and statistics while providing the fundamental notation for the book. We introduce some specific models, methods, and examples that are used throughout the book. Chapter 3 introduces decision analysis. We focus primarily on the concepts required to understand and appreciate the decision theoretic notion of VOI. Chapter 4 describes spatial statistics models. We motivate spatial modeling through several illustrative examples that are used for VOI analysis in subsequent chapters. These three topics – probability and statistics, decision analysis, and spatial modeling – lay the foundation for the subsequent formulation of methods and tools for VOI in Earth sciences applications. Each of these three topics is, of course, extensive with vast amounts of dedicated literature. The aim is not to cover them comprehensively; rather, the goal in Chapters 2, 3, and 4 is to provide an overview and lay the groundwork for the following chapters.

Chapter 5 integrates concepts from previous chapters. We define spatial decision situations and demonstrate VOI analysis for information-gathering schemes for various categories of spatial decision situations. We describe a taxonomy based on the different combinations of opportunities for spatial decision making and information gathering. Several examples are used to illustrate the concepts.

Chapter 6 provides a number of real-world examples of VOI analysis. We demonstrate applications using data from petroleum, mining, and groundwater applications. For petroleum exploration and development, information-gathering schemes include exploration wells, seismic data, or electromagnetic data. For mine development and safety, various kinds of borehole information could provide valuable information. In the hydrology example, geophysical electromagnetic data could be used to better characterize the subsurface and thus help make better decisions about groundwater recharge.

Chapter 7 contains a number of exercises and hands-on projects. On the book website (srb.stanford.edu/voi), we provide further information about these examples, including various data sets, Netica project examples, and a collection of MATLAB m-files to help readers reproduce many of the results described in the book. The code will also be useful for the hands-on projects.

In every chapter, we end with bibliographic notes, where we present our views on connections with relevant literature as well as some references. Mathematical details about the most important models used in the book are provided in the appendix.

A number of examples recur throughout the book. Table 1.1 summarizes these examples in terms of key assumptions, application domains, and the relevant sections of the book.

1.5 Intended audience and prerequisites

This book is primarily intended for practitioners, professionals, and graduate and advanced undergraduate students in domains associated with the Earth sciences. It may be used as a supplementary text for a class on spatial statistics or decision making in the Earth sciences