SPATIAL STATISTICS AND DIGITAL IMAGE ANALYSIS

Panel on Spatial Statistics and
Image Processing
Board on Mathematical Sciences
Commission on Physical Sciences,
Mathematics, and Applications
National Research Council

National Academy Press
Washington, D.C. 1991

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Frank Press is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Robert M. White is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Samuel O. Thier is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Frank Press and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

The National Research Council established the Board on Mathematical Sciences in 1984. The objectives of the Board are to maintain awareness and active concern for the health of the mathematical sciences and serve as the focal point in the National Research Council for issues connected with the mathematical sciences. In addition, the Board is designed to conduct studies for federal agencies and maintain liaison with the mathematical sciences communities and academia, professional societies, and industry.

Support for this project was provided by the Office of Naval Research under grant number N00014-89-J-1641.

Library of Congress Catalog Card No. 90-63215
International Standard Book Number 0-309-04376-X

Available from
National Academy Press,
2101 Constitution Avenue, NW
Washington, DC 20418
S237

Printed in the United States of America

CONTRIBUTORS

ADRIAN BADDELEY, Centre for Mathematics and Computer Science
JULIAN BESAG*, University of Washington
HERMAN CHERNOFF*, Harvard University
PETER CLIFFORD, Oxford University
NOEL A. CRESSIE, Iowa State University
DONALD J. GEMAN, University of Massachusetts
BASILIS GIDAS*, Brown University
LAWRENCE S. GILLICK*, Dragon Systems, Inc.
NICHOLAS GREEN, Oxford University
PETER M. GUTTORP, University of Washington
TED KADOTA, AT&T Bell Laboratories
ALAN LIPPMAN, University of Washington
JAMES SIMPSON*, Scripps Institution of Oceanography

Staff
SCOTT T. WEIDMAN, Senior Staff Officer

---

*Member of Panel on Spatial Statistics and Image Processing, Julian Besag and James Simpson, co-chairs

BOARD ON MATHEMATICAL SCIENCES

COMMISSION ON PHYSICAL SCIENCES, MATHEMATICS, AND APPLICATIONS*

* The project that is the subject of this report was initiated under the predecessor group of the Commission on Physical Sciences, Mathematics, and Applications, which was the Commission on Physical Sciences, Mathematics, and Resources, whose members are listed on the following page.

COMMISSION ON PHYSICAL SCIENCES, MATHEMATICS, AND RESOURCES

PREFACE

Spatial statistics is one of the most rapidly growing areas of statistics, rife with fascinating research opportunities. Yet, many statisticians are unaware of those opportunities, and most students in the United States are never exposed to any course work in spatial statistics. This report aims at illustrating the wide scope of spatial statistics to provide an introductory snapshot of the field to researchers and graduate students in both statistics and related areas. It is hoped that these readers will go on to explore the many research opportunities in the subject, or bring appropriate problems to the attention of practicing spatial statisticians.

This panel was specifically charged to prepare a cross-disciplinary report on spatial statistics and image analysis that would (1) describe the contributions of the mathematical sciences, (2) summarize the current state of knowledge and open problems, and (3) identify likely future fruitful directions for research.

CONTENTS

Plates for chapters 2, 3, and 10 precede page 71.

1

Introduction

Spatial statistics is concerned with the study of spatially referenced data and associated statistical models and processes. It is therefore relevant to most areas of scientific and technological inquiry. In addition, there are many problems that occur in subjects that are not overtly spatial, for example in speech recognition or in the construction of expert systems, that can be given useful spatial interpretations or can benefit in some other way from research in spatial statistics. Indeed, the abundance of application areas has meant that the task of the panel in preparing this report has been not only stimulating but also difficult, in that a limited number of topics had to be chosen for detailed discussion.

The title of the report clearly implies that the panel places considerable emphasis on the relationship of spatial statistics to digital image analysis. This emphasis reflects the recent surge of interest among mathematicians and statisticians in this exciting area, which is destined to play an increasingly important role, not only in science and technology but in everyday life. For example, sequences of satellite images of regions of the Earth are now collected routinely. Each individual image is concerned with only a small part of the Earth's surface and itself is subdivided into a rectangular array of picture elements or "pixels," typically 1024×1024. For every pixel several or many measurements are taken, each of which corresponds to a reflectance value in a particular range of the visible or near-visible electromagnetic spectrum. The eventual aim might be to convert this vast quantity of two-dimensional, multivariate data into a simple crop inventory that can be used, for instance, to estimate the total potential winter-wheat harvest of a country. Satellite images are also used for other purposes, such as locating and monitoring the condition of rocket silos in foreign territories; here, there are analogies in computer vision, where object recognition is one of many

important tasks. Significantly, conceptually similar problems occur in (nuclear) magnetic resonance imaging (MRI) of the brain and of other human organs, where it is required to produce tissue classifications (e.g., into white matter, gray matter, spinal fluid, and tumor) from multispectral data.

Magnetic resonance imaging, mentioned above, represents just one of several different imaging modalities in nuclear medicine. Other examples include the CAT-scan, in which X-ray images taken from several different positions are combined to reconstruct views of cross-sections of the anatomy of a patient; positron emission tomography (PET) and single photon emission computed tomography (SPECT), which are used to measure perfusion (blood flow) and metabolic activity in specific organs; and ultrasonic imaging, for measuring reflective and refractive gradients, such as organ boundaries, within the body. Here, we briefly describe SPECT, a low-cost technique within the reach of most medical facilities, as opposed to PET, which requires an on-site cyclotron and is available only in roughly 100 hospitals worldwide. Of course, there is a price to pay: SPECT currently produces much cruder images. However, this inadequacy stems in part from poor use of underlying, well-understood physical principles and it is here that mathematical and statistical modeling can play a fundamental role.

In SPECT, a patient is injected with a radiopharmaceutical that has been tagged with a radioactive isotope. The pharmaceutical is chosen for its propensity to concentrate in the organ of interest in a way that is related to the particular phenomenon under study. The aim is to map the concentration of the pharmaceutical throughout the target region, usually on a slice-by-slice basis; time may also be a factor, as when different phases of a heart cycle are being monitored. SPECT relies on the radioactive decay of the isotope, which causes photons to be emitted according to a Poisson process in space and time, with intensity at any particular location being proportional to the concentration of the pharmaceutical there. A bank of gamma cameras, usually in a 64×64 array, counts the photon emissions that reach it and, by repeating the procedure for typically 64 positions around the patient's body, data that correspond to 64 different projections are collected. Mathematical interest centers on how the $64 \times 64 \times 64$ array of counts can be used to reconstruct an accurate estimate of the true intensity map, suitably discretized. Commercially available reconstruction methods are based on "filtered back projection" (FBP), a technique borrowed from transmission (e.g., X-ray) tomography. However, FBP is not appropriate to SPECT, because of the very low signal-to-noise ratio and the importance of non-uniform attenuation and depth-dependent scatter and blur. These

forces combine to produce unsatisfactory reconstructions. At first sight, it might appear sufficient to build a proper physical model, in which the data are independent observations from Poisson random variables with means determined by a particular transform of the true intensity map. Unfortunately, the inverse problem of inferring (a discretized version of) the true intensities is too ill-posed for this to provide a satisfactory solution. An additional regularization assumption must be made, which prevents the local behavior of the reconstruction from becoming too disjoint, yet does not impose undue smoothness on the image. The Bayesian solution to this dilemma is one of the topics tackled in chapter 2, but the basic idea is to specify a stochastic model for the true image that is at once globally flexible, yet locally constrained to produce severe discontinuities only when there is convincing evidence of their existence in the data. Incidentally, a somewhat similar problem occurs in the epidemiology of rare, noncommunicable diseases, such as particular forms of cancer, when incidence rates are observed over a specific period of time in a large number of contiguous administrative regions and the objective is to estimate underlying differences in risk. In each region, the number of cases can be viewed as an observation from a Poisson distribution, with mean proportional both to the population and to the risk there. When the means are small, the observed rates are very noisy and provide a poor measure of risk, so that some form of smoothing is required to produce a more readily interpretable map. Note that this problem is simpler than SPECT in that it involves direct rather than indirect sensing and also the number of observations is much smaller. As a result, it is possible to implement computationally intensive methods of spatial statistical analysis that are not yet feasible for genuine images. Such problems are therefore not only valuable in their own right but provide useful insights for the future.

However, attention should also be paid to the origins of spatial statistics, as well as to its present and future. Perhaps the best known and most accessible among early examples is that of Dr. John Snow, a medical practitioner, who traced the precise source or cholera epidemic in central London in 1864 by plotting the locations of water pumps and of deaths from the disease on the same map (see Tufte, 1983, p. 24). Such simple graphical techniques are still very important, though it should be noted that they are often of little value in modern epidemiology because of variations in background population density. One could cite many other isolated examples, but it is probably fair to say that spatial statistics did not emerge as an identifiable discipline until 1960, with the publication by Bertil Matern of his doctoral

dissertation entitled *Spatial Variation*. Much of the material was well ahead of its time, although, remarkably enough, some of it had been completed as early as 1947. The treatise has recently been republished (Matern, 1986) and still provides much useful guidance, regarding both statistical theory and practice. However, it was not until the 1970s and 1980s that spatial statistics began to receive widespread attention from other mathematicians and statisticians. Nonetheless, it may be said to have "come of age," as it now provides a major focus of contemporary research. Applications are many and varied and generate a steady stream of new problems.

Historically, observational programs that use the analysis methods of spatial statistics (e.g., earth sciences, agriculture, and epidemiology) have been limited by sparse sampling. As recently as 20 years ago, for example, as few as 10 observations of sea surface temperature per day over a 200-km^2 area of the ocean was considered state of the art. From a statistical point of view, the inadequacy of such sparse sampling in a domain of large spatial and temporal variation (such as in the ocean) was clear. Modern data acquisition methods (e.g., satellite observations from space) now have greatly circumvented this sampling limitation. Data rates as high as 10^6–10^7 bits per second are routinely achieved with this new technology. A similar situation exists in nuclear medicine. The overall result of this improvement in data acquisition is the development of data bases that provide a high spatial resolution and a synoptic realization of a given process under study. Such data bases are manageable, however, only because of contemporaneous advances in digital computer processing and mass storage/retrieval devices.

Computer resources are also required to execute the large number of repetitive operations typically required in the application of a technique of spatial statistics or digital image analysis to a field of science. Advances in work station technology, data base management, data compression, and data archiving, coupled with the expansion of computer network topologies, now provide the necessary technical infrastructure for the development of joint university curricula in spatial statistics and digital image analysis and in its cross-disciplinary application of methods to a broad range of scientific, engineering, and medical problems.

The vast amounts of data collected by satellites, radar, and sonar measurements needs to be organized and reduced in complexity. While statistics originally emphasized obtaining maximal information from minimal data, the challenge from these new data sources is to summarize eloquently and to increase understanding of enormous quantities of information. Pictures need to be sharpened, new summary measures need to be developed, and

different forms of storing, organizing, and retrieving information need to be implemented. The interface between statistics and computer science is particularly important here: data structures, such as geographic information systems (GIS), can help in both organization and display of spatially expressed data, and graphical tools that visually link overlaid data components are useful in detecting and exhibiting relationships. There are many open research problems in the area of visual data-analytic techniques. For example, how does one display the uncertainty connected with contour lines on a statistical map, and what is an effective way of displaying more than one spatially expressed variable?

The remainder of this volume consists of 10 scientific chapters. Chapter 2 describes the Bayesian/spatial statistics framework in image analysis and computer vision. Particular attention is paid to image reconstruction. Chapter 3 addresses the application of non-Bayesian digital image analysis methods to oceanography and atmospheric science. Examples of image segmentation (i.e., cloud detection in complex natural scenes), near-surface velocity computation from image sequences, and ice boundary detection in satellite data are given. Chapter 4 applies methods of spatial statistics to a broad range of environmental science issues: spatial variation in solar radiation, environmental impact design, and modeling of precipitation using space-time point processes. Chapter 5 provides a basis for geostatistical analysis of earth science data. The variogram and kriging are then exploited to study the flow of groundwater from a proposed nuclear waste site and the spatial distribution of acid rain over the eastern half of the United States. The uses of spatial statistics to analyze data from agricultural field experiments are explored in chapter 6. The objective of such analyses is to compare the effectiveness of different treatments (e.g., fertilizers) on a particular crop variety or to make comparisons between different varieties of the same crop but with a valid assessment of error. Chapter 7 examines the traditional use of point process methods in ecology and analyzes some of the weaknesses in this applications area. Spatial statistics as a signal processing tool for radar and sonar systems in the ocean is the topic of chapter 8, while chapter 9 uses spatial statistics to examine chemical kinetics of active chemical systems. Chapter 10 provides a statistical basis for the field of stereology and discusses statistical modeling of stereological data. Finally, chapter 11 discusses problems of speech recognition with the ultimate goal of enabling machines to emulate human speech.

Although this report attempts a broad overview of main areas of spatial statistics and digital image analysis, there are many areas that we have not

covered. For example, spatial aspects of epidemiology have been used extensively in attempting to relate disease incidence to potential causes, but this application is not addressed in this report. Likewise, in astronomy, Neyman and Scott (1958) initiated the uses of the clustered point processes (*cf.* Chapter 7) to describe the distribution of galaxies and clusters of galaxies. Recent work on processing images of galaxies (Molina and Ripley, 1989) provides a substantial improvement over traditional (maximum entropy) methods in the area. Methods of digital image analysis have recently found application in population genetics to deal with complex pedigrees (Sheehan, 1989). Sampling techniques for spatial data, emphasizing systematic designs, are reviewed in Ripley (1981, Ch. 3).

At first glance, this diversity in the applications of spatial statistics and digital image analysis may mask some of the underlying concepts common to most of the applications. Historically, spatial statistics and digital image analysis have tried to extract the most information from limited data sets. Modern data acquisition systems (e.g., remote-sensing of the Earth using satellites, nuclear medicine) now provide well-sampled spatial data. Hence, a relatively new role for spatial statistics and digital image analysis is to synthesize and reduce large volumes of data into manageable pieces of information. "Modeling," used in a most generic sense, is perhaps the most fundamental concept unifying the diverse applications base of modern spatial statistics. Models attempt to provide a coherent framework for the interpretation of complex data sets. Statistical models, which generally are noncausal in nature, draw conclusions about data sets without necessarily providing future predictive capability. Physical models, which generally include time dependence, attempt to provide a prognostic capability about a physical process based on available data sets. It is likely that significant advances in science and engineering will be made by judiciously combining these two types of models. It is also important to note that some phenomena (and/or data sets) may not be amenable to modeling. In this case, spatial statistics attempts to develop the best representation of the data set from which the maximum statistically robust information can be extracted. Some of the methods found in these applications are (1) exploitation of local specification models (i.e., Markov random fields), (2) use of covariance estimation, and (3) the revolutionary role of the Gibbs Sampler in Bayesian statistics.

Recent advances in computer science and computer technology have contributed significantly to the efficient and effective utilization of spatial statistics by other fields (e.g., engineering sciences). Present computer architec-

tures, however, are still far from ideal for several other classes of important problems in spatial statistics. For example, although massively parallel computers are applicable to some problems in spatial statistics and digital image analysis, current designs are not particularly appropriate to the important tomographic reconstruction problem.

Modern advances in image display technology, visualization techniques (e.g., dithering), and the theory and implementation of compact, efficient data structures now allow scientists and engineers to store, retrieve, and display efficiently the large amounts of spatial data now being routinely recorded in such diverse fields as medicine, oceanography, and astronomy. Advances in spatial statistics will also be closely linked to these advances in computer science, especially within the subfield of data structures. Continued research in data structures should be directed toward determining the most compact and efficient data structures for the storage and representation of information related to problems in two-dimensional signal analysis and image analysis.

Spatial statistics and digital image analysis will play important future roles in science and industry. For example, nondestructive evaluation (NDE) methods will be used more extensively by industry and government to perform quality control and assurance on a spectrum of applications ranging from manufacturing of circuit boards to metal fatigue tests on airplane fuselages. Often such applications of NDE involve ultrasound detection and tomographic reconstruction, and/or two-dimensional signal processing and methods of digital image analysis. Methods of digital image analysis and image reconstruction also will be used to analyze the large-volume spatial data sets necessary to study a variety of issues (e.g., acid rain, ozone depletion, and global warming) related to quantitatively understanding climate and global change processes on planet Earth. Continued advances in data acquisition and digital image analysis will have a significant impact on such diverse fields a medicine and astronomy.

At present, the United States has limited indigenous expertise in spatial statistics and its relation to modern methods of digital image analysis. Few university programs exist that properly accommodate the inherent cross-disciplinary nature of the field. The panel believes that careful consideration should be given to the development of joint curricula in spatial statistics and digital image analysis, which should accurately reflect their diverse applications in the fields of science, engineering, and medicine.

Bibliography

[1] Matern, B., *Spatial Variation*, Meddelanden fran Statens Skogsforskningsinstitut **49**. Second edition published as Springer Lecture Notes in Statistics, vol. 36, Springer, New York, 1986.

[2] Molina, R., and B. D. Ripley, Using spatial models as priors in astronomical image analysis, *J. Appl. Stat.* **16** (1989), 193–206.

[3] Neyman, J., and E. L. Scott, Statistical approach to problems of cosmology, *J. R. Stat. Soc., B* **20** (1958), 1–43.

[4] Ripley, B. D., *Spatial Statistics*, John Wiley and Sons, New York, 1981.

[5] Sheehan, N. A., *Image Processing Procedures Applied to the Estimation of Genotypes on Pedigrees*, Technical Report No. 176, Department of Statistics, University of Washington, Seattle, 1989.

[6] Tufte, E., *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Conn., 1983.

2

Image Analysis and Computer Vision

Donald Geman, University of Massachusetts
and
Basilis Gidas, Brown University

2.1 Introduction to Image Analysis

The goal of computer vision is to build machines that reconstruct and interpret a three-dimensional environment based on measurements of radiant energy, thereby automating the corresponding processes performed by biological vision systems, and perhaps eventually extending some abilities of these systems. This goal is far from being attained, and indeed most of the fundamental problems remain largely unsolved. The field is generally immature by comparison with standard scientific disciplines, the present methodology being a hybrid of those from artificial intelligence, classical signal processing, pattern theory, and various branches of mathematics, including geometry and statistics. Still, important advances have been made that are beginning to substantially affect such areas as industrial automation, earth science, medical diagnosis, and digital astronomy.

Computer vision tasks are generally divided into "low-level" and "high-level" problems to differentiate between those that (apparently) are largely data-driven ("early vision") and those that (apparently) rely heavily on stored knowledge and symbolic reasoning. More specifically, low-level vision includes such problems as coding and compressing data for storage and transmission; synthesizing natural and man-made patterns; restoring images degraded by blur, noise, digitization, and other sensor effects; reconstructing images from sparse data or indirect measurements (e.g., computed tomography); computing optical flow from motion sequences; and reconstructing

three-dimensional surfaces from shading, motion, or multiple views (e.g., binocular stereo) or multiple wavelengths (e.g., multichannel satellite data). In contrast, high-level vision tends to be driven by more specific goals and generally involves object recognition and scene interpretation. These are perceived as the most difficult of the "natural" processes to replicate; in particular, one of the factors that inhibits the introduction of industrial and exploratory robots is their inability to "see," in particular, to infer enough information about objects to navigate in complex, cluttered environments. Indeed, invariant object recognition is one of the most challenging modern scientific problems whose resolution may require new conceptual principles, computational procedures, and computer architectures.

Biological vision systems, in particular the human eye and brain, analyze scenes in an apparently effortless way. This ability appears miraculous (especially to workers in computer vision) and is attributed to several factors, not the least of which is the large proportion of the human brain devoted to vision. Still, we actually know very little about the principles of biological vision despite the information gathered by physiological, psychophysical, and neurophysiological studies. We do know that our visual system is able to integrate cues from many sources (e.g., binocular stereo, motion, and color) and that we exploit a priori expectations, specific scene knowledge, and contextual clues to reduce ambiguities and "correctly" perceive the physical world. It also appears that low-level analysis of retinal information and high-level cognition are done interactively, sometimes referred to as the integration of "bottom-up" and "top-down" processing. Finally, there is little doubt that our recognition and interpretation system is largely scale invariant and at least partially rotation invariant.

Computer vision systems are usually quite inferior to biological ones. This may be due in part to the lack of raw processing power or suitably parallel computation, but also, and perhaps more important, to the inability of synthetic systems to integrate sources of information and place appropriate global constraints. At the moment, automated visual systems rarely make "interpretation-guided" or "knowledge-driven" decisions, due probably to a lack of sufficiently invariant representations and to feedback mechanisms between these representations and the raw data. It appears inefficient, if not fruitless, to attempt to represent a given object in all its possible presentations. Instead, object representations and recognition algorithms should possess certain invariances with respect to scale, location, and rotation.

Despite these shortcomings, important successes have been achieved in actual applications. In some areas, the sheer amount of data leaves no choice

except automated analysis. For instance, a single Landsat Multi-spectral Scanner image may contain 30 Mbytes of data. In fact, some of the earliest and most publicized successes of computer vision occurred during the 1960s and 1970s when images received from orbiting satellites and space probes were substantially improved with linear signal processing techniques such as the Wiener filter. More recently, significant advances have been made in the classification of satellite data for weather and crop yield prediction, geologic mapping, and pollution assessment, to name but three other areas of remote-sensing. Another domain in which automated systems are achieving success is the enhancement and interpretation of medical images obtained by computed tomography, nuclear magnetic resonance, and ultrasonics. Other applications include those to optical astronomy, electron microscopy, silicon wafer inspection, optical character recognition, robot navigation, and robot manipulation of machine parts and toxic material.

By and large, the algorithms used in machine vision systems are specifically dedicated to single applications and tend to be ad hoc and unstable. From a practical viewpoint, problems arise when algorithms are so critically "tuned" to the particulars of the environment that small perturbations in the output of the sensors, or ordinary day-to-day variations, will significantly reduce their level of performance. Obviously, there is a need for algorithms that are more robust and that are based on solid theoretical foundations. But these are ambitious goals; the problems are very difficult, ranging, in mathematical terms, from ill-conditioned (unstable) to ill-posed (underdetermined), the latter due to the loss of information in passing from the continuous physical world to sampled and quantized two-dimensional arrays. In order to reduce the ambiguity, it is necessary to reduce the set of plausible interpretations by incorporating a priori knowledge and integrating multiple cues. One then seeks a mathematical representation for structure and regularity.

Standard regularization theory, as applied, for example, to inverse problems in particle scattering, is deterministic and lacks flexibility. One major current trend is toward "stochastic regularization." This is not surprising in view of the fact that many natural regularities are in fact *nondeterministic*: they describe correlations and likelihoods. Spatial statistics, in general, and lattice-based random field models, in particular, provide a promising framework for capturing these regularities and quantifying generic and a priori knowledge. Such models, properly conceived, impose severe but natural restrictions on the set of plausible interpretations. Thus spatial processes and Bayesian inference have provided a coherent theoretical basis for cer-

tain inverse problems in low-level vision. Moreover, this framework supports robust and feasible computational methods (e.g., Monte Carlo algorithms), measures of optimality and performance, and well-designed principles of inference. This methodology is described in more detail in §2.4. Section 2.2 contains a brief review of digital images, and §2.3 describes four specific image analysis tasks.

2.2 Digital Images

The data available to an automated vision system are one or more images acquired by one or more sensors. The most familiar sensors are optical systems, with ordinary cameras and lenses, that respond to visible light. Other sensors respond to electromagnetic radiation corresponding to other parts of the spectrum (e.g., infrared, X-rays, and microwaves), or to other forms of energy such as ionized high-energy particles (protons, electrons, alpha particles), ultrasound, and pressure (tactile sensors). Many applications employ multiple sensors; for example, navigation robots may be equipped with video camera(s), range, and tactile sensors, and the Landsat multispectral scanners collect data in bands of both visible light and infrared radiation. *Sensor fusion* is a current trend in many technologies and inferential procedures.

Regardless of the form of energy acquired and the specific processes of detecting, recording, and digitizing, the output of all sensors has a common structure: it is a finite collection of measurements, $y = \{y_t : t \in T\}$, indexed by a finite set T. With few exceptions (e.g., photon emission tomography), y is a two-dimensional array, i.e., T is a grid of points (pixels) in the two-dimensional image plane. Each y_t is integer-valued or, as in the case of multispectra satellite and color images, a vector with integer-valued components. Except in photon counting devices, the values of y may be regarded as "quantizations" of a continuum signal $g = \{g_t : t \in T\}$, which, in turn, is a discrete approximation or sampled version of a function $g(u), u \in \mathbf{R}^2$, defined on the two-dimensional image plane (or some domain $S_0 \subset \mathbf{R}^2$). In addition to the errors introduced by *digitization* (= sampling + quantization), g involves various types of *degradation* (e.g., blur and noise; see below). In the absence of these degradations and other artifacts, i.e., in an "ideal" system, we would observe an ideal energy pattern $f(u)$, $u \in \mathbf{R}^2$.

The data y or $\{g_t : t \in T\}$ may be thought of as a representation of the physical scene being imaged. The task of computer vision is to estimate or infer properties ("attributes") of the scene on the basis of the data and a

priori knowledge or expections about the physical world. Attributes of interest may be the true pattern $f(u)$ itself (as in *image restoration*); geometric features (e.g., orientation, depth, curvature) of objects in the scene; or semantical labels for scene entities, such as in object recognition or remote sensing, in which regions are classified as "forest," "cropland," "water," and so on. The relation of specific attributes (e.g., shape) to the true pattern $f(u)$ is problem-specific, and some concrete cases will be treated in §§2.3, 2.4.

More specifically, the true pattern $f(u)$ corresponds to the distribution of energy flux (*radiance*) emitted by objects, either because they are "illuminated" by an energy source, or because they are a primary source of energy themselves; it is often referred to as "scene intensity" or "brightness." The measured values g correspond to the energy flux (or *irradiance*) intercepted, detected, and recorded by the sensor, and are usually referred to as "image intensities" or again simply as "brightness." Between emission and detection, various types of distortion and artifacts occur. These are usually lumped into three major categories (before digitization): *blur*, which may be introduced by scattering within the medium (e.g., atmosphere, human body), de-focused camera lenses, or motion of the medium, objects, or cameras; *noise*, introduced by the random nature of photon propagation ("quantum noise") or by the sensing and recording devices (e.g., film grain noise or current fluctuations in electronic scanners); *nonlinear transformations* of the signal (referred to as "*radiometric distortion*") introduced again by the sensing and recording devices.

These degradations amount to a transformation from $f(u)$ to $g(u)$. The most general transformation encountered is

$$g(u) = \psi\{\phi[K(f)(u)]\,,\, \vec{\eta}(u)\}\,. \tag{2.1}$$

Here, K accounts for blur and $K(f)(u) = \int K(u, v, f(v))dv$. In the linear case, $K(u, v, z) = K(u, v)z$ and $K(u, v)$ is called the *point spread function* (PSF); the function ϕ accounts for radiometric distortions, $\vec{\eta}$ is a collection of noise processes, and ψ defines the noise mechanism (e.g., additive, multiplicative). These parameters have been studied extensively for many sensors (e.g., Vidicon and CCD cameras) and media (e.g., the atmosphere and human tissues). We refer to [9] and references cited there for more details.

2.3 Some Problems in Image Analysis

In this section we describe four specific problems, which are representative of those in low-level computer vision: (1) image restoration; (2) boundary detection; (3) tomographic reconstruction; (4) three-dimensional shape reconstruction. These problems demonstrate the mathematical difficulties encountered in converting information which is implicit in the recorded digital image to explicit properties and descriptions of the physical world. By and large, these problems are naturally nonparametric, and, as inverse problems, range from ill-conditioned to ill-posed. Consequently, as discussed in §2.1, one seeks *a priori* assumptions, or *a priori* information about the physical world (i.e., separate from the data and imaging equations), to constrain the set of possible, or at least plausible, solutions. An approach to some of these problems based on stochastic regularization is presented in §2.4. Other approaches are briefly indicated in §§2.3 and 2.4, and a few references are cited for more details. However, these problems have been studied extensively in the computer vision and engineering literature, and we refer the reader to [9] (and other surveys) for more complete references to related approaches based on stochastic regularization and Bayesian inference. Finally, the field of mathematical morphology [26], which is not considered here, offers a quite different methodology for some of these problems.

2.3.1 Image Restoration

The classical image restoration problem for intensity data is that of recovering a true, two-dimensional distribution of radiant energy, $f(u)$, $u \in \mathbf{R}^2$, from the actual recorded image values. In a "continuous-continuous" setup, the problem is posed by equation (2.1). However, since the number of recorded values is finite (in fact space-discretization is inherent in many sensors), the continuous-discrete formulation is more realistic. Ignoring, for the moment, quantization of the measurement values, we can assume the actual recorded data constitute a two-dimensional array $g = \{g_i : i \in S\}$ of positive real numbers on a regular grid $S = \{i = (i_1, i_2) : 1 \leq i_1, i_2 \leq N\}$, in fact an $N \times N$ integer lattice. Then (2.1) is replaced by

$$g_i = \psi\{\phi[(Kf)_i], \vec{\eta}_i\}, \quad i \in S. \tag{2.2}$$

Assuming the degradation mechanism to be known (or previously estimated), the problem of recovering $f(u)$ from $\{g_i\}$ is nonparametric and obviously ill-posed in general. For computational purposes, the domain of $f(u)$ is

discretized into a larger integer lattice S', concentric to S, of dimension $N' \times N'$, $N' \geq N$, and $f(u)$ is replaced by $f = \{f_i : i \in S'\}$. Then K becomes a discrete representation of the point spread function, and for linear space-invariant systems we have

$$(Kf)_i = \sum_{j \in S'} K(i-j) f_j \,, \quad i \in S \,.$$

Assuming K has bounded support (a reasonable assumption in most cases), then S' should be taken large enough so that the summation over $j \in S'$ includes all terms for which $K(i-j) > 0$. This is not exactly a convolution; one can imagine f convolved with K on the infinite lattice, but only observed on S, and reconstructed on S'.

The problem of estimating $\{f_i\}$ from $\{g_i\}$ is still ill-posed. To see this, consider the simple case of a linear model: $g = Kf + \eta$, with only blur and a single noise process. By relabeling the sites of S and S', we can regard g, η, and f as vectors of dimension N^2, N^2, and N'^2, and K as an $N^2 \times N'^2$ matrix. For $\eta = 0$ and $N < N'$, the problem is underdetermined and K^{-1} is not well-defined. But even if $N = N'$ and K were invertible (e.g., toroidal convolution), the matrix is usually nearly singular, so the existence of measurement and quantization errors then renders the problem unstable in the sense that the propagation of errors from the data to the solution is not controlled. Put differently, given g and K, two images with blurred values very close to g can be very different.

Consequently, one seeks additional information to constrain or "regularize" the problem. Traditional approaches (see section 4.2 of [9] and the references there) may be divided into *linear* methods, such as the traditional Wiener filter and other constrained least-squares methods, and *nonlinear* methods, such as the maximum entropy technique. Linear methods are typically ill-conditioned and the reconstructions are often compromised by large oscillatory errors. Maximum entropy has been extensively examined and is widely popular in certain areas, such as digital astronomy [16, 8, 19, 24]. The regularization underlying the standard constrained least-squares filter is equivalent to a Gaussian (improper) prior distribution, whereas that of maximum entropy is equivalent to a particular random field model whose variables interact only through their sum (= total energy). None of these methods addresses the crucial issue of discontinuities, which often convey much of the information in the image, and which are difficult to recover with standard methods. A general framework for nonlinear estimation based on spatial stochastic processes is outlined in §2.4.

2.3.2 Boundary Detection

The boundary detection problem is that of locating (discrete) contours in a digital image that correspond to sudden changes of physical properties of the underlying three-dimensional scene such as surface shape, depth (occluding boundaries), surface composition (texture boundaries), and surface material. A common complication is that sharp physical boundaries may appear in an image as slow intensity transitions or may not appear at all due to noise and other degradation effects. In addition, extraneous "edges" may appear from artifacts of the imaging system, such as sensor nonlinearities and digitization, or from "nonphysical" boundaries, such as shadows. *Boundary classification* refers to detecting and labeling boundaries according to their physical origins, but it is rarely attempted and appears essentially impossible, at least without additional information from multiple views or sensors, temporal sequences, or specific scene knowledge.

Segmentation is a closely related problem; one seeks to partition an image into disjoint regions (pixel classes) on the basis of local properties such as color, depth, texture and surface orientation, or on the basis of more global (or even semantical) criteria, for instance involving dichotomies such as "object-background" or "benign-malignant." Clearly, each partition induces a unique boundary "map," whereas only boundary maps that are sufficiently organized yield useful segmentations.

These problems are studied extensively in computer vision, and there are many concrete applications. For example, texture is a dominant feature in remotely sensed images, and texture segmentation is important in the analysis of satellite data for resource classification, crop assessment, and geologic mapping. Other applications include automated navigation and industrial quality control; for example, in silicon wafer inspection, low magnification views of memory arrays appear as highly structured textures. Nonetheless, most work on boundary detection and segmentation has been of a general nature, separated from specific problems, and regarded as a "preprocessing" step toward further analyses such as extracting three-dimensional shape attributes (see §2.3.4), object recognition, and full-scale scene interpretation. In the view of some researchers, including the authors, this modular approach to image analysis is highly suspect, and generic segmentation is overemphasized.

Algorithms abound for "edge detection," which refers to the problem of locating the individual changes in intensity independently of the overall scene geometry. Most of these algorithms are heuristic, partly because it is

difficult to state these problems in precise mathematical terms. Traditional edge detectors [17, 20] are deterministic procedures based on (discrete) differential operators. Boundaries are then regarded as well-organized subsets of edges and the construction of boundary maps is usually considered as a second phase in which the detected edges are "cleaned," "smoothed," and otherwise massaged into structures we associate with real boundaries. A variational method proposed in [22] for boundary detection has lead to interesting mathematical problems, but its practical utility is uncertain. Statistical methods for segmentation in remote sensing are very prevalent [9, 25] and often successful. However, until recently most techniques employed *non-spatial* methods, such as linear discriminant analysis. Statistical methods that are truly spatial are currently gaining popularity, and an example involving texture segmentation is presented in §2.4.2.

2.3.3 Tomographic Reconstruction

Tomography is an imaging technology widely used in medical diagnosis (and also in industrial inspection and other areas). The two basic types are *transmission* and *emission* tomography. The most commonly used form of transmission tomography is the "CAT-scan," whereby a radiation source rotates about the patient's body and bombards it with X-rays or other atomic particles. Those particles that pass through the body are counted, and an image (or series of images) is formed from the combined counts; fewer counts correspond to regions of higher attenuation, which may, for example, indicate the presence of tumors.

In emission tomography, a pharmaceutical product is combined with a radioactive isotope and directed to a location in the patient's body, usually by injection or inhalation. The pharmaceutical is selected so that its concentration at the target location is proportional to some organic function of interest, for example, metabolic activity or local blood flow. The objective is then to reconstruct the (internal) two-dimensional or three-dimensional isotope concentration based on counting the number of released photons that escape attenuation and are registered by arrays of detectors placed around the body. For instance, in *positron emission tomography* (PET), the isotope emits a positron, which, upon colliding with a nearby electron, produces two photons propagating in opposite directions.

From here on the focus is on *single photon emission computed tomography* (SPECT), in which the isotope releases a single photon each time a radioactive decay occurs. Arrays of collimators are placed around the area

of interest and capture photons that escape attenuation and whose trajectories carry them down the collimator. The problem is then to reconstruct the isotope concentration from the photon counts and is similar to the inverse problems mentioned in §2.3.1.

The dominant physical effect is *photon attenuation*, by which photons are annihilated and their energy absorbed by body matter. Other significant effects are photon scattering and background radiation, as well as those effects induced by the sensors. Attenuation is accurately described by a single function μ whose values are known for bone, muscle, and so on, and for various photon energies. The incorporation of scattering and other effects is more subtle [13].

We formalize the SPECT reconstruction problem as follows [12, 27]. Let $x(u)$ be the isotope density defined on some domain S_0, which is usually taken as a two-dimensional region corresponding to a cross section of the body. The detectors σ_j, $j = 1, \ldots, m$, are arranged on a linear array at equally spaced intervals, and the array is positioned at equally spaced angles θ_k, $k = 1, \ldots, n$, for the same time duration at each angle. Let $T = \{(\sigma_j, \theta_k) : j = 1, \ldots, m;\ k = 1, \ldots, n\}$. The total number of observed photons is an array $y = \{y_t : t \in T\}$. Assuming that the photons generated in regions S_0 are governed by a spatially nonhomogeneous Poisson process with mean $x(u)$ at the point $u \in S_0$, the observation y is a realization from another nonhomogeneous Poisson process, $Y = \{Y_t : t \in T\}$, with mean

$$EY = Ax\,, \tag{2.3}$$

where the linear operator A incorporates attenuation (via the *attenuated Radon transform*) and other physical effects [13]. Hence, the conditional probability distribution of Y given x is

$$P(Y = y|x) = \prod_{t \in T} \frac{(Ax)_t{}^{y_t}}{y_t!} \exp\{-(Ax)_t\}\,. \tag{2.4}$$

For computational purposes, the region S_0 is discretized into a grid S of pixels. Then $x = \{x_i : i \in S\}$ represents a piecewise constant approximation of $x(u)$, and the operator A becomes a matrix $A = \{a_{t,i}\}$, $t \in T$, $i \in S$.

The oldest, and still current, method used for reconstructing x from y is *back-projection*, which essentially inverts (2.3), and is not very accurate. A more recent method [27] is that of *maximum likelihood* (ML), i.e., maximize $P(y|x)$ with respect to x, which is implemented with the expectation maximization (EM) algorithm. However, it has been widely realized that the ML

estimator is too "rough." Various procedures for smoothing this estimator have been suggested (see the references in [9]), including penalized ML and the method of sieves. Another alternative within the Bayesian framework is described in §2.4.2.

2.3.4 Three-Dimensional Shape Reconstruction

The problem of estimating or reconstructing properties of three-dimensional surfaces, such as orientation, height, Gaussian curvature, and Euler characteristics, from digital images is also widely studied in computer vision. In a common paradigm, especially in robot vision, these features are regarded as indispensable steps toward object recognition and other goals; the extraction of geometric features is followed by the imposition of relational structures among the features (using "grammars," symbolic graphs, and so on), and in turn by matching these data structures against stored models of real objects and spatial relationships among them. Again, the paradigm consists of distinct steps, and the extraction of geometric features is itself preceded by "pre-processing," which encompasses noise removal and other aspects of restoration, edge and boundary detection, and perhaps segmentation.

The main sources of information ("cues") for three-dimensional shape reconstruction are the intensity changes themselves ("shape-from-shading"), pairs of images corresponding to multiple views ("stereopsis") or wavelengths, motion sequences ("shape-from-motion"), texture analysis ("shape-from-texture"), and range data. Stereopsis is thought to be the most important process by which human beings obtain depth information, but has played a lesser role in automated systems due to computational demands and lack of accuracy. In contrast, shading has played a larger role in machine vision than in human vision, and range data is becoming an important source of information for automated systems but is unavailable to humans.

The mathematical problems encountered in three-dimensional shape reconstruction are similar to those in §§2.3.1–2.3.3. We conclude §2.3 with two examples in more detail.

Stereopsis

Stereo vision uses two cameras (corresponding to our two eyes) and a single light source. The relative difference, or so-called *disparity*, in the positions of an object in the two images is useful information for extracting surface orientation and relative distances between objects. A physical point in the viewed

scene has corresponding points in the two images and the *correspondence problem* is to find these pairs. The disparity between these points, together with simple geometric arguments, is then used [20] to estimate orientation or relative depth. The standard implementation of stereopsis [20] consists of the following steps: (1) detect significant intensity changes at various resolutions in both images, specifically, for example, "zero-crossings" of a Laplacian or other discrete differential operator or "filter"; (2) match these zero-crossings or properties thereof in the two images; (3) estimate from the disparity data the desired property of associated and sparse three-dimensional points; and (4) combine these data with regularization procedures to estimate entire surfaces. In addition, there are various pre-processing steps.

Shape-from-Shading

It is easier to state the problem in the fully continuous setup. The aim is to estimate a surface $z = z(u)$, $u = (u_1, u_2) \in \mathbf{R}^2$, from an observed image irradiance function $g(u)$ on the image plane $u \in \mathbf{R}^2$. The radiance (see §2.2) $f(u)$ is related to the geometry of the surface via the "irradiance equation" [17]:

$$f(u) = R[\vec{N}(u), \vec{S}, \vec{V}, \rho(u)], \tag{2.5}$$

where $\vec{N}(u)$ is the surface unit normal at the physical point (u_1, u_2, z), $\vec{S}$ and $\vec{V}$ are the directions of the illumination source and the camera, respectively, $\rho(u)$ is a property of the surface material (called albedo), and R is called the *reflectance map*. This function has been studied extensively [17] for many materials and illumination conditions, and we assume it to be known. In practice, not only $z(u)$ but also $\vec{S}$ and $\rho(u)$ may be unknown and require estimation. However, assuming that these are also known and that we observe $f(u)$ (or derive it from g as in §2.3.1), then equation (2.5) is a first-order differential equation for $z = z(u)$ over its domain $S_0 \subset \mathbf{R}^2$. The problem of estimating z is underdetermined unless one knows the normal vectors along occluding boundaries or along certain contours. Various numerical schemes have been used for solving this boundary value problem and for dealing with underlying instabilities. (These involve first detecting occluding boundaries.) Other approaches [18, 23] employ deterministic regularization, which leads to a second-order (elliptic) differential equation. An important issue in the discrete implementation of these methods is the incorporation of the *integrability condition*, i.e., of the discrete version of $z_{u_1,u_2} = z_{u_2,u_1}$.

Finally, there are many related problems, such as *photostereo*, in which one has two images acquired by a single camera but with separate light sources, and *shape-from-motion*, in which one is given a sequence of images induced by relative motion between the scene and the camera.

2.4 Bayesian/Spatial Statistics Framework

2.4.1 General Framework

Real scenes exhibit a variety of regularities: nearby locations typically have similar brightness; boundaries are usually smooth and persistent; textures, although possibly random locally, have spatially homogeneous regions; entities such as roads, leaves, and arteries have characteristic structures; object surfaces consist of locally smooth patches on which orientation and curvature change smoothly, whereas abrupt changes appear along object boundaries. The statistical variability of such regularities suggests a Bayesian formulation in which a priori knowledge and expectations are represented by a *prior distribution*. Spatial processes in general, and Markov random fields (MRF) in particular, provide flexible candidate distributions and the resulting framework supports reasonable computational algorithms, measures of performance, and inference procedures.

This framework consists of six basic steps: attribute modeling (i.e., choice of the prior), degradation modeling, computation of the posterior distribution, model identification, attribute estimation, and algorithmic implementation. These are described below; examples are given in §2.4.2. A more detailed exposition of the methodology and its applications is given in [9] (to which we shall refer for original and other references). See also [5, 24] for recent reviews and additional applications, [3] for seminal work on the role of MRF's in spatial statistics, and [15] for an early and influential paper on Bayesian scene modeling and stochastic relaxation.

Attribute Modeling

Scene attributes may often be regarded as two-dimensional arrays. Examples are intensity values corresponding to the "true" distribution of radiant energy; labels of textures, boundaries, or objects; and values of surface orientation, depth, or curvature. In some problems, we are interested in more than one array simultaneously. The collection of arrays of interest, denoted by X, is modeled as a discrete-parameter stochastic process indexed by the

vertices ("sites") of a graph $\mathcal{G}$. The set S of vertices of $\mathcal{G}$ serves simply to index the process, $X = \{X_i : i \in S\}$, whereas the edges or "bonds" of $\mathcal{G}$ capture the *interactions* among the individual random variables. If $\mathcal{N}_i$ denotes the set of vertices connected to site $i \in S$, then $\mathcal{N} = \{\mathcal{N}_i : i \in S\}$ defines the *neighborhood system* for $\mathcal{G}$, and we identify $\mathcal{G}$ with $(S,\mathcal{N})$. The graph $\mathcal{G}$ is usually sparse in the sense that the neighborhoods $\mathcal{N}_i$ are small compared to the graph size, and are usually "local" as well in the sense that the neighbors of i are spatially near i, which is rather natural for modeling spatial coherence and spatial context, and it is very convenient in computations.

The probability law, $\pi(x) = P(X = x)$, is the prior distribution, usually chosen to be a Gibbs distribution (i.e., X is a MRF with respect to $\mathcal{G}$) meaning that

$$\pi(x) = \frac{1}{Z} e^{-U(x)}, \tag{2.6}$$

Z being a normalizing constant called the *partition function* and $U(x)$ an *energy function* which is usually locally composed: $U(x) = \sum_{i \in S} \Phi_i(x_i, x(\mathcal{N}_i))$, $x(\mathcal{N}_i) = \{x_j : j \in \mathcal{N}_i\}$. The choice of the neighborhoods and "interactions" Φ_i is problem-specific. As a simple example, amplified in later sections, let X denote the true intensity values, S the regular pixel grid with a four nearest neighbor system (i.e., $\mathcal{N}_i = \{j \in S : |i-j| = 1\}$), and consider the problem of modeling spatial cohesion. We then might choose

$$\Phi_i(x_i, x(\mathcal{N})) = \sum_{j \in \mathcal{N}_i} \phi(|x_i - x_j|),$$

where $\phi(\cdot)$ is increasing on $[0, \infty)$. In this way, the measure π favors configurations in which nearby pixels have similar gray levels.

Sometimes it is convenient (if not necessary) to allow "infinite energies" (zero probabilities) in the prior. For example, in boundary detection, rather than stochastically inhibiting "blind" boundary endings and redundant boundary representations (see §2.4.2), it is more appropriate to simply disallow, or forbid, such configurations. These "constraints" are realized with a nonnegative "penalty function" $V(x)$. Then the prior is defined on the "allowed" set $\{x : V(x) = 0\}$, i.e.,

$$\pi(x) = \frac{1}{Z}\, \delta_{\{V=0\}}(x) e^{-U(x)}. \tag{2.7}$$

Degradation Model

By design, $X = \{X_i : i \in S\}$ contains all the relevant information for inference and decisionmaking. The goal is to estimate X based on the prior distribution and the observed data y, which is assumed to be a realization of an *observation process* $Y = \{Y_t : t \in T\}$ indexed by a discrete set T that might be different from S. The observation y may be a collection of multiple arrays available from multiple sensors, views, wavelengths, etc. The process Y is related to X by a conditional probability $P(Y|X)$—the *degradation model*—which may or may not be degenerate. This model, or the transformation from x to y, is problem-specific, and most often nonlinear. In the restoration problem, the degradation model is induced by (2.2); in boundary detection and segmentation, it is a projection; in tomography, it is given by (2.4); in shape-from-shading, it is induced by (2.5) (and (2.2)); and in other problems, it may involve "missing" observations (e.g., due to obscuration).

Posterior Distribution

The prior $\pi(x)$ and degradation model $P(Y|X)$ determine the joint distribution of (X, Y), and in particular the *posterior distribution* $\pi(X|Y)$ of X given Y. For a given observation $Y = y, \pi(X|y)$ contains all the relevant information about X, i.e., it reveals the likely and unlikely states of the "true" but unobserved attributes X. For a wide class of degradation models, $\pi(X|y)$ is again a Gibbs distribution over a new graph $\mathcal{G}^P$, which is in general larger than $\mathcal{G}$ but still sparse. However, exceptions occur; for example, in tomography (§2.4.2), $\mathcal{G}^P$ is highly nonlocal, and its complexity depends on the matrix A of (2.3).

Given $\pi(X)$ and $P(Y|X)$, the posterior $\pi(X = x|Y = y)$ is derived by Bayes' rule, and has the form

$$\pi(x|y) = \frac{1}{Z(y)} e^{-\tilde{U}(x|y)} \tag{2.8}$$

in case (2.6), and

$$\pi(x|y) = \frac{1}{Z'(y)} \delta_{\{V=0\}}(x) e^{-\tilde{U}(x|y)} \tag{2.9}$$

in case (2.7). Here, $\tilde{U}(x|y)$ is the energy associated with the posterior distribution; it is computed from $U(x)$ and the degradation model.

Model Identification

Both the prior $\pi(X)$ and the degradation $P(Y|X)$ may contain unknown parameters which need to be estimated from the data. The estimation of the parameters is often difficult and computationally expensive, and it is regarded by some as a serious drawback of the Bayesian framework. Its difficulty stems in part from the complexity of the likelihood function, i.e., the marginal distribution of Y. Other complicating factors include the high dimensionality of the data and the strong dependence of the individual random variables.

Several methods have been developed for estimating the parameters (see references in [9]): for *complete data* (i.e., observable X), ML via a stochastic gradient algorithm, maximum pseudo-likelihood (MPL), variational methods [2], coding, and logistic-like methods; for *incomplete data*, ML via the EM algorithm [12], and the method of moments [12, 13]. The problem of parameter estimation has given rise to interesting mathematical questions, and to an interplay between statistical inference and the phenomena of phase transitions [7].

Attribute Estimation

The ultimate goal, of course, is to choose a particular estimate, $\hat{x} = \hat{x}(y)$, of the attributes X given the data y. One choice is the MAP (maximum a posteriori) estimator, which is the mode of $\pi(x|y)$; it is the Bayes estimator corresponding to the zero-one loss function. Another Bayes estimate is the mean of $\pi(x|y)$, which derives from squared-error loss. Estimates of X are obtained by the basic algorithms described next. This estimation is distinct from the parameter estimation problem, but, in some cases, the two have been treated simultaneously [4].

Algorithms

The conditional distribution $\pi(x|y) = \tilde{\pi}(x)$ is usually too complex to allow a direct computation of $\hat{x} = \hat{x}(y)$. Instead, Monte Carlo type algorithms (motivated by analogies with physical systems in statistical mechanics) are used to generate sample realizations from $\tilde{\pi}(x)$, approximate global expectations with ergodic averages, and estimate modes by "annealing."

For example, the MAP estimate of (2.8) amounts to finding a minimal energy state of $\tilde{U}(x|y) \equiv \tilde{U}(x)$. Physically, a low-energy state is achieved

by heating and then slowly cooling a substance—a procedure called *annealing*. This suggests searching for global minima of $\tilde{U}(x)$ by simulating the dynamics of annealing using the *Metropolis Algorithm* (MA) (see references in [1]) or variants such as the *Gibbs Sampler* (GS) [10] (also called stochastic relaxation) and the *Langevin equation*. These algorithms generate a Markov chain $X(t)$ with transition probabilities arranged so that the equilibrium distribution is $\tilde{\pi}(x)$. For example, in GS one chooses a sequence of sites $i(1), i(2), \ldots$ so that each site in S is visited infinitely often. If, say, $X(t) = x$, then $X_j(t+1) = x_j$ for all $j \neq i(t)$, and $X_{i(t)}(t+1)$ is a sample from the conditional probability

$$\tilde{\pi}\left(X_{i(t)} = \cdot \,|\, X_j = x_j,\, j \neq i(t)\right) = \tilde{\pi}\left(X_{i(t)} = \cdot \,|\, X_j = x_j,\, j \in \mathcal{N}^P_{i(t)}\right), \quad (2.10)$$

where $\{\mathcal{N}^P_i\}_{i \in S}$ is the posterior neighborhood system. An important feature of these algorithms is that there is no need to compute the partition function $\tilde{Z}(y)$, which is intractable in general.

To simulate annealing, one introduces an artifical "temperature" (or control parameter) $T(t)$ into the posterior distribution. Let

$$\tilde{\pi}_{T(t)}(x) = \frac{1}{\tilde{Z}_T(y)} \exp\{-\frac{1}{T(t)}\tilde{U}(x)\}. \quad (2.11)$$

Now let $T(t) \downarrow 0$ as $t \to \infty$ sufficiently slowly (e.g., $T(t) \geq C(1 + \log t)^{-1}$, C small) so that the nonstationary Markov chain $X(t)$ converges weakly to a distribution supported by the global minima of $\tilde{U}(x)$ [10, 1]. The annealing algorithm (AA) has also been modified to deal with the constrained optimization problem underlying (2.9); see [9].

These algorithms are computationally demanding, but parallelizable. In practice, one compromises between the theoretical algorithms and practicality via such methods as low-temperature sampling [11], iterated conditional modes (ICM) [4], iterated conditional expectation (ICE) [13], and the renormalization group algorithm [14].

2.4.2 Examples

Image Restoration

The basic degradation model is given by (2.2). For simplicity we assume the presence of only one noise process $\{\eta_i\}$, and $N = N'$. Although not necessary, we assume that the intensities are quantized. We will denote the

pixel lattice by $S^P = \{i = (i_1, i_2): 1 \le i_1, i_2 \le N\}$, the grey-level intensity process by $X^P = \{X_i^P : i \in S^P\}$, and the observation process by $Y = \{Y_i : i \in S^P\}$. Then

$$Y_i = \psi\{\phi[(KX^P)_i], \eta_i\}, \ i \in S^P . \tag{2.12}$$

We assume that K, ϕ, ψ, and the law of η are known.

The basic idea proposed in [10] is to use MRF models to account for the spatial coherence among the intensity levels of nearby pixels, as well as for the existence of discontinuities between intensity surfaces. To this end, in addition to the intensity process X^P, a second "edge" process X^E is introduced. The process X^E accounts for discontinuities, and is indexed by the dual lattice S^E of S^P; S^E consists of all nearest-neighbor pairs $< i, j >$ from S^P, and an element ("site") $t \in S^E$ corresponds to a putative edge between the corresponding pixel sites. Thus, $X^E = \{X_t^E : t \in S^E\} = \{X_{<i,j>}^E : i, j \in S^P\}$, $X_t^E \in \{0, 1\}$ where $X_t^E = 1$ (resp. 0) indicates presence (resp. absence) of an edge at $t \in S^E$. The process X^E is neither part of the data nor the target of estimation; rather, it is an auxiliary process designed to bring exogenous information into the model, and it is coupled to X^P in such a manner that in the likely states of the joint probability distribution of $X = (X^P, X^E)$, the intensity function is locally smooth with possibly sharp transitions, and the locations of the edges satisfy our a priori expectations about the behavior of boundaries.

The process $X = (X^P, X^E)$ is indexed by $S = S^P \cup S^E$, and is chosen to be a MRF with energy function of the form

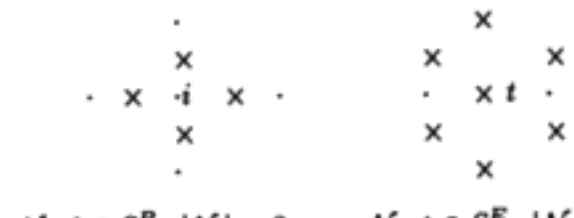
$$U(x) = U(x^P, x^E) = U_1(x^P, x^E) + U_2(x^E), \tag{2.13}$$

where U_1 reflects our expectations about interactions between intensities and edges (where edges "belong"), while U_2 reflects boundary organization. Both terms are constructed from "local energies" corresponding to a neighborhood system $\mathcal{N} = \{\mathcal{N}_\alpha : \alpha \in S\}$. The simplest neighborhood system ("nearest neighbors") is shown in Figure 2.1, where dots denote pixels and crosses edge sites.

The energy $U_1(x^P, x^E)$ is defined so that the low energy states will have $x_{<i,j>}^E = 1$ (resp. 0) when $|x_i^P - x_j^P|$ is large (resp. small). More specifically,

$$U_1(x^P, x^E) = \theta_1 \sum_{<i,j>} \phi(x_i^P - x_j^P)(1 - x_{<i,j>}^E), \tag{2.14}$$

with $\theta_1 > 0$, $\phi(0) = -1$, and ϕ even and nondecreasing on $[0, \infty)$. Note that when $x_{<i,j>}^E = 1$, the bond ("interaction") between pixels i and j is broken;



$\mathcal{N}_i,\ i \in S^P,\ |\mathcal{N}_i| = 8$ $\qquad$ $\mathcal{N}_t,\ t \in S^E,\ |\mathcal{N}_t| = 8$

FIGURE 2.1

also, the properties of ϕ ensure that when $x_i^P = x_j^P$, $|i-j|=1$, then $x^E_{<i,j>}=0$ is a lower energy state than $x^E_{<i,j>}=1$ and, in the absence of boundaries, there is cooperation between nearby pixel intensities. Note also that when $\phi(x_i^P - x_j^P)=0$, we have no preference about the state of an edge at site $<i,j>$. A specific $\phi(\cdot)$ that has been used in restoration problems [9] is

$$\phi(\xi) = 1 - 2\left[1 + \left(\frac{\xi}{\Delta}\right)^2\right]^{-1}, \quad \xi \in [-K, K], \tag{2.15}$$

where Δ is a scaling constant.

The energy $U_2(x^E)$ reflects our prior expectations about boundaries: most pixels are not at boundary regions; boundaries are usually persistent (no isolated or abandoned segments); boundary intersections (junctions), sharp turns, and "small structures," are relatively unlikely. For specific choices of $U_2(x^E)$, we refer to [9] (and references cited there). Some of the above generic properties of boundaries may also be captured with penalty functions (see (2.7)). For problems, such as image restorations and shape-from-shading, where the edge process X^E is auxiliary, using "soft constraints" such as $U_2(x^E)$ is satisfactory, whereas for problems in which the boundaries are of central interest (e.g., texture segmentation; see the next subsection), penalty functions are often more appropriate.

The energy function (2.13) determines the prior $\pi(x)$. Assuming that the noise process $\{\eta_i\}$ is independent of $X = (X^P, X^E)$, the transformation (2.12) determines the degradation model $P(Y|X)$ in a straightforward manner. The posterior distribution $\pi(X|Y)$ is then computed from the joint distribution $P(Y|X)\pi(X)$. For example, suppose that η_i is Gaussian with mean μ and variance σ^2, and that $b \mapsto \psi(a,b)$ is invertible, say,

$\eta_i = F(y_i, \phi[(Kx^P)_i])$ for some $F(u,v)$. Then assuming that $F(u,v)$ is strictly increasing and $F_1(u,v) = \frac{\partial}{\partial u}F(u,v)$ exists, the posterior energy (see (2.8)) is given by [9]

$$\tilde{U}(x|y) = U(x^P, x^E) + \frac{1}{2\sigma^2}\sum_{i\in S^P}\left|\mu - F\left(y_i\,,\,\phi[(Kx^P)_i]\right)\right|^2 - \sum_{i\in S^P}\log F_1\left(y_i\,,\,\phi[(Kx^P)_i]\right)\,. \qquad (2.16)$$

Boundary Detection

We summarize the procedure developed in [11] for detecting textural boundaries. The method also applies to the problem of locating boundaries representing sudden changes in depth and surface orientation.

The procedure involves two processes: the *intensity process* $X^P = \{X_i^P: i \in S^P\}$, as in the first subsection of §2.4.2, and the *boundary process* $X^B = \{X_t^B: t \in S^B\}$, which is indexed by a regular lattice S^B with sites interspersed among the pixels of S^B. Again, $X_t^B = \{0,1\}$ with $X_t = 1$ (resp. 0) indicating presence (resp. absence) of a boundary at $t \in S^B$. The prior distribution for $X = (X^P, X^B)$ is chosen to be of the form (2.7) with $U(x) = U(x^P, x^B)$, and $V(x) = V(x^B)$.

The intensity-boundary term $U(x^P, x^B)$ is chosen to promote placement of boundaries between regions in the image that demonstrate distinct spatial patterns. In [11] it was chosen to be of the form

$$U(x^P, x^B) = \sum_{<t,s>}\Phi_{t,s}(x^P)(1 - x_t^B x_s^B),\ t,s \in S^B\,, \qquad (2.17)$$

where $< t,s >$ denotes nearest-neighbors in S^B. The function $\Phi_{t,s}$ is critical; it is a *measure of disparity* between the gray-level values in two blocks of pixels adjacent to $< t,s >$; large disparities ($\Phi \gg 0$) encourage the presence of an active boundary (i.e., $x_t^B = x_s^B = 1$), while small disparities ($\Phi \ll 0$) discourage the presence of a boundary (i.e., $x_t^B x_s^B = 0$). Specific choices of $\Phi_{t,s}$ in [11] are constructed in terms of the Kolmogorov-Smirnov distance applied to either the raw data ("first-order" statistics), or to transformed data corresponding to higher-order statistics (e.g., window means, range, variance, and "directional residuals"). As a function of x^B, (2.17) is similar to "spin glass" models in statistical mechanics.

The penalty function $V(x^B)$ is chosen to inhibit unwanted configurations such as blind endings of boundaries, redundant boundaries, sharp turns, and other forbidden patterns (see [11] for details).

Assuming that there are no degradations that preclude direct observation of x^P, then the data $y \equiv x^P$. The degradation model $P(Y|X^B)$ is singular (the point mass at $y = x^P$), and the posterior energy is $\tilde{U}(x^B|y = x^P) = U(y{=}x^P, x^B)$. The MAP estimate is equivalent to global optimization with constraints.

Plates 2.1 and 2.2 (from [11]) show two experiments. Plate 2.1 is an L-band synthetic aperture radar (SAR) image of ice floes in the ocean: (a) original image, 512 × 512, (b) sixteen "snapshots" from sixty sweeps of stochastic relaxation with constraints. Plate 2.2 is a collage composed of nine Brodatz textures: leather, grass, and pigskin (top row), raffica, wool, and straw (middle row), and water, wood, and sand (bottom row). Two of the textures, leather and water, are repeated in the two circles; (a) original 384 × 384, (b) detected boundaries obtained by deterministic (left) and stochastic (right) algorithms.

Single Photon Emission Tomography

The digitized isotope intensity (see §2.3.3) is thought to be a realization of a spatial process $X = \{X_i : i \in S\}$. The idea of [12, 13] is to use a Gibbs prior to reflect the common observation that neighboring locations of the isotope surface typically have similar concentration levels, whereas sharp changes in concentration may occur, for instance, across an arterial wall or across a boundary between two tissue types. In contrast to the procedure in restoration, sharp changes are not represented explicitly by an edge or boundary process; instead, the intensity model is designed to allow such changes. Specifically (*cf.* [12]),

$$U(x) = \beta \sum_{<i,j>} \phi(x_i - x_j) + \frac{\beta}{\sqrt{2}} \sum_{[i,j]} \phi(x_i - x_j)\,, \tag{2.18}$$

where $< i, j >$ denotes a nearest-neighbor bond, $[i, j]$ represents a nearest-neighbor-diagonal bond, and ϕ is given by (2.15).

The degradation model is given by (2.4), and the posterior energy is then

$$\tilde{U}(x|y) = U(x) + \sum_{t \in T} [(Ax)_t - y_t \log(Ax)_t]\,.$$

Although $U(x)$ has a local structure, the graph for $\tilde{U}(x|y)$ is highly nonlocal due to A.

The choice of the "smoothing" parameter β is critical. For $\beta = 0$, the MAP estimator is just the ML estimator and, hence, typically too rough. For

large β, the MAP estimator is too faithful to the prior and, hence, typically too smooth. The value of β is estimated in [12, 13] via the EM algorithm or the method of moments. The parameter Δ in ϕ is also important, but its statistical estimation from the data appears to be difficult. Fortunately, reconstructions are not sensitive to moderate changes in Δ, and empirical values based on information about range intensities work well [12].

Plates 2.3, 2.4, and 2.5 show three experiments from [13] with real (hospital) data, using the ICE algorithm. For comparison, the reconstruction with the filtered back projection (FBP) method is also shown. In all three cases the β is estimated by the ML method. Plate 2.3 shows a slice of a human skull across the eyes: (a) FBP, (b) ICE with $\beta = 2.7$. Note that in (b) one can distinguish details such as the nose bone, eyes, brain region; also the skull border is sharp. Plate 2.4 displays a SPECT reconstruction of a simulated phantom. The model used in this experiment was developed by the Nuclear Medicine Department of the University of Massachusetts Medical Center, in Worcester. This is a comprehensive model that captures the effects of photon scattering, photon attenuation, camera geometry, and quantum noise: (a) original phantom, (b) FBP reconstruction, (c) ICE reconstruction with $\beta = 1$. Plate 2.5 is a human liver/spleen scan: (a) FBP, (b) ICE with $\beta = 3$. The value $\beta = 3$ is the ML estimation; (c) and (d) are ICE reconstructions with $\beta = 0$ and $\beta = 20$ respectively; they demonstrate the significance of the parameter β.

Shape-From-Shading

We focus on the estimation of surface orientation. For simplicity, we assume that $\vec{S}$ and ρ are known, that the reflectance map is spatially homogeneous and known, and that $\vec{V}$ is constant throughout the image (orthographic projections). However, the procedure presented below can be modified [28] to estimate also $\vec{S}$ and ρ. There are three basic processes: The true (undegraded) intensity process $X^P = \{X_i^P : i \in S\}$, the *shape process* $N = \{\vec{N}_i : i \in S\}$ where $\vec{N}_i$ is the unit normal at the surface "point" corresponding to pixel i, and the observation process $Y = \{Y_i : i \in S\}$. Here, S is a discretization of the domain of $z(u)$ (see §2.3.4). The shape process N (the target of estimation) is related to X^P via the discrete version of (2.5), and X^P is related to Y via (2.12). X^P is an indeterminate process and will play no direct role here.

In the absence of degradation we have $Y_i = R(\vec{N}_i)$, which is a deterministic constraint on N. The shape-from-shading problems usually refer (see

[17]) to this case (i.e., observable X^P), and we refer to [17,18] for various approaches to the problem including deterministic regularization techniques.

The procedure below was developed in [28] and applies to both undegraded and degraded data. The basic idea is to use Gibbs distributions to articulate general properties of shapes: surfaces are locally smooth, orientation may exhibit jumps because of changes in depth or presence of surface discontinuities. As in restoration, the process N is coupled to an edge process $X^E = \{X_t^E : t \in S^E\}$, where X_t^E and S^E are as in the first subsection of §2.4.2. The coupled process $X = (N, X^E)$ is a MRF with an energy function $U(x) = U_1(N, x^E) + U_2(x^E)$ (compare with (2.13)), where $U_2(x^E)$ is chosen as in the first subsection of §2.4.2, and

$$\begin{aligned} U_1(N, x^E) &= \theta_1 \sum_{<ij>} (1 - x^E_{<i,j>})\, \phi(|\vec{N}_i - \vec{N}_j|) \\ &+ \theta_2 \sum_{[i,j]} (1 - h^E_{[i,j]})\, \phi(|\vec{N}_i - \vec{N}_j|), \end{aligned} \tag{2.19}$$

with $\theta_1, \theta_2 > 0$, $< i, j >, [i, j]$ as in (2.18), and $h^E_{[i,j]} = 1$ if any of $x^E_{t_1} x^E_{t_2} = 1$, $x^E_{t_1} x^E_{t_3} = 1$, $x^E_{t_2} x^E_{t_4} = 1$, $x^E_{t_3} x^E_{t_4} = 1$, are true, zero otherwise; here i, j, t_1, t_2, t_3, t_4 are as in Figure 2.2.

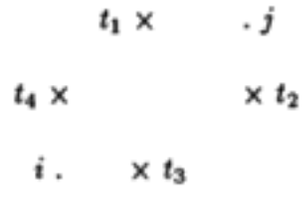


FIGURE 2.2

The function ϕ was chosen [28] to be $\phi(|\vec{N}_i - \vec{N}_j|) = -1 + \frac{1}{2}|\vec{N}_i - \vec{N}_j|^2 = -\vec{N}_i \cdot \vec{N}_j$. Because of the constraint $|\vec{N}_i| = 1$, the prior resulting from this choice of ϕ is non-Gaussian even if $\theta_2 = 0$ and $x^E = 0$. In fact, model (2.19) has worked well in some cases [28] even without the edge process (i.e., $x^E = 0$).

In the absence of degradation, the distribution $P(Y|X)$ is degenerate and amounts to the constraint $V_1(N) = \sum_{i \in S} |Y_i - R(\vec{N}_i)|^2 = 0$. In the presence of degradation, $P(Y|X)$ is computed as in the first subsection of §2.4.2. In both cases, there is an extra deterministic constraint $V_2(N) = 0$

corresponding to the integrability condition. In the former case, the shape-from-shading problem reduces to minimizing $U(N, x^E)$ under the constraint $V_1(N) + V_2(N) = 0$; in the latter case, to minimizing the posterior energy $\tilde{U}(N, x^E)$ under the constraint $V_2(N) = 0$.

The procedure does not require boundary conditions. However, in the absence of degradation and a single image, the quality of reconstruction is better [28] if one assumes correct boundary conditions along the image border. In the presence of noise, the results are satisfactory if one uses two images (photostereo) obtained with a single camera but two light sources of different origin.

Plate 2.6 shows an experiment from [28] with an egg imaged under uncontrolled illumination. The surface of the egg was assumed to be matte, and the algorithm estimated, in addition to N, the albedo ρ and an effective light source direction $\vec{S}$. The reconstruction used a combination of constrained annealing and ICM: (a) original image, 64×64, (b) reconstruction, (c) reconstructed egg illuminated from x-direction, and (d) reconstructed egg illuminated from y-direction.

Deformable Templates

In this subsection we briefly describe a powerful and elegant methodology introduced by Ulf Grenander for pattern synthesis and analysis of biological shapes. It provides a promising geometric/Bayesian paradigm for medical and industrial applications.

The procedure [6] is based on global shape models designed to incorporate prior (biological) knowledge about the variability and global properties of shapes, and quantitative information about the variability of the geometric object. The shape model has three basic components: (a) a "geometry" consisting of a space G of generators, a connector graph σ, a "regularity" relation R, and a transformation group $S{:}G \to G$; (b) a template (or templates) describing the overall architecture of the shape; and (c) a group-valued stochastic (typically Markov) process, which articulates the statistical variations of the shape. The choices of the template, transformation group, and stochastic process control the desired global and local geometric properties of shapes. The choice of (a) is application-specific.

We refer to [6] for the general framework, experiments, and references. Here we outline the method for the special case of two-dimensional (planar) shapes (as in the HAND experiment [6]). Assuming that all the relevant information is contained in the boundary, and approximating the boundary

by a polygon with, say, n edges, the components of (a) are as follows: $g_i \in G$, $j = 0, 1, \ldots, n-1$ are polygonal edges; σ is a cyclic graph consisting of the n nodes; R may, for example, be the condition that the polygon is closed and simply connected (other regularity conditions are often desirable); S may be chosen to be the general linear group $GL(2)$, or the Euclidean group (i.e., rotations $O(2)$ and translations), or the group $US(2)$ of uniform dilations (scale), or $US(2) \times O(2)$.

The configuration space of interest is $C(R) = \{c = \sigma(g_0, \ldots, g_{n-1})\colon\ g_i \in G,\ j = 0, \ldots, n-1,\ c \text{ satisfies } R\}$, i.e., the set of boundaries of closed, simply connected polygons. The interior of a polygon defines a *pure image* I. A template is a specific configuration $c^{(0)} = \sigma(g_0^{(0)}, \ldots, g_{n-1}^{(0)}) \in C(R)$ that represents the prototypical shape being considered; for example, in the HAND experiment [6], $c^{(0)}$ is an "ideal" male hand computed by "averaging" several male hands. The purpose of the group-valued process is to define a prior distribution on $C(R)$. This is done as follows: let μ be a fixed measure on S, and $\pi(s_0, \ldots, s_{n-1})$, $s_j \in S$, a probability density (w.r.t. μ) on S^n. In [6], π was chosen to be of Gibbs type with nearest-neighbor interactions, i.e.,

$$\pi(s_0, \ldots, s_{n-1}) = \prod_{i=0}^{n-1} A(s_i, s_{i+1}),$$

with $A \in L_2(S \times S, \mu \times \mu)$, so that $\pi(S^n) = 2$ (the actual form of A is dictated by applications; see [6] for examples). This probability distribution is now restricted (conditioned) by using the template $c^{(0)}$, and considering only those s-sequences for which $\sigma(s_0 g_0^{(0)}, \ldots, s_{n-1} g_{n-1}^{(0)}) \in C(R)$. For special cases, this conditioning is straightforward; in general, it involves subtle limit arguments. This results in a probability measure P—the *prior*—on $C(R)$.

In the above framework, shape synthesis amounts to simulation of P. Samples from P reflect the variability of the shape under consideration. For example, in the HAND experiment [6], the variability accounts for differences between hands of individuals, as well as for possible hand shapes (e.g., position of fingers) of a given individual.

For analysis tasks such as restoration, segmentation, detection of anatomical pathologies, recognition, and so on, inferences are made on the basis of the prior P and the data. Let us consider restoration, for example. The procedure not only gives a restored image, but it also yields a "structured" restoration, in the sense that it provides the configuration analysis of it. This automatically makes possible, for instance, more challenging problems such as finding statistically meaningful abnormalities. Suppose that we ob-

serve a degraded version I^D of a true pure image I which is our target of estimation. The degradation mechanism $I \rightarrow I^D$ defines the degradation model $P(I^D|I)$, equivalently $P(I^D|c)$, as in the first subsection of §2.4.2. Then the posterior distribution $P(c|I^D)$ is computed as before, and contains all the relevant information about the unknown image I (equivalently c). The computational burden of estimating I is demanding, but feasible. Asymptotic arguments and other specific properties have been exploited in [6] to reduce the computations. One important aspect of the approach is that it can often be combined with dynamic programming to speed up the processing considerably.

Bibliography

[1] Aarts, E., and J. Korst, *Simulated Annealing and Boltzmann Machines*, John Wiley and Sons, 1989.

[2] Almeida, M., and B. Gidas, A variational method for estimating the parameters of MRF from complete or incomplete data, preprint, Brown University, 1989.

[3] Besag, J., Spatial interaction and the statistical analysis of lattice systems (with discussion), *J. R. Stat. Soc., B* **36** (1974), 192–236.

[4] Besag, J., On the statistical analysis of dirty pictures (with discussion), *J. R. Stat. Soc., B* **48** (1986), 259–302.

[5] Besag, J., Towards Bayesian image analysis, *J. Appl. Stat.* **16** (1989), 395–407.

[6] Chow, Y., U. Grenander, and D. Keenan, *HANDS, A Pattern Theoretic Study of Biological Shapes*, to be published by Springer-Verlag, 1990.

[7] Comets, F., and B. Gidas, Parameter estimation for Gibbs distributions from partially observed data, submitted to *Ann. Appl. Prob.* (1989).

[8] Frieden, B. R., Restoring with maximum likelihood and maximum entropy, *J. Opt. Soc. Amer.* **62** (1972), 511–518.

[9] Geman, D., *Random Fields and Inverse Problems in Imaging*, to appear in Lecture Notes in Mathematics, Springer-Verlag, 1990.

[10] Geman, S., and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Machine Intell.* **6** (1984), 721–741.

[11] Geman, D., S. Geman, C. Graffigne, and P. Dong, Boundary detection by constrained optimization, *IEEE Trans. Pattern Anal. Machine Intell.* **12** (1990), 609–628.

[12] Geman, S., and D. McClure, Statistical methods for tomographic image reconstruction, in Proceedings of the 46th Session of the Int. Stat. Institute, *Bulletin Int. Stat. Inst.* **52** (1985).

[13] Geman, S., D. McClure, K. Manbeck, and J. Mertus, Comprehensive statistical model for single photon emission computed tomography, preprint, Brown University, 1990.

[14] Gidas, B., A renormalization group approach to image processing problems, *IEEE Trans. Pattern Anal. Machine Intell.* **11** (1989), 164–180.

[15] Grenander, U., *Tutorial in Pattern Theory*, Technical Report, Brown University, 1983.

[16] Gull, S. F., and J. Skilling, Maximum entropy methods in image processing, *IEE Proc.* **131** (1984), 646–659.

[17] Horn, B. K. P., *Robot Vision*, MIT Press, 1986.

[18] Horn, B. K. P., Height and Gradient from Shading, MIT A.I. Memo No. 1105.

[19] Jaynes, E. T., On the rationale of maximum entropy methods, *Proc. IEEE* **70** (1982), 939–952.

[20] Marr, D., *Vision*, W. H. Freeman, 1982.

[21] Marroquin, J. L., S. Mitter, and T. Poggio, Probabilistic solution of ill-posed problems in computational vision, *J. Am. Stat. Assoc.* **82** (1987), 76–89.

[22] Mumford, D., and J. Shah, Boundary detection by minimizing functionals, talk presented at the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, 1985.

[23] Poggio, T., V. Torro, and C. Koch, Computational vision and regularization theory, *Nature* **317** (1985), 314–319.

[24] Ripley, B. D., *Statistical Inference for Spatial Processes*, Cambridge University Press, 1988.

[25] Ripley, B. D., Statistics, images, and pattern recognition, *Can. J. Stat.* **14** (1986), 83–111.

[26] Serra, J., *Image Analysis and Mathematical Morphology*, Academic Press, New York, 1982.

[27] Shepp, L. A., and Y. Vardi, Maximum likelihood reconstruction in positron emission tomography, *IEEE Trans. Medical Imaging* **1** (1982), 113–122.

[28] Torreão, J., A Bayesian Approach to Three-Dimensional Shape Estimation for Robot Vision, thesis, Brown University, 1989.

3

Oceanographic and Atmospheric Applications of Spatial Statistics and Digital Image Analysis

James J. Simpson
Scripps Institution of Oceanography

3.1 Introduction

Historically, remote sensing of the environment has helped process-oriented studies examine individual aspects of the physics, chemistry, and biology of the earth-ocean-atmosphere system. From such studies, several problems of global significance have emerged that are cross-disciplinary in nature. Examples include acid rain, the increase in atmospheric carbon dioxide, anticipated depletion of the ozone layer, El Niño-related modifications in weather and ocean circulation with their resultant effects in agriculture and fisheries, and tropical rain forest destruction by fires of human origin. It has been recognized that the key to progress on these and other cross-disciplinary issues in earth science during the decade of the 1990s probably will be addressing those questions that concern the integrated functioning of the earth as a system (EOS, 1984). The hydrologic cycle, the biogeochemical cycle, and climate processes are the fundamental processes that integrate the earth as a system, and thus each of these cycles must be examined in detail and on a global scale if meaningful progress is to be made on the problems cited above.

Large-scale, synoptic observations of a wide variety of phenomena (e.g., sea surface temperature, sea level, wind stress, ozone concentration, radi-

ation heat balance, sea ice, vegetation, and near-surface current velocity) are required to study these cycles properly, and remote sensing provides the only practical way to collect synoptically the necessary data on the global scale. The magnitude of the data set, however, may preclude understanding unless it can be distilled and synthesized into organized patterns that can be related meaningfully to the underlying governing physics, chemistry, or biology of these global cycles and processes. Spatial statistics and mathematical methods of digital image analysis provide mechanisms for such a synthesis.

This chapter discusses a few areas of spatial statistics and data representation useful in digital image analysis that have immediate application in the area of remote sensing of the earth. Some special considerations needed for the correct analysis of remotely-sensed data are presented and a few critical research areas are identified.

3.2 Selected Analysis Areas

In this section three mathematical methods useful in digital analysis of sequences of remotely sensed images are presented. These methods were chosen because they have a broad range of applications in earth science. These methods are non-Bayesian in nature; a presentation on Bayesian methods used in image analysis is given in chapter 2 of this report.

Several abbreviations commonly used in remote sensing and the earth sciences appear in this chapter. Each is defined at the place it first occurs in the text. For easy reference, there is an appendix containing abbreviations and their definitions at the end of this chapter.

3.2.1 Principal Component Analysis

General Concepts

Principal component analysis (PCA) is a multivariate statistical technique that can be applied to all forms of multispectral or multi-temporal image data and is most commonly applied in the general context of arbitrary multivariate data. Forms of PCA have been used to study pattern classification (Geladi *et al.*, 1989), sequential segmentation (Esbensen and Geladi, 1989), spatial patterns of variability in sea surface temperature and phytoplankton pigment observed in satellite data (Lagerloef, 1986), and in algorithms for cloud removal from satellite data (e.g., Gallaudet and Simpson, 1991a).

The principal component transformation (PCT) is the basis for all these analyses. The PCT is a linear transformation that isolates uncorrelated linear combinations of a given set of variables in such a way that each element in this combination represents a decreasing amount of variance in the original variables (Ingerbritsen and Lyon, 1985). This linear transformation defines a new set of coordinate axes for the data such that (1) the transformed origin is at the mean of the data distribution (Lillesand and Kiefer, 1987), (2) the transformed coordinate axes are mutually orthogonal (Jensen, 1986), and (3) the transformed coordinate axes are in the directions of maximum variance (Jensen, 1986). Below, two examples of PCA are developed, one using multispectral image data and the other using multi-temporal image data. An extensive treatment of the use of PCA techniques in atmospheric sciences and oceanography is given by Preisendorfer (1988).

Multispectral Data Application

The ability to accurately and automatically segment clouds in remotely sensed imagery is critically important to a broad range of disciplines in earth science. Clouds significantly affect the net heating of the atmosphere and the underlying ocean-land surface by modifying solar and terrestrial radiation (Ohring and Clapp, 1980). This net radiative heating governs the thermodynamics and dynamics of the atmosphere, which in turn influence the formation and dissipation of clouds (e.g., Matveev, 1984). The potential feedback effects associated with this cloud-radiation interaction are among the greatest sources of uncertainty in determining the relation between changes in external conditions such as solar radiation and atmospheric carbon dioxide concentration and changes in climate (e.g., Henderson-Sellers, 1982; Ramanathan, 1987). Clouds also affect our ability to remotely sense the properties of the atmosphere, ocean, and land; such observations are needed, for example, in weather prediction (e.g., Pailleux, 1986), oceanography (e.g., Eckstein and Simpson, 1990a,b), and agriculture (e.g., rainfall, Browning, 1986). The PCT can be used with multispectral image data to robustly segment clouds from natural images. An example of such an application is given below.

The Principal Component Transformation Split-and-Merge Clustering (PCTSMC) Algorithm. Here, the development of the PCTSMC algorithm is presented in abbreviated form. Mathematical details of the algorithm are given in Gallaudet and Simpson (1991a). An AVHRR infrared

image, designated **T**, consisting of brightness temperatures calibrated to degrees Celsius in bands 3, 4, and 5 (hereinafter referred to as T_3, T_4, T_5) is first differenced to construct a 2-banded differenced image **D** (i.e., $T_3 - T_4$, $T_3 - T_5$). This differenced image now contains all the information of the original infrared image needed for cloud detection, but only in two linearly independent bands. Next, this differenced image is transformed using the PCT. The result is a 2-banded transformed differenced image in which all interband correlation is destroyed (Mather, 1987), thus making it a logical preprocessor to the segmentation operation (step 3) of the PCTSMC algorithm.

Step 3 of the PCTSMC algorithm performs image segmentation using a split-and-merge clustering procedure (e.g., Pavlidis, 1977; Seddon and Hunt, 1985; Richards, 1986) on the PC transform of the differenced image. This results in a segmented image in which the natural spectral classes in the original image are separated into distinct groups (i.e., land versus ocean versus cloud). The method of clustering that is employed in the PCTSMC algorithm combines both the partitional and hierarchical approaches. It consists of a partitional clustering algorithm augmented by a splitting-and-merging step at each iteration. Combining a partitional with a hierarchical method has several advantages over the use of either method alone: (1) pure hierarchical methods are not appropriate for complex data (Muerle and Allen, 1968; Fukada, 1980; Jain and Dubes, 1988); (2) pure hierarchical methods are more appropriate for data that is to be partitioned on the basis of both local and global information, rather than global information only (Jain and Dubes, 1988); (3) pure hierarchical methods impose a taxonomic structure on the data (Anderberg, 1973), which is not characteristic of cloud-containing AVHRR imagery; (4) pure hierarchical methods are *order* dependent—i.e., the resulting segmentation will vary depending upon the order in which the regions are split and merged (Cheevasuvit *et al.*, 1986); this often results in less than optimal segmentations of the data; (5) pure partitional algorithms often converge to local minima of the clustering criterion function (Pairman and Kittler, 1986; Jain and Dubes, 1988); (6) the combined approach is more efficient than pure merging or pure splitting methods of region detection (Pavlidis, 1977; Richards, 1986); and (7) the combined approach is less dependent on the initial segmentation, and therefore is more capable of recovering all of the natural clusters in the data (Seddon and Hunt, 1985; Jain and Dubes, 1988); this is because the number of clusters in the initial segmentation need not be the same as those

that actually exist in the given data (Seddon and Hunt, 1985; Pairman and Kittler, 1986).

In the final step of the PCTSMC cloud screening procedure, the data are retransformed back to the feature space, and each of the clusters is labeled as either a cloud or non-cloud. In the first three steps discussed above, the operations performed on the data were entirely unsupervised—i.e., no *a priori* knowledge was required. Hence, they apply to an arbitrary image segmentation. In this fourth step, expert knowledge is introduced to perform a boolean classification. Rules appropriate for land versus cloud versus ocean separation in AVHRR image data are given in Simpson and Humphrey (1990) and Gallaudet and Simpson (1991a).

Plate 3.1a shows AVHRR Band 2 data; clouds appear as white or gray tones. The coastline is white in this panel. Plate 3.1b shows AVHRR Band 4 infrared temperature; coldest temperature is white and warmest temperature is black. In this panel, the coastline is black. The segmented image produced by the PCTSMC algorithm is shown in Plate 3.1c and the final cloud-masked sea surface temperature is shown in Plate 3.1d. In this final panel, the warmest ocean temperature is white, cooler ocean temperatures appear as shades of gray, and cloud contaminated pixels and land are shown as black. Land was masked from the Plate 3.1 images using a recursive polygon fill algorithm (Simpson, 1991) and is rendered either white or black depending on the gray scale mapping used in the individual panel.

Multi-Temporal Data Application

A major objective of earth science studies is to identify spatial patterns of variance in temporal sequences of images. Examples include the analysis of variability in sea surface temperature structure in oceanic current systems (e.g., Lagerloef, 1986) and in seasonal and interannual variation in phytoplankton abundance (e.g., Strub *et al.*, 1990). The form of PCA used in such studies is generally referred to as empirical orthogonal function (EOF) analysis: "empirical" because the functions arise from the data themselves and "orthogonal" because they are uncorrelated. Note that closed-form mathematical functions generally cannot be used as the basis functions representing the complex images observed in nature.

Empirical orthogonal functions are useful in work with large data sets. The method separates a data set $D(x,t)$ into spatial components $F_i(x)$ and

temporal components $A_i(t)$ such that

$$D(x,t) = \sum_{i=1}^{NT} F_i(x) \cdot A_i(t), \tag{3.1}$$

where NT is the number of EOFs computed. Note that x and t represent generalized spatial and temporal coordinates, i.e., x represents the set $\{x_1, \ldots, x_{NX}\}$ and t represents the set $\{t_1, \ldots, t_{NT}\}$. In matrix notation, $\mathbf{D}(x,t)$ is an $NX \times NT$ matrix, with NX representing the number of positions x and NT the number of time steps t. Each column of the $\mathbf{F}$ matrix is an EOF. Each EOF has NX values, and there are NT EOFs. Thus, $\mathbf{F}$ is an $NX \times NT$ matrix. The time series matrix $\mathbf{A}$ has NT rows and NT columns.

Forms of Normalization. The computations may be implemented with two different normalization schemes because the EOF representation decomposes the space-time series, using separation of variables, into a sum of products of temporal amplitudes, $A_i(t)$, modulating spatial patterns of variance, $F_i(x)$. Note that, in physics, separation of variables occurs widely because of the form of the underlying differential equations. Here, the motivation is to provide a compact statistical representation of the data in which spatial patterns in variance can be distinguished from temporal patterns. Either the $A_i(t)$ or the $F_i(x)$ can retain the same physical units as the original data. Historically, a normalization scheme (method 2) has been used in which the temporal components of the decomposition retained the same units of the data. More recent studies in oceanography and atmospheric science (e.g., Barnett and Patzert, 1980) prefer to use EOFs with the same units as the original data for easier interpretation of the spatial patterns of variance resulting from the EOF analysis (method 1). Both methods of normalization are equivalent and both are given here for completeness. (The summation notation is used for consistency with the vast majority of published studies.)

In the first method

$$F_i(x) \cdot F_j(x) = \sum_{k=1}^{NX} F_i(x_k) F_j(x_k) = \lambda_i \delta_{ij}, \tag{3.2}$$

where F_i and F_j are columns of the $\mathbf{F}$ matrix and δ_{ij} is the Kronecker delta function. The coefficient λ_i is an eigenvalue of the covariance matrix $\mathbf{C} = \frac{\mathbf{D} \cdot \mathbf{D}^{\mathrm{T}}}{NT}$, where $\mathbf{D}^{\mathrm{T}}$ is the transpose of $\mathbf{D}$. The rows of the time series

matrix are orthonormal, i.e.,

$$\frac{A_i(t) \cdot A_j(t)}{NT} = \frac{1}{NT} \sum_{k=1}^{NT} A_i(t_k) A_j(t_k) = \delta_{ij} , \tag{3.3}$$

where A_i and A_j are rows of $\mathbf{A}$. The EOFs then give the variance in the same units as the data.

The second normalization method forces the EOFs to be orthonormal, i.e.,

$$F_i(x) \cdot F_j(x) = \delta_{ij} . \tag{3.4}$$

Hence,

$$\mathbf{F}^{\mathrm{T}} \cdot \mathbf{F} = \mathbf{I} , \tag{3.5}$$

where $\mathbf{I}$ is the identity matrix. The rows of the time series matrix are orthogonal:

$$\frac{A_i(t) \cdot A_j(t)}{NT} = \lambda_i \delta_{ij} . \tag{3.6}$$

With this normalization, the EOFs no longer have the same units as the data.

Theoretical Basis. The objective of EOF analysis is to represent a given matrix of data $\mathbf{D}$ by the product $\mathbf{F} \cdot \mathbf{A}$. In the discussion that follows, the second method of normalization is used, and thus the matrix $\mathbf{F}$ satisfies equation (3.5).

The equations governing EOF theory can be derived from the eigenequation of the covariance matrix,

$$\mathbf{C} \cdot \mathbf{F} = \mathbf{F} \cdot \mathbf{\Lambda} , \tag{3.7}$$

where $\mathbf{C}$ is the $NX \times NX$ covariance matrix $= \frac{\mathbf{D} \cdot \mathbf{D}^{\mathrm{T}}}{NT}$, $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues, and $\mathbf{F}$ is the matrix of EOFs. Because $\mathbf{C}$ is a real symmetric positive definite matrix, the elements of $\mathbf{\Lambda}$ are positive. By definition of the covariance matrix $\mathbf{C}$, (3.7) can be rewritten as

$$\left\{ \frac{\mathbf{D} \cdot \mathbf{D}^{\mathrm{T}}}{NT} \right\} \cdot \mathbf{F} = \mathbf{F} \cdot \mathbf{\Lambda} . \tag{3.8}$$

The eigenvectors in the matrix $\mathbf{F}$ are now used to define the matrix $\mathbf{A} = \mathbf{F}^{\mathrm{T}} \cdot \mathbf{D}$. This is the principal component transform. Thus, $\mathbf{F} \cdot \mathbf{F}^{\mathrm{T}} = \mathbf{I}$ and necessarily

$$\mathbf{D} = \mathbf{F} \cdot \mathbf{A} . \tag{3.9}$$

At this point the number of EOFs is equal to NT. The purpose of EOF analysis, however, is to produce a representation of the data that is more compact than the original data set. For this goal to be met, it is necessary for most of the variance (65 to 80%) in the original data set to be contained in the first few EOFs. (By convention, the eigenvalues are ordered with the largest being first.) If this is true, then $\mathbf{F}$ may be truncated to produce a new matrix $\hat{\mathbf{F}}$ having just $\widehat{NT}$ columns, where $\widehat{NT} < NT$. Then $\hat{\mathbf{D}}$ is the approximation to $\mathbf{D}$ computed from the EOF decomposition:

$$\mathbf{D} \approx \hat{\mathbf{D}} = \hat{\mathbf{F}} \cdot \mathbf{A} . \tag{3.10}$$

The decomposition is the most efficient representation of $\mathbf{D}$ with regard to a mean square error criterion (Davis, 1976).

Finally, it should be noted that if $\mathbf{D}$ contains a noninteger number of cycles of a sinusoidally varying variance mode, then the variance represented by the associated eigenvalue may not agree with the actual variance. For example, the variance of a sine wave differs for one-half and one full cycle, but the covariance matrix $\mathbf{C} = \frac{\mathbf{D} \cdot \mathbf{D}^{\mathrm{T}}}{NT}$ is the same. Hence, the EOF will return the variance for an entire cycle when the data represent one-half cycle.

A Computational Method. The empirical orthogonal functions can be computed from a singular value decomposition (SVD) of the data $\mathbf{D}$. When $NX \geq NT$, then (Press *et al.*, 1986)

$$\mathbf{D} = \mathbf{U} \cdot \mathbf{W} \cdot \mathbf{V}^{\mathrm{T}} , \tag{3.11}$$

where $\mathbf{U}$ is an $NX \times NX$ orthogonal matrix with only NT linearly independent columns, $\mathbf{W}$ is an $NX \times NT$ matrix with an $NT \times NT$ diagonal upper portion having positive or zero elements, and $\mathbf{V}$ is an $NT \times NT$ orthogonal matrix. Note that the diagonal elements in the $NT \times NT$ upper portion of $\mathbf{W}$ contain the singular values of $\mathbf{D}$. Call this portion of $\mathbf{W}$ the matrix $\mathbf{W}'$. The $\mathbf{U}$ and $\mathbf{V}$ matrices are orthogonal in the sense that their columns are orthonormal, that is,

$$\begin{aligned} \mathbf{U}^{\mathrm{T}} \cdot \mathbf{U} &= \mathbf{I} \\ \mathbf{V}^{\mathrm{T}} \cdot \mathbf{V} &= \mathbf{I} . \end{aligned} \tag{3.12}$$

Using this decomposition, the covariance eigen-equation can be written as

$$\mathbf{C} \cdot \mathbf{U} = \left\{ \frac{\mathbf{D} \cdot \mathbf{D}^{\mathrm{T}}}{NT} \right\} \cdot \mathbf{U} = \frac{1}{NT} \left\{ \mathbf{U} \cdot \mathbf{W} \cdot \mathbf{V}^{\mathrm{T}} \right\} \cdot \left\{ \mathbf{V} \cdot \mathbf{W}^{\mathrm{T}} \cdot \mathbf{U}^{\mathrm{T}} \right\} \cdot \mathbf{U} , \tag{3.13}$$

which simplifies to

$$\mathbf{C} \cdot \mathbf{U} = \frac{1}{NT} \mathbf{U} \cdot \mathbf{W}^2 . \tag{3.14}$$

This is equivalent to the eigen-equation (3.7) with $\mathbf{F} \equiv$ the highest order columns of $\mathbf{U}$ and $\mathbf{\Lambda} \equiv \mathbf{W}^2/NT$, using the second method of normalization. Hence, the diagonal elements W_i of $\mathbf{W}'$ relate to the eigenvalues λ_i of the covariance matrix $\mathbf{C}$ via the equation

$$\frac{W_i'^2}{NT} = \lambda_i . \tag{3.15}$$

Thus, the SVD returns the EOFs as the highest order columns of $\mathbf{U}$, and $\mathbf{A} = \mathbf{W}' \cdot \mathbf{V}^{\mathrm{T}}$.

Generalizations to Two Dimensions. The EOF decomposition is not confined to space-time series where space is one-dimensional. If the positions are actually (x, y) coordinates, such as latitudes and longitudes or lines and samples from images, then if each time step covers the same locations, $\mathbf{D}(x, y, t)$ can be reduced to $\mathbf{D}(x', t)$. The sequence of positions x' is constructed by setting $x'_1 = (x_1, y_1)$, $x'_2 = (x_2, y_1)$, ..., $x'_{NX} = (x_{NX}, y_1)$, ..., $x'_{NX \cdot NY} = (x_{NX}, y_{NY})$. If one considers x and y to be rows and columns of an individual time step matrix, then x' is equivalent to concatenating the rows of $\mathbf{D}$ together. Whether one concatenates the rows or the columns is immaterial. (Note, EOF analysis with two-dimensional images is actually three dimensional: x, y, and t. To render the problem tractable, the spatial dimensions must be concatenated so that one can work with a two-dimensional space-time matrix.) Then the individual EOFs F_i are also vectors in x' and, in order to map the patterns of variation associated with the EOFs, this vector must now be parsed back into its rows and columns.

Example. The purpose of this example is to show that EOF analysis can be useful for determining the dominant patterns of spatial variance in a sequence of images. (The normalization scheme used in this example is that of method 1.) For this purpose, a sequence of eight images was constructed by superimposing an image of an inclined plane with that of a disk that is out of phase with the inclined plane. Note also that the range of data in the image of the inclined plane is about twice that of the image of the circle. One-half cycle of the image sequence so constructed is shown in Plate 3.2a. The EOF decomposition of this sequence (Plate 3.2b) identifies

two dominant patterns of spatial variance in the test sequence. EOF #1 is an inclined plane that accounts for 67.7% of the total variance in the sequence. EOF #2 is a circle that accounts for 32.2% of the total variance. The corresponding time series for EOF #1 and EOF #2 (Plate 3.2b) correctly establish the phase relationship between the two patterns.

Additional Considerations. In remote-sensing applications, multiple observations are generally spread over different times. Moreover, data are generally not evenly distributed either in time (e.g., due to variations in orbit) or in space (e.g., cloud cover may obscure some pixels in the scene). These circumstances can bias the results of an EOF analysis because the EOF analysis is predicated on evenly distributed data in both space and time. If the NT images in the data set D (equation (3.1)) are not equispaced in time, the resultant EOF decomposition may be biased toward specific time periods. A practical way to minimize such temporal bias is to construct a weighting scheme for the images in the data set. Generally, temporal weights are constructed in such a way as to preserve the total mean and total variance of the data while simultaneously minimizing the temporal bias (e.g., Kelly, 1985). Spatial data (i.e., one or more of the NX pixels in a given image) are often replaced by compositing data from other images that surround the given image in time. Care must be taken to ensure that the composite time scale is very much less than the time step between images in the sequence undergoing EOF analysis. Other interpolation schemes (e.g., kriging) can also be used to minimize spatial and temporal bias in EOF analyses resulting from imperfectly sampled data.

There are three ways to compute EOFs: (1) large covariance matrix approach, (2) small covariance matrix approach, and (3) singular value decomposition. The first two methods involve a direct solution of the eigenvalue equation of the covariance matrix of the data set D. There are two ways to do this because the data in the image sequence can be stored in the two-dimensional data array in two different ways. If the data are stored such that there are NX rows and NT columns, the covariance matrix will be $NX \times NX$. If, however, the data are stored such that there are NT rows and NX columns, then the covariance matrix will be $NT \times NT$. For most remote sensing applications, the number of spatial points NX in a given image in the sequence is usually very much greater than the number of images, NT, in the sequence. Thus the method of data storage giving rise to an $NX \times NX$ covariance matrix is often called the large covariance ma-

trix approach and generally cannot be used in remote sensing applications. Likewise, the method of storage resulting in an $NT \times NT$ covariance matrix is generally called the small covariance matrix approach. It has been used often in oceanographic and atmospheric applications (e.g., Preisendorfer, 1988). Solution of the eigenvalue equation of the covariance matrix by singular value decomposition was discussed earlier in this chapter. A detailed discussion of the various methods of solution of the eigenvalue equation of the covariance matrix, in the context of EOF analysis, is given by Gallaudet and Simpson (1991b).

3.2.2 Velocity Estimates from Image Sequences

General Comments

All methods of motion estimation based on image sequence analysis depend in some way on the detection of image brightness gradients. These gradients are defined as normal in direction to contours of constant brightness (or sea surface temperature [SST] for the case of the Advanced Very High Resolution Radiometer [AVHRR] flying on the NOAA series of operational satellites). The total velocity $\mathbf{v}$ at any point on a contour can be written as

$$\mathbf{v} = \nu\hat{\mathbf{n}} + \tau\hat{\mathbf{t}}\,, \tag{3.16}$$

where ν and τ are the magnitudes of the normal and tangential velocity components, respectively, and $\hat{\mathbf{n}}$ and $\hat{\mathbf{t}}$ are the unit vectors in the directions normal and tangential to the contour, respectively. The total velocity vector can also be decomposed into ordinary Cartesian coordinates

$$\mathbf{v} = u\hat{\mathbf{i}} + v\hat{\mathbf{j}}\,, \tag{3.17}$$

where u and v are the magnitudes of the x and y velocity components, respectively, and $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are the unit vectors in the x and y directions, respectively. Below, objective methods for computing the total, normal, and tangential components of near-surface oceanic flow from sequences of AVHRR data are presented. The formulation follows closely that of Wahl and Simpson (1990a,b). Note that atmospheric motions can be computed using these same methods from sequences of either AVHRR or Geostationary Operational Environmental Satellite (GOES) data.

Total Flow Field From Pattern Matching

Consider a pair of real, discrete, two-dimensional random functions, $s(m,n)$ and $p(m,n)$. The expectation value of these random functions can be approximated as the normalized sum over the tile occupied by the random function in the image. Thus,

$$\mathbf{E}[(s(m,n)] = \frac{1}{NL} \times \frac{1}{NS} \sum_{m=1}^{NL} \sum_{n=1}^{NS} s(m,n)\,, \tag{3.18}$$

where NL and NS are the number of lines and the number of samples in the tile. Assuming stationarity of their first two cross moments, the auto-covariance and cross-covariance of the two functions are given by

$$\begin{aligned} C_{ss}(m_0,n_0) &= \mathbf{E}[(s(m,n)-\eta_s)(s(m+m_0,n+n_0)-\eta_s)] & (3.19)\\ C_{sp}(m_0,n_0) &= \mathbf{E}[(s(m,n)-\eta_s)(p(m+m_0,n+n_0)-\eta_p)]\,, & (3.20) \end{aligned}$$

respectively, where $\mathbf{E}[\cdot]$ is the expected value, (m_0,n_0) is the spatial lag between the functions, and η_s and η_p are given by the function means

$$\begin{aligned} \eta_s &= \mathbf{E}[s(m,n)] & (3.21)\\ \eta_p &= \mathbf{E}[p(m,n)]\,. & (3.22) \end{aligned}$$

It can be seen that the function variances σ_s^2 and σ_p^2 are the zero lag auto-covariances. The correlation coefficient is defined as

$$r_{sp}(m_0,n_0) = \frac{C_{sp}(m_0,n_0)}{\sigma_s \sigma_p}\,, \tag{3.23}$$

such that

$$\begin{aligned} |r_{sp}(m_0,n_0)| &\leq 1 & (3.24)\\ r_{ss}(0,0) &= 1\,. & (3.25) \end{aligned}$$

If the second signal is an exact spatially lagged version of the first signal,

$$s(m,n) = p(m+m_0,n+n_0)\,, \tag{3.26}$$

then equations (3.20) and (3.25) require that the correlation coefficient achieve an absolute maximum value at this lag, or $r_{sp}(m_0,n_0) = 1$. For any physical signals, the maximum correlation will be a value less than 1 since the second signal is not necessarily an exact lagged version of the first.

A detailed discussion of the two-dimensional cross-correlation function is given by Dudgeon and Mersereau (1984).

Now consider two consecutive satellite SST images mapped to the same spatial grid. These images can be thought of as two-dimensional discrete functions. Given an image subsection $s(m,n)$ from the second image, the problem is to determine if it contains a region similar to a smaller subsection $p(i,j)$ from the first image. Let $p(i,j)$ be called the pattern tile, and let $s(m,n)$ be called the search tile. The pattern tile is a section of the first image that occupies the same spatial coordinates as the center region of the search tile in the second image. Then, the correlation matrix between the pattern and search tiles is given by

$$r_{sp}(k,l) = \frac{\sum_i \sum_j [s(i+k,j+l) - \eta_s(k,l)]\,[p(i,j) - \eta_p]}{\left\{\sum_i \sum_j [s(i+k,j+l) - \eta_s(k,l)]^2 \sum_i \sum_j [p(i,j) - \eta_p]^2\right\}^{1/2}}, \tag{3.27}$$

where $\eta_s(k,l)$ is the average value of $s(m,n)$ in the subregion coincident with $p(i,j)$, and the summations are over the coordinates common to both s and p. The value of η_p is computed once outside the summations and is given by (3.22), where the pattern tile $p(m,n)$ replaces the search tile $s(m,n)$ in (3.18). The ranges of k and l correspond to the regions of correlation in which $p(i,j)$ is completely contained in $s(m,n)$.

The pattern matching method determines the spatial lag between the pattern tile from the first image and the search tile in the second image by finding the location of their maximum correlation. With this spatial lag and the time between the images, the average velocity of the features in the pattern tile can be computed. The most basic assumption of the method is that the spatial displacement of thermal gradient features can be tracked as if the shape were time invariant. This assumption would be rigorously true if the correlation (3.27) were equal to 1 for each pattern-search tile pair. Unfortunately, this condition is never met. Hence, it becomes necessary to determine an acceptable minimum correlation or correlation threshold (see Wahl and Simpson, 1990a). Velocities obtained from the pattern matching technique and proper choice of correlation threshold show good correspondence with observations.

The pattern matching method may also be implemented in the wave number domain. This implementation uses the discrete Fourier transform (DFT) property

$$s(m,n) * p(m,n) \leftrightarrow S'(k_x,k_y)P(k_x,k_y), \tag{3.28}$$

where $*$ represents the correlation operation, $S'(k_x, k_y)$ is the conjugate of the DFT of $s(m,n)$, and $P(k_x, k_y)$ is the DFT of $p(m,n)$. This is the case because the correlation in the space-time domain and the spectrum in the frequency-wavenumber domain are Fourier transform pairs (Dudgeon and Mersereau, 1984). Thus, the correlation between two discrete signals can be computed by taking the product of the DFT of a zero-padded pattern tile with the conjugate of the DFT of the search tile and taking the inverse DFT of this product. It may seem that computing the DFT of the discrete functions via the fast Fourier transform (FFT) is a more efficient approach. In most applications, however, the pattern tile usually occupies a much smaller area than the search tile. If the number of non-zero terms in the pattern tile is less than 132, it is more efficient to implement (3.27) than to use the FFT algorithm to compute the correlation function (Campbell, 1969). (For odd size tiles the Winograd implementation of the FFT is recommended.)

The method used here is similar to other template matching schemes, such as area correlation and matched filtering (Jain, 1989). Matched filtering involves the construction of a linear filter that maximizes the output signal-to-noise ratio. Using the matched filtering technique, the area surrounding the pattern is assumed to be colored noise. If the power spectral density of the noise is known, the signal-to-noise ratio can be maximized by passing the signals through a high-pass filter before performing the area correlation.

Minimum distortion methods have been used to do interframe registration in video camera systems (Jain and Jain, 1981). Simpson and Bloom (1990) have applied this method to the computation of near-surface velocity from sequences of images and have shown that the distortion is simply related to the correlation. In effect, maximizing the correlation is equivalent to minimizing the distortion assuming that the variance and standard deviations of the different search tiles remain the same. Both methods yield the same velocity fields (Simpson and Bloom, 1990) for a given image sequence. However, the minimum distortion method executes faster because it does not require a standard deviation computation.

The basic assumption of these methods is shape invariance of the pattern under translation. Rotational motion of the pattern, however, also can occur. Pattern matching techniques that detect rotational motion and/or combined translational-rotational motion of the pattern have been developed (e.g., Jain, 1989). These methods are computationally very expensive but will perform well in the presence of curvature in the motion field of the pattern. Only translational techniques were used herein because they are sufficient to determine the general flow pattern in the image sequence under consideration.

The Normal Component of Flow

Marr and Ullman (MU) Method. Early visual primitives can provide clues to establish the motion of elements in a visual field (Marr and Ullman, 1981). The simplest such primitives are the image raw intensity values, but these provide no information about the shapes of objects. The next higher-order primitives are the zero crossing segments produced by the convolution of an image with the Laplacian of Gaussian (LOG) operator, $\nabla^2 G$ (Marr and Hildreth, 1980). The LOG operator is defined by

$$\nabla^2 G = k \left[1 - \frac{(x^2 + y^2)}{2\sigma^2}\right] \exp\left[\frac{-(x^2 + y^2)}{2\sigma^2}\right], \tag{3.29}$$

where x and y are the number of rows and columns from the function center, σ is the Gaussian width parameter, and k is a normalization constant. The parameter σ determines the spatial scales of intensity changes detectable by the $\nabla^2 G$ operator. The $\nabla^2 G$ operator is the optimal smoothing bandpass filter in the sense that it minimizes the product of bandwidth and spatial localization (e.g., Marr and Hildreth, 1980). If I is the demeaned image function, then locations of zero crossings of the convolution of the $\nabla^2 G$ operator with I will correspond to locations of intensity changes (i.e., gradients). Let this convolution be denoted $\nabla^2 G * I = I'$, where I' is the output of the convolution. Note the units of I and I' are the same (°C for AVHRR data). Note also that the LOG operator (a commonly used edge detector) produces a set of zero crossings in the image. The locations of the zero crossings (i.e., the edges) are then determined by a zero crossing operator that detects the positions of the edges by finding the locations of changes in sign in the LOG-convolved image.

The idea of directionally sensitive units, which can establish the direction of movement of an edge detected by the $\nabla^2 G$ operator, was introduced by Marr and Ullman (1981, hereafter referred to as MU) to determine the motion of visual elements. Given an approximation to the time derivative of I' and the spatial rate of change of I', the direction of motion of a zero crossing in either the line or sample direction can be determined. The locations of zero crossings in the first convolved image of the sequence are determined by a zero crossing operator. Then, at each zero crossing pixel, the normal component of flow can be estimated using the equation

$$v = \frac{-I'_t}{|\vec{\nabla} I'|}, \tag{3.30}$$

where the subscript represents differentiation with respect to time, and the unit normal vector $\hat{\mathbf{n}}$ is in the direction of the gradient $\vec{\nabla} I'$. Equation (3.30) is the conservation of heat equation ($\frac{DT}{Dt} = 0$) for the convolved image I'. If the magnitude of the gradient is negative and I'_t is positive, then motion of the edge is in the positive x direction. If I'_t is negative, motion is in the negative x direction. The opposite is true if the magnitude of the gradient is positive. Thus, the negative sign in (3.30) gives the correct direction of the normal component of flow. It can be seen that a directionally sensitive unit is represented by a transition in the sign of I'_t combined with the sign of the gradient.

Spatial-Scale Considerations. Methods used to compute the normal component of flow employ small neighborhoods, typically 4×4 pixels or smaller in size. The total velocity is computed over a much larger neighborhood. Typically, pattern tile sizes vary between 16×16 and 32×32 pixels. Thus, estimates of the total velocity via pattern recognition represent the mean motion of the centroid of the pattern as measured by the displacement of the two-dimensional cross-correlation maximum of the pattern. The normal velocities, however, provide local estimates of the displacement of small spatial-scale gradient, typically computed over 4×4 tiles. This basic difference in spatial scales further constrains the computation of the tangential flow.

Other Representations

Optical Flow (OF) Method. Optical flow (hereafter referred to as OF) is an estimate of the motion of solid bodies based on a first-order variation of brightness patterns in an image (e.g., Horn and Schunck, 1981). This method of computing the velocity field from a sequence of images is based on the solution of two constraint equations. The first constraint equation relates the velocity in an image to the image brightness (or temperature) pattern and is called the "motion constraint equation":

$$\frac{DT}{Dt} = T_t + \mathbf{v} \cdot \nabla T = 0, \tag{3.31}$$

where $\frac{D}{Dt}$ is the material derivative operator, T_t is the partial derivative of brightness (or SST) with respect to time, and $\mathbf{v}$ is the total velocity vector. This equation can only estimate the normal component of velocity because the tangential component solution is an annihilator (i.e., $\tau(\hat{\mathbf{t}} \cdot \vec{\nabla} T)$).

The form and physical interpretation of the second constraint used in OF methods is one that has been widely debated. In general, it seems that the second constraint imposed is not based on any physical characteristics of the flow field, but, rather, it is a mathematical constraint imposed to produce a unique solution. Horn and Schunck (1981) used a constraint based on the smooth variation of the flow field to derive an iterative scheme for computing what they referred to as the "total flow." Various other schemes have been introduced to estimate the total flow (see Aggarwal and Nandhakumar (1988) for a review of OF methods). It is noted (e.g., Horn and Schunck, 1981; Hildreth, 1983; Verri and Poggio, 1989) that the estimate of the total flow field may be very far from the actual velocity field, depending on various factors influencing the time series of images.

Given the material derivative constraint (3.31), one assumes that the flow is continuous and varies smoothly over small spatial scales. One way to define the measure of smoothness is to examine the squares of the magnitudes of the spatial rate of change of the OF velocity. This can be written as a departure from smoothness error

$$E_c^2 = u_x^2 + u_y^2 + v_x^2 + v_y^2 , \tag{3.32}$$

where (u, v) is the local total velocity vector, and the subscripts represent differentiation with respect to the spatial coordinates (x, y). There will also be errors in the estimation of the partial derivatives of brightness because noise is amplified by differentiation. Thus the equality of (3.31) will not be exact. Define this error term as

$$E_b = T_t + uT_x + vT_y , \tag{3.33}$$

where subscripts indicate differentiation with respect to either a spatial (x, y) or temporal (t) coordinate. The objective function to be minimized can be written as the integral

$$J = \iint (E_b^2 + \alpha^2 E_c^2)\, dx\, dy , \tag{3.34}$$

where E_b is the error in computing the material derivative, E_c is the measure of smoothness, and α^2 is a weighting parameter. The calculus of variations can be used to minimize this integral. Then the variation equations can be rewritten as spatial iterative equations

$$u^{k+1} = \bar{u}^k - \frac{T_x(T_x\bar{u}^k + T_y\bar{v}^k + T_t)}{\alpha^2 + T_x^2 + T_y^2} ,$$

$$v^{k+1} = \bar{v}^k - \frac{T_y(T_x\bar{u}^k + T_y\bar{v}^k + T_t)}{\alpha^2 + T_x^2 + T_y^2}, \tag{3.35}$$

where the superscript k is an iteration index. The first step is to compute the brightness derivatives at all points in the image using centered finite differences. Then, starting with an initial estimate of zero velocity, the method spatially iterates on the velocity values until velocity residual between estimates is small. The resulting velocity field is then used as the new initial velocity estimate for the next time step if more than one time step is available. It is interesting to note that updated velocity values in the iteration equations (3.35) do not rely solely on the previous values at a given point, but rather on the local averages of velocity. Note that the local averages of velocities typically are computed over small spatial domains and are computationally efficient. The parameter α^2 is seen to be important in regions of small gradient. If the gradients are small relative to α^2 then α^2 will dominate any perturbations in the estimation of the derivatives at this point.

Minimum Norm (MN) Solution for Normal Flow. Equation (3.31) gives one equation for the two unknowns (u, v) of the total velocity. This underconstrained system has fewer equations than unknowns and thus has an infinite number of solutions. One way to solve such underconstrained systems is to find the solution with the minimum vector length, or norm (Luenberger, 1969). Equation (3.31) was solved for the normal component of velocity using the singular value decomposition to obtain the solution of minimum norm (hereafter referred to as MN) at every point in the given image subsection (Wahl and Simpson, 1990b). It can again be seen that the solution yields only the normal component of flow because the tangential component is an annihilator of (3.31). The MN solution of normal velocity was done on a point-by-point basis on raw temperature data using a centered finite difference for the derivatives of temperature.

Equivalence of Minimum Norm (MN) and Optical Flow (OF) for $\alpha^2 = 0$. Both the MN and OF estimates of the normal component of flow are based on the same motion constraint equation (i.e., equation (3.31)). The first error equation used in the OF method (equation (3.32)) is a measure of the smoothness of the flow field. The second error equation (3.33) is the error in the estimation of the motion constraint equation. The objective function minimized in the OF method is the integral of the sum of these

errors (equation (3.34)) where the parameter α^2 weights the smoothness error. If α^2 is set to zero then the OF method essentially minimizes the error in the estimation of the motion constraint (3.31). When $\alpha^2 = 0$ the solution for (3.34) using the OF method converges to the solution of (3.31) using the MN method.

The Tangential Component of Velocity

General Considerations. In the previous section, only the component of flow normal to isobrightness contours was discussed because the tangential component of flow cannot be calculated directly. The difficulty arises from what is known as the aperture problem and manifests itself in different ways for each motion estimation algorithm. In the MU case, the problem occurs when the motion of an oriented edge is detected by a direction-sensitive unit that is small compared to the moving edge. Then the only information that can be extracted is the component of motion perpendicular to the local orientation of the edge. Hence the component of motion oriented along the edge is invisible. In both the OF and MN methods for estimating the normal component of velocity, the aperture problem manifests itself as the annihilation of the tangential component in the basic constraint equation (3.31). An alternative interpretation of the aperture problem occurs if a point of brightness along an isobrightness contour moves along that contour from time t_1 to time t_2; this motion cannot be detected. These considerations show that a direct method for computing the tangential component of flow is not possible. Vector decomposition of the known total velocity field, given a known normal component of flow, however, can yield an estimate of the tangential component of flow. These considerations are consistent with a proof given by Verri and Poggio (1989).

An Indirect Solution. Given the total flow field computed on a rectangular grid from the pattern matching method mentioned previously, one can take the normal component of flow and perform a vector subtraction to obtain the tangential component. This decomposition was performed using the three normal component representations discussed above. The MU normal component of velocity was chosen for this purpose because Wahl and Simpson (1990b) have shown that it produces better estimates of the normal component of velocities than either the OF or MN method.

The MU method velocities are computed only at the points of zero crossings of edges. These vectors must be spatially aligned on the same grid as

that of the total velocity prior to the decomposition. To estimate the normal components of velocity at the locations of the total flow, the MU normal components were subsectioned into the same size tiles as the total flow. The normal velocities which fell within the region of the pattern tile area of the first image were then averaged to produce an overall average normal velocity for a given tile. This procedure is consistent with the assumption that the total velocity vector produced by the pattern matching technique represents the average velocity of the feature in a pattern tile. The mean normal component of flow was then decomposed into its Cartesian components and these components were used in the final vector decomposition to compute the tangential component of flow.

Example

A sequence of cloud-free AVHRR images for a region off the central California coast was co-registered to a latitude-longitude grid and calibrated to SST (Plate 3.3). The image is stored as a matrix where, by tradition in image analysis, the row index is referred to as a line and the column index is referred to as a sample. Co-registration is the process of mapping images observed in the line and sample domain to the same latitude and longitude domain using an appropriate map projection (e.g., Brush, 1985; Jezching and Jain, 1989). Calibration is the process of converting the raw measured brightness counts in one or more of the images to a geophysical variable (e.g., Kaufman, 1988). This sequence is characterized by a cold-water filament extending southward from the top of the sampled region. Thermal structure edge maps for time step 2 of the image sequence were computed using the LOG operator with a value of $\sigma = 5$ (Plate 3.4a). Motion inferred from these edge maps agrees well with estimates of the total flow field computed using the pattern matching method (Plate 3.4b).

These edge maps were then used to compute the normal component of velocity of the thermal structure over time using (3.30) at the zero-crossing points. A centered finite difference scheme was used to compute the spatial gradient of I', and a single time-centered difference was used to approximate the temporal derivative of I'. It is important to emphasize that the MU method (i.e., equation (3.30)) only yields an estimate of the normal component of velocity near a well-defined edge (Plate 3.4a). The normal component of flow so obtained (Plate 3.4c) accurately represents motion inferred from the edge maps. Note especially the north-south oriented feature in the center right region (see region marked 3 in Plate 3.4c). The tangen-

tial component of flow (Plate 3.4d) again shows good correspondence with motion inferred from the edge maps.

Ideally the estimated normal and tangential components of flow should be orthogonal. The need for spatial averaging of the normal components, however, may introduce directional errors in the approximation of the tangential component. Wahl and Simpson (1990b) have shown that typically the angles between the normal and tangential components of flow are between 75° and 80°. Thus, this method generally will not produce an exact tangential solution. It does, however, produce an approximate tangential solution, which can be useful in many oceanographic applications (e.g., computation of the offshore transport of nutrients associated with coastal upwelling). It should be reemphasized that there is no direct method for computing the tangential component of motion from sequences of image data.

3.2.3 Ice Floe Identification and Principal Curves

Overview of Banfield and Raftery Algorithm

Knowledge of the spatial distribution, size, and shape of ice floes is critical to understanding physical processes in polar regions and the potential role of these processes in studies of global warming. Moreover, in high-latitude zones, shipping, naval operations, fishing, and the successful deployment of offshore structures are all strongly influenced by the distribution of the polar ice pack. Banfield and Raftery (1989) have developed an automated procedure for identifying ice floes in Landsat data. Automated procedures are needed for several operational reasons: (1) to handle the huge volume of data; (2) to eliminate intercalibration problems associated with subjective analyses; and (3) to improve on the poor performance records of human analysts working under the adverse weather conditions often associated with polar operations. The Banfield and Raftery method uses principal curves (Hastie and Stuetzle, 1989), an erosion propagation algorithm, and a method for clustering about principal curves to automatically identify the floes. Only the major elements of the method are reviewed here: the interested reader is referred to Banfield and Raftery (1989) for details of the procedure.

Hastie and Stuetzle (1989) developed the concept of a principal curve. A principal curve can be thought of as a smooth one-dimensional curve that passes through the middle of an m-dimensional data set. It is nonparametric, and its shape is suggested by the data; it thus provides a nonlinear summary

of the data (Banfield and Raftery, 1989). A one-dimensional curve in m-space is an m-vector consisting of m functions of a single variable λ, called coordinate functions. The variable λ parameterizes the curve and provides an ordering along it; λ often is the arc length along the curve. Let $\boldsymbol{\chi} \in \mathbf{R}^m$ be a continuous random vector. Then $\mathbf{f}(\lambda)$ is a principal curve if

$$\mathbf{E}[\boldsymbol{\chi} | \mathbf{f}^{-1}(\boldsymbol{\chi}) = \lambda] = \mathbf{f}(\lambda),$$

where

$$\mathbf{f}^{-1}(\mathbf{x}) = \max_{\lambda} \left\{ \lambda : \|\mathbf{x} - \mathbf{f}(\lambda)\| = \inf_{\mu} \|\mathbf{x} - \mathbf{f}(\mu)\| \right\}.$$

Given the distribution of $\boldsymbol{\chi}$, Hastie and Stuetzle (1989) proposed the following algorithm for finding f:

$$\mathbf{f}_{i+1}(\lambda) = \mathbf{E}[\mathbf{x} | \mathbf{f}_i^{-1}(\boldsymbol{\chi}) = \lambda],$$

where $\mathbf{f}_i$ is the i^{th} iterate. If the distribution of $\boldsymbol{\chi}$ is unknown, then an estimate of $\mathbf{f}$ can be obtained from the data set $\{\mathbf{x}_i\}$ by estimating $\mathbf{E}[\mathbf{x} | \mathbf{f}_i^{-1}(\boldsymbol{\chi}) = \lambda]$. Hastie and Stuetzle (1989) obtain this estimate by scatterplot smoothing.

Banfield and Raftery (1989) noted that scatterplot smoothers generally produce curves that are biased toward the center of curvature. They modified the Hastie and Stuetzle (1989) principal curve estimation algorithm to reduce bias and variance by using projections of the data rather than the data itself to model the principal curves when the distribution is unknown.

Next, Banfield and Raftery (1989) used an erosion-propagation (EP) algorithm to select potential edge pixels and to provide an initial grouping of the edge pixels into floe outlines. The EP algorithm operates on a binary image. Hence, the Landsat data must be binarized by thresholding prior to EP analysis. Banfield and Raftery (1989) justified this procedure by noting that the marginal distribution of pixel intensities in the high-resolution polar Landsat data is highly bimodal. They further noted that the final result is relatively insensitive to the precise value of the threshold chosen. The erosion part of the EP algorithm identifies potential edge elements by using standard concepts from mathematical morphology (Serra, 1982), while the propagation part keeps track of the floe to which an edge pixel belongs by locally propagating the information about the edge elements into the interior of the floe as it is eroded. The algorithm is applied iteratively to the binarized image.

Banfield and Raftery (1989) noted that the EP algorithm tends to subdivide floes. Therefore, they developed a method, based on clustering about the closed principal curves, for determining which of the floes identified by the EP algorithm should be merged. This final component of the overall procedure to identify ice floes in polar Landsat data is hierarchical and agglomerative.

Example

Shown in Plate 3.5a is a polar Landsat image. This image is 200×200 pixels, where each pixel is an 80-m square. The entire image represents a 15×15 km area. Ice floes appear as the gray features against the darker background. Ice floe outlines for this image, obtained using the Banfield and Raftery (1989) algorithm, are shown in Plate 3.5b. The algorithm accurately identifies the distribution, size, and shape of the floes on space scales of 200–300 m and larger.

3.3 Storage and Image Representation

A typical AVHRR image consists of five channels of matrix data. The matrix size typically is in the range of 4,000 lines by 2,000 samples. Generally, the 10-bit sample is stored as a 16-bit integer with the upper 6 bits of each sample filled with zeros. All of these factors work out to about 80 to 100 megabytes (Mb) of archived data for a typical single scene. Use of satellite data sets in the analysis of problems related to global-scale climate processes may require the analysis of literally thousands of such images. Thus, there is a need to have efficient storage and economical data structures for proper and efficient representation of the image data.

3.3.1 Storage Considerations

The primary archive of satellite data is generally the raw digitized telemetry stream directly received from the satellite. For multispectral images, the data are usually band interleaved rather than band sequential, usually contain embedded calibration information, and often include other data needed for proper Earth location of the scene. This data structure is inherently one-dimensional and has little resemblance to the two-dimensional image data structures normally associated with satellite images. (Note, however, that

the more familiar two-dimensional satellite images are subsequently constructed from this telemetry stream using some set of mathematical transformations.) These data generally are called "level 1" data and are mandated for primary archives because higher-level data (e.g., the two-dimensional images) often are produced by irreversible transformations. Moreover, level 1 data usually consist of 10-bit (e.g., AVHRR, CZCS) or less (e.g., DMSP) data strings packed into zero-filled 16-bit integer format for convenience.

Data compression techniques and optical disc storage technology clearly are required if the data storage issue is to be adequately addressed. (Data compression is the process of reducing the number of bits required to store a given amount of information without loss of information.) For example, tests with a recent compression algorithm using Lempel-Ziv-Welch (LZW) coding (Welch, 1984) conservatively show that the average 80-Mb scene can be compressed to 32 Mb. This represents a 60% reduction in size from the original data set. Some scenes may be compressed by as much as 75% with LZW coding.

The LZW compression algorithm has two main competitors currently in common use: Huffman coding and run length coding (RLC). The Huffman coding algorithm is not well suited to satellite data: preliminary tests indicate that only a 14–20% reduction in size is achievable (Jain, 1989). Furthermore, it is much slower than the LZW algorithm. The RLC algorithm was originally designed to vectorize bit maps. It is designed to work on strings of bits which are *all* 1s or 0s (Jain, 1989). This makes it impractical for satellite data, which tends to vary too much (i.e., the strings of uniform 1s or 0s are too short). While the upper 6 bits of each 16-bit sample can be coded efficiently with the RLC algorithm, the remaining 10 bits pose a serious problem for RLC methods. Preliminary tests show that the estimated size reduction obtained from RLC algorithms will be in the 10–25% range. The speed of the RLC is comparable to that of LZW. These considerations show that the LZW compression algorithm is best suited to the proposed task.

Preliminary tests also indicate that decompressing a typical satellite pass so that it can be used for analysis can take as much as 20 minutes per pass. This time factor depends on the speed of the disc and processing unit. No compression algorithm can get around this problem; increased access time is the trade-off for decreased space usage.

Data compression techniques are also needed for the two-dimensional higher level satellite images. Unfortunately, most two-dimensional algorithms either achieve speed by creating distortion in the data or achieve lack

of distortion by requiring excessively long execution times. Considerable research is needed to develop efficient two-dimensional compression/decompression algorithms which do not distort the data.

3.3.2 Image Representation

The data structure selected to represent the spatial data in an image will have a critical effect on the implementation and final performance of the analysis algorithm. The quadtree and octree are hierarchical data structures often used to represent spatial data. The term *quadtree* is used to describe a class of hierarchical data structures whose common property is that they are based on the principle of recursive decomposition of space (Samet, 1989). They can be differentiated on the following bases: (1) the principle used to determine the decomposition process, (2) the resolution (variable or constant), and (3) the type of data they are used to represent. The prime motivation for the development of the quadtree is the need to reduce the amount of space necessary to store data through the use of aggregation of homogeneous blocks (Samet, 1989). An important by-product of this aggregation is the reduction in execution time of an analysis process. Quadtrees have proved to be useful data structures for dithering algorithms, computing geometric properties of images, implementation of linear image transformations, development of hierarchical hidden-surface algorithms, and ray tracing. The quadtree is only one of several digital data structures useful in spatial statistics and digital image analysis.

The constraints on and the need for the large-scale use of remotely sensed images in studies of global change is clear. Research is required in areas of both data storage and image representation if optimal use of remotely sensed data by the earth sciences community is to be achieved.

3.4 Special Considerations

Remote sensing of the environment with earth-observing satellites poses some additional considerations beyond those normally encountered in laboratory-based applications of digital image analysis. In the laboratory, both illumination and viewing geometry can be controlled. Moreover, the imaging detector is close to the object being detected, and any interfering influence between imaging detector and object can be minimized. Finally, one generally has a good notion of what constitutes the detected object. In contrast, earth observing satellites typically fly 800–900 km above the surface of the

earth. Viewing geometry and illumination are not controlled and can vary greatly from orbit to orbit. The 800–900 km layer of atmosphere between target and detector acts as a filter that varies spatially and temporally, often partially corrupting image quality. There is the need to accurately project the image taken 900 km above the Earth onto a flat map projection suitable to the particular application under study. Thus atmospheric correction algorithms (e.g., Curran and Dungan, 1989; Kaufman, 1988; Gratzki and Gerstl, 1989; Simpson and Humphrey, 1990), sensor calibration algorithms (e.g., Gordon *et al.*, 1983; Eckstein and Simpson, 1990a), and earth location algorithms (e.g., Brush, 1985; Goshtasby *et al.*, 1986; Jezching and Jain, 1989) are all pre-processing steps essential prior to meaningful mathematical analysis.

3.5 Summary

Remote sensing provides the only practical way to obtain the large-scale synoptic data sets necessary to address major problems of global significance in earth science that are fundamentally cross-disciplinary in nature. The magnitude of the data set, however, may preclude meaningful understanding unless it can be distilled and synthesized into organized patterns of variance that can be meaningfully related to the underlying governing physics, chemistry, and biology of global-scale cycles and processes. Spatial statistics and mathematical methods of digital image analysis provide mechanisms for such a synthesis. The examples cited herein included principal component analyses, which are useful for image segmentation and for determining spatial patterns of variance in large data sets; edge detection; pattern matching; optical flow methods, which are useful for determining fields of motion from sequences of image data; and principal curves, which are useful for determining the spatial distribution, size, and shape of ice floes observed from spacecraft data. Atmospheric correction algorithms, sensor calibration algorithms, and earth location algorithms generally are required as pre-processes to digital image analysis of remotely-sensed images. Each of these pre-processing areas contains challenging mathematical problems which will have to be solved before earth sciences can benefit from the full potential of remote-sensing technology.

Acknowledgments

This work was sponsored in part by the Marine Life Research Group of the Scripps Institution of Oceanography and by a grant from the Office of Naval Research. AVHRR data (Plates 3.1, 3.3, and 3.4) were provided by the Scripps Satellite Oceanography Center. Landsat data (Plate 3.5) and permission to use material from the Banfield and Raftery ice floe study were generously provided by Professor Raftery of the University of Washington. Dr. Barbara Eckstein produced the sample EOF analysis (Plate 3.2) while working as an ONR-sponsored postdoctoral research assistant in the author's research group at Scripps. Several other present or former members of this group (L. Al-Rawi, D. Atkinson, J. Bloom, T. Gallaudet, C. Humphrey, J. Toman, and D. Wahl) contributed in one way or another to the completion of this chapter. S. McBride typed initial drafts of the manuscript, and F. Crowe and G. Tupper assisted with figure preparation.

Bibliography

[1] Aggarwal, J. K., and N. Nandhakuymar, On the computation of motion from sequences of images—A review, *Proceedings of IEEE* **76** (1988), 917–935.

[2] Anderberg, M. R., *Cluster Analysis for Applications*, Academic Press, New York, 1973.

[3] Banfield, J. D., and A. E. Raftery, *Ice Floe Identification in Satellite Images using Mathematical Morphology and Clustering about Principal Curves*, Technical Report No. 172, Department of Statistics, University of Washington, Seattle, 1989.

[4] Barnett, T. P., and W. C. Patzert, Scales of temperature variability in the tropical Pacific, *J. Phys. Oceanogr.* **10** (1980), 529–540.

[5] Browning, K. A., Use of radar and satellite imagery for the measurement and short-term prediction of rainfall in the United Kingdom, in *Remote Sensing Applications in Meteorology and Climatology*, R. A. Vaughn, ed., NATO ASI Series, vol. 201, 1986.

[6] Brush, R. J. H., A method for real time navigation of AVHRR imagery, *IEEE Trans. Geosci. and Remote Sensing* **23** (1985), 876–887.

[7] Campbell, J. D., Edge Structure and the Representation of Pictures, Ph.D. dissertation, Department of Electrical Engineering, University of Missouri, Columbia, 1969.

[8] Cheevasuvit, F., H. Maitre, and D. Vidal-Madjar, A robust method for picture segmentation based on a split and merge procedure, *Comput. Vision, Graph. Image Proc.* **34** (1986), 268–281.

[9] Curran, P. J., and J. A. Dungan, Estimation of signal-to-noise: A new procedure applied to AVIRIS data, *IEEE Trans. on Geosci. and Remote Sensing* **27** (1989), 620–627.

[10] Davis, R. E., Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean, *J. Phys. Oceanogr.* **6** (1976), 249–266.

[11] Davis, R. E., Techniques for statistical analysis and prediction of geophysical fluid systems, *Geophys. Astrophys. Fluid Dyn.* **8** (1977), 245–277.

[12] Deschamps, P. Y., M. Herman, and D. Tanre, Modeling of the atmospheric effects and its application to remote sensing of ocean color, *Appl. Opt.* **22** (1983), 3751–3778.

[13] Dudgeon, D. E., and R. M. Mersereau, *Multidimensional Digital Signal Processing*, Prentice Hall, Englewood Cliffs, N.J., 1984.

[14] Eckstein, B. A., and J. J. Simpson, Aerosol and Rayleigh radiance contributions to Coastal Zone Color Zone Scanner images, *Int. J. Remote Sensing*, in press (1990a).

[15] Eckstein, B. A., and J. J. Simpson, Cloud screening Coastal Zone Color Zone Scanner images using channel 5, *Int. J. Remote Sensing*, in press (1990b).

[16] EOS, Science and Mission Requirements Working Group Report, Volume I and Appendix, National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, Md., 1984.

[17] Esbensen, K., and P. Geladi, Strategy of multivariate image analysis (MIA), *Chemometrics and Intelligent Laboratory Systems* **7** (1989), 67–86.

[18] Fukada, Y., Spatial clustering procedures for region analysis, *Pattern Recognition* **12** (1980), 395–403.

[19] Gallaudet, T. C., and J. J. Simpson, Automated cloud screening of AVHRR imagery using split-and-merge clustering, *Remote Sensing Environ.*, submitted (1991a).

[20] Gallaudet, T. C., and J. J. Simpson, Seasonal and non-seasonal variability in sea surface temperature off Punta Eugenia, to be submitted to *Remote Sensing Environ.* (1991b).

[21] Geladi, P., H. Isaksson, L. Lindquist, S. Vold, and K. Esbensen, Principal component analysis of multivariate images, *Chemometrics and Intelligent Laboratory Systems* **5** (1989), 209–220.

[22] Gordon, H. R., J. W. Brown, O. B. Brown, R. H. Evans, and D. K. Clark, Nimbus 7 CZCS: Reduction of its radiometric sensitivity with time, *Appl. Opt.* **22** (1983), 3929–3931.

[23] Goshtasby, A., G. C. Stockman, and C. V. Page, A regional-based approach to digital image registration with subpixel accuracy, *IEEE Trans. Geosci. and Remote Sensing* **24** (1986), 390–399.

[24] Gratzki, A., and S. A. W. Gerstl, Sensitivity of an atmospheric correction algorithm for non-Lambertian vegetation surfaces to atmospheric parameters, *IEEE Trans. Geosci. and Remote Sensing* **27** (1989), 326–331.

[25] Hastie, T., and W. X. Stuetzle, Principal curves, *J. Am. Stat. Assoc.* **84** (1989), 502–516.

[26] Henderson-Sellers, A., Defogging cloud determination algorithms, *Nature* **298** (1982), 419–420.

[27] Hildreth, E. C., *The Measure of Visual Motion*, MIT Press, Cambridge, Mass., 1983.

[28] Horn, B. K. P., and B. G. Schunck, Determining optical flow, *Artif. Intell.* **17** (1981), 185–203.

[29] Ingerbritsen, S. E., and J. P. Lyon, Principal component analysis of multi-temporal image pairs, *Int. J. Remote Sensing* **6** (1985), 687–696.

[30] Jain, A. K., *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, N.J., 1989.

[31] Jain, A. K., and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, N.J., 1988.

[32] Jain, J. R., and A. K. Jain, Displacement measurement and its application to interframe image coding, *IEEE Trans. Commun.* **29** (1981), 1789–1808.

[33] Jensen, J. R., *Introductory Digital Image Processing*, Prentice Hall, Englewood Cliffs, N.J., 1986.

[34] Jezching, T., and A. K. Jain, Registering Landsat images by point matching, *IEEE Trans. Geosci. and Remote Sensing* **27** (1989), 642–651.

[35] Kaufman, Y. J., Atmospheric effects on spectral signature—Measurements and corrections, *IEEE Trans. Geosci. and Remote Sensing* **26** (1988), 441–449.

[36] Kelly, K. A., The influence of winds and topography on surface temperature patterns over the northern California slope, *J. Geophys. Res.* **90** (1985), 11,783–11,793.

[37] Lagerloef, G. S. E., EOF analysis of AVHRR and CZCS data, *Third Conference on Satellite Meteorology and Oceanography*, 1986.

[38] Lillesand, T. M., and R. W. Kiefer, *Remote Sensing and Image Interpretation*, 2nd ed., John Wiley and Sons, New York, 1987.

[39] Luenberger, D. G., *Optimization by Vector Space Methods*, John Wiley and Sons, New York, 1969.

[40] Marr, D., and E. C. Hildreth, Theory of edge detection, *Proc. R. Soc. London, B* **207** (1980), 187–217.

[41] Marr, D., and S. Ullman, Directional sensitivity and its use in early visual primitives, *Proc. R. Soc. London, B* **211** (1981), 151–180.

[42] Mather, P. M., *Computer Processing of Remotely-Sensed Images: An Introduction*, John Wiley and Sons, New York, 1987.

[43] Matveev, L. T., *Computer Processing of Remotely Sensed Images: An Introduction*, John Wiley and Sons, New York, 1984.

[44] Muerle, J. L., and D. C., Allen, Experimental evaluation of techniques for automatic segmentation of objects in a complex scene, pp. 3–13 in *Pictorial Pattern Recognition*, G. Cheng *et al.*, eds., Thompson, Washington, D.C., 1968.

[45] Ohring, G., and P. F. Clapp, The effects of changes in cloud amount on the net radiation at the top of the atmosphere, *J. Atmos. Sci.* **37** (1980), 447–454.

[46] Pailleux, F. S., The impact of satellite data on global numerical weather prediction, in *Remote Sensing Applications in Meteorology and Climatology*, R. A. Vaughn, ed., NATO ASI Series, vol. 201, 1986.

[47] Pairman, D., and J. Kittler, Clustering algorithms for use with images of clouds, *Int. J. Remote Sensing* **7** (1986), 855–866.

[48] Pavlidis, T., *Structural Pattern Recognition*, Springer-Verlag, Berlin, 1977.

[49] Preisendorfer, R. W., *Principal Component Analysis in Meteorology and Oceanography*, C. D. Mobley, compiler ed., Developments in Atmospheric Science **17**, Elsevier, Amsterdam, 1988.

[50] Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes*, Cambridge University Press, New York, 1986.

[51] Ramanathan, V., The role of earth radiation budget studies in climate and global circulation research, *J. Geophys. Res.* **92** (1987), 4075–4095.

[52] Richards, J. A., *Remote Sensing Digital Image Analysis: An Introduction*, Springer-Verlag, New York, 1986.

[53] Samet, H., *Applications of Spatial Data Structures, Computer Graphics, Image Processing and GIS*, Addison-Wesley Publishing Corporation, New York, 1989.

[54] Seddon, A. M., and G. E. Hunt, Segmentation of clouds using cluster analysis, *Int. J. Remote Sensing* **6** (1985), 717–731.

[55] Serra, J., *Image Analysis and Mathematical Morphology*, Academic Press, New York, 1982.

[56] Simpson, J. J., Image segmentation using recursive polygon fill operations, *Remote Sensing Environ.*, submitted (1991).

[57] Simpson, J. J., and J. Bloom, Objective estimates of velocity from sequences of satellite data using log search and minimum distortion methods, *J. Geophys. Res.*, submitted (1990).

[58] Simpson, J. J., and C. Humphrey, An automated cloud screening algorithm for daytime AVHRR imagery, *J. Geophys. Res.* **95** (1990), 13,459–13,481.

[59] Singh, A., and A. Harrison, Standardized Principal Components, *Int. J. Remote Sensing* **6** (1985), 883–896.

[60] Strub, P. T., C. James, A. C. Thomas, and M. A. Abbott, Seasonal and non-seasonal variability of satellite-derived surface pigment concentration in the California Current, *J. Geophys. Res.* **95** (1990), 11,501–11,530.

[61] Verri, A., and T. Poggio, Motion field and optical flow: Qualitative properties, *IEEE Trans. Pattern Anal. Machine Intell.* **11** (1989), 490–498.

[62] Wahl, D. D., and J. J. Simpson, Physical processes affecting the objective determination of near-surface velocity from satellite data, *J. Geophys. Res.* **95** (1990a), 13,511–13,528.

[63] Wahl, D. D., and J. J. Simpson, Satellite-derived estimates of the normal and tangential components of near-surface flow, *Int. J. Remote Sensing*, in press (1990b).

[64] Welch, Terry A., A technique for high performance data compression, *IEEE Computer* **17** (June 1984), 8–19.

Appendix to Chapter 3

Listed below are definitions and explanations for abbreviations commonly used in remote sensing.

AVHRR—The Advanced Very High Resolution Radiometer on the National Oceanic and Atmospheric Administration's polar-orbiting weather satellite. It measures cloud cover and infrared sea surface temperature.

DFT—Discrete Fourier transform.

EOF—Empirical orthogonal function.

EOS—Earth Observing System. A proposed National Aeronautics and Space Administration program for earth observing systems to be launched between 1997 and 2007.

FFT—Fast Fourier transform.

GOES—Geostationary Operational Environmental Satellite. An operational weather satellite used to measure cloud cover. Estimates of solar radiation often are computed from GOES data.

LOG—Laplacian of the Gaussian operator.

LZW—Lempel-Ziv-Welch coding used in data compression algorithms.

MN—Minimum norm solution.

MU—The Marr-Ullman solution for the normal component of velocity.

OF—Optical flow method of computation.

RLC—Run length coding used in data compression algorithms.

SST—Sea surface temperature.

4

Spatial Statistics in Environmental Science

Peter Guttorp
University of Washington

4.1 Introduction

During the last 15 years much attention has been focused on environmental problems, such as tree and lake death from acidic precipitation, global warming due to increased carbon dioxide concentration, and a possible reduction of the ozone layer in the stratosphere. For example, the problem of long-term trends in atmospheric deposition was the subject of a recent report of the National Research Council (1986). Many statistical problems are emerging from research in the environmental sciences. This chapter addresses the estimation of spatial covariance, with an application to a solar radiation network. Also discussed briefly are some aspects of monitoring network design and the usefulness of point process models in developing global climate models.

4.2 Estimating Spatial Covariance

The fundamental problem of environmetrics is that the observable processes of interest are highly variable. Noise typically overwhelms the signal. For example, when studying wet deposition of sulfate or nitrate at a location, the variability of rainfall constitutes a large fraction of the observed variability (Pollack *et al.*, 1989). Statistically precise methods for signal extraction are vital for policymakers.

In order to assess the severity of an environmental insult, the researcher typically has access to monitoring data from a relatively sparse network of stations, while assessment of the mean level (averaged both temporally and spatially) is needed over unobserved locations. Thus it is necessary to use spatial interpolation methods. The most common such method, namely kriging, is discussed in Chapter 5 of this report. A Bayesian nonparametric method for interpolation, called regularization, has been developed by Zidek and coworkers (Weerahandi and Zidek, 1988; Ma *et al.*, 1986) with environmental applications in mind. Common to these methods is the necessity to determine the spatial covariance.

The development of nonparametric procedures for interpolating observed spatial covariances of a random function sampled at a finite number of locations has lagged well behind the development of interpolation methods for the expected value of the underlying function. The kriging and regularization methods mentioned above depend explicitly on the spatial covariance or variogram functions. Most approaches to modeling spatial covariance structure have been parametric and have assumed isotropy and/or stationarity. The best-known models are parametric forms for the variogram originating in Matheron's theory of regionalized variables. The common assumption of a spatially stationary variogram in kriging analyses was called the "intrinsic dispersion law" by Matheron. Switzer and Loader (1989) propose a less parametrically oriented method to fit empirical dispersion or covariances. Since the empirical site-pair covariances may themselves be subject to sampling variability, some degree of parametric modeling is required, which at the same time respects the apparent heterogeneity in the covariance field. Basically, a parametric covariance model is forced on the available empirical covariances, and modified covariance estimates are obtained by shrinking toward the parametric covariances.

A nonparametric approach to global estimation of the spatial covariance structure of a random function $Z(x,t)$ observed repeatedly at a finite number of sampling stations x_i, $i = 1,2,\ldots,N$, in the plane has been developed by Sampson and Guttorp (1990). The true covariance structure is assumed to be neither isotropic nor stationary, but rather a smooth function of the geographic coordinates of station pairs (x_i, x_j). Using a variant of multidimensional scaling (MDS), a two-dimensional representation for the sampling stations is computed for which the *spatial dispersions* $\mathrm{Var}(Y(x_i) - Y(x_j))$ are approximated by a monotone function of interpoint distances. That is, in terms of this second two-dimensional representation, the spatial covari-

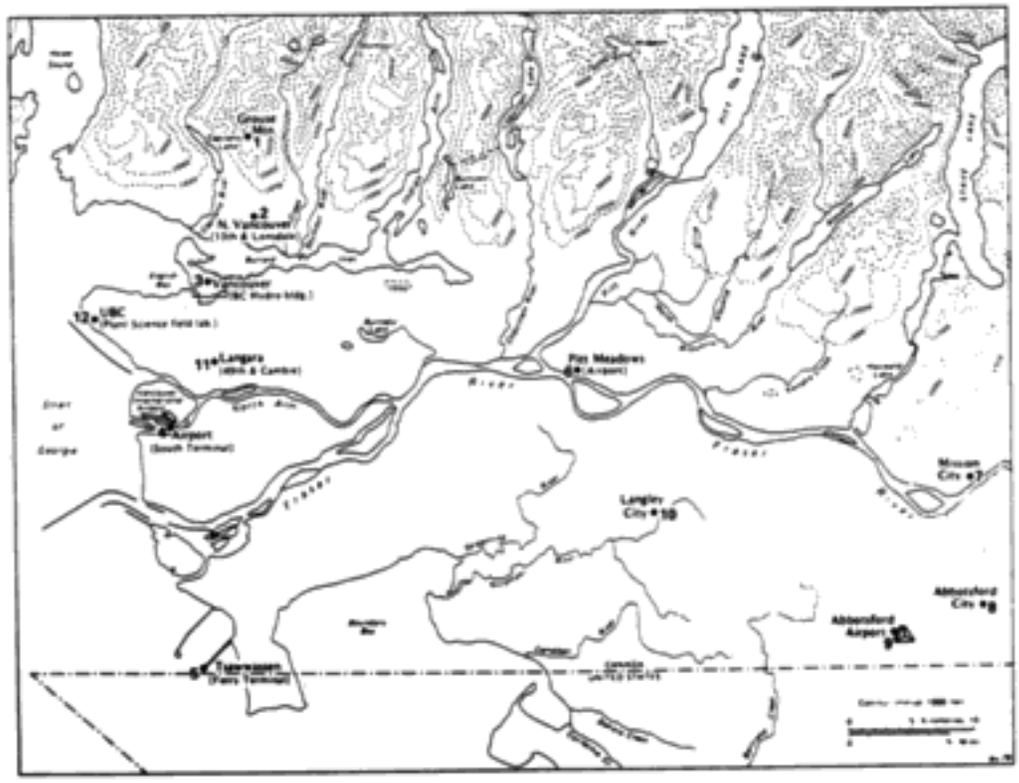


FIGURE 4.1: The 12-station solar radiation monitoring network in Lower Mainland, British Columbia, Canada. Reprinted, by permission, from Hay (1984). Copyright © 1984 by Pergamon Press.

ance structure as represented by the spatial dispersions is stationary and isotropic. (These variances are usually fitted by parametric models for the *variogram*.) Thinplate splines are applied to compute a smooth mapping of the geographic representation of the sampling stations onto the MDS representation. *Bi-orthogonal grids*, introduced by Bookstein (1978) in the field of morphometrics, can be used to depict the mapping. This mapping yields a nonparametric method for estimating $\mathrm{Var}(Y(x_a) - Y(x_b))$ for any two unsampled locations x_a and x_b in the geographic plane, and a graphical representation of the global spatial covariance structure. The resulting nonparametric models for spatial covariance are constrained to be positive-definite—or, in the terminology of geostatistics, the variogram models are conditionally non-negative-definite. This is obtained by fitting a mixture of covariance functions of Gaussian type in the MDS step of the algorithm.

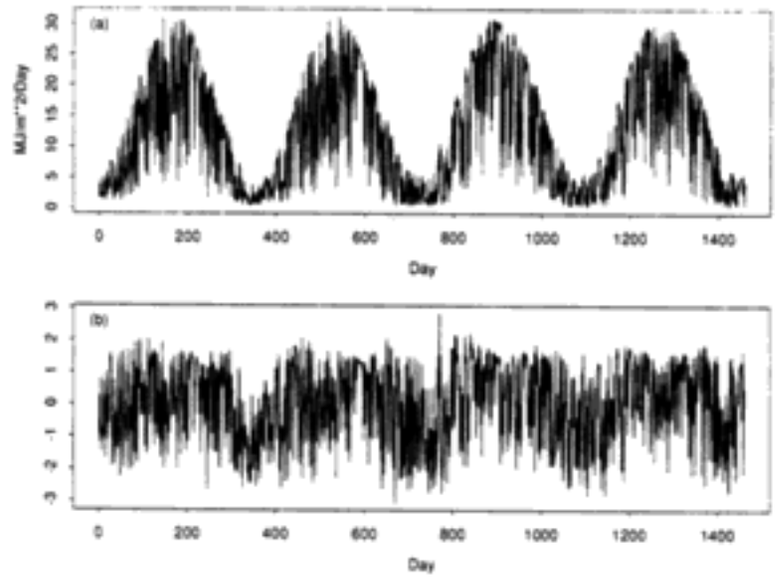


FIGURE 4.2: Daily solar radiation totals for Vancouver International Airport (site 4): (a) raw; (b) transformed.

4.2.1 Example: Spatial Variation in Solar Radiation

We present here a preliminary analysis of data collected from a solar radiation monitoring network in southwestern British Columbia, Canada (Hay, 1984), with a view toward determining the feasibility of solar power generation in British Columbia. This example manifests a somewhat extreme but easily understood form of nonstationarity in the spatial covariance structure of the solar radiation field. Figure 4.1, taken from Hay (1983), displays the locations of the 12 monitoring stations.

The data consist of daily solar radiation totals (MJ $m^2 day^{-1}$) for the years 1980–83. Figure 4.2 plots the data for the monitoring station at Vancouver International Airport. Note the relatively sharp upper bound on the maximum solar radiation as a function of season. Sivkov (1971, Chap. 7) explains how and why the maximum solar radiation (observed on cloudless days) varies approximately as a sine function with minimum at the vernal equinox. A reasonable stochastic model for observations at one location is thus

$$Z_{i,t} = \theta_{i,t}\left(\alpha + \beta \sin\left[\frac{2\pi}{365}(t - 80)\right]\right)(1 + \epsilon_{i,t}),$$

where observations are taken daily ($t = 1, 2, \ldots, 365$), $\theta_{i,t}$ is a random variable taking values on the interval $(0, 1]$ to express atmospheric attenuation effects, and $\epsilon_{i,t}$ represents a mean zero measurement error effect. Cloudiness is the principal factor determining θ_t. As the first step in our analysis, we estimate the parameters α and β, which define the maximum expected solar radiation as a function of day of year. We then scale all the data as a percentage of the estimated seasonally adjusted maximum possible solar radiation. Thus we attempt to focus on analyzing the spatial structure of θ_t. These data have a concentration of values near the maximum of 100%, and so we compute covariances among monitoring stations using a logit transformation of the percentage-of-maximum data. These transformations removed the major aspect of seasonality associated with the orientation of the earth with respect to the sun. However, the spatial covariance structure retains seasonal structure because of variation in the atmospheric processes. We therefore analyze the spatial structure of the data separately by season. Here we present only the results for the combined spring and summer quarters (vernal equinox, March 22, through autumnal equinox, September 22).

Interstation correlations are very high for these data, and the dispersions are closely related to geographic distances among the stations. Figure 4.3 shows the distribution of monitoring stations in the D-plane as determined by MDS applied to the matrix of dispersions. The most obvious deviation between the two planar representations is in the relative location of station 1, Grouse Mountain. The Grouse Mountain station is at an elevation of 1128 meters while all other stations lie below 130 meters. This orographic feature explains the relatively high dispersions (low covariances) between station 1 and all the others as reflected in the scaling in Figure 4.3.

4.3 Network Design

The purpose of a monitoring network is to detect potential changes in key environmental parameters. The designer of a long-term monitoring network cannot fully foresee all of the benefits that may be derived from the network by its future users. Environmental engineers, resource developers, biologists, human health agents, and so on, will need the data for a variety of purposes, some of which will not even have been identified. In addition to the hypothesis testing mentioned above, there is a need for inference about changes in areal averages, and about the areal maximum of such changes. The network may be regarded simply as an information gathering device.

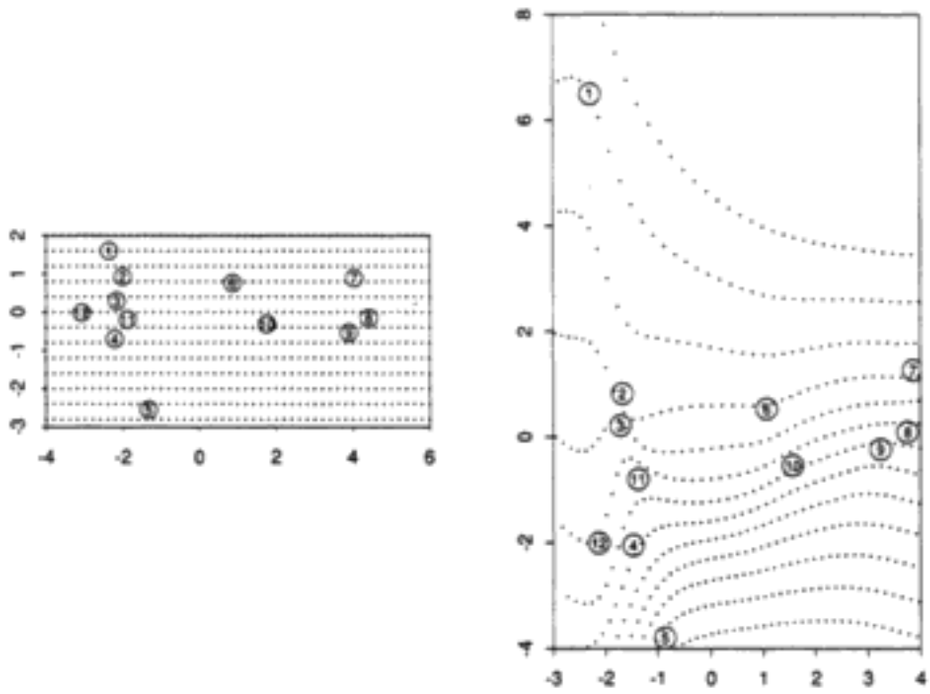


FIGURE 4.3: Transformation of the G-plane configuration of solar radiation monitoring stations (left) into the D-plane configuration (right).

There are objectives where the choice of design may not be critical. Switzer (1979) argues that for estimating areal averages, the search for a design that minimizes mean squared estimation error is unnecessary, since the criterion is relatively insensitive to design changes among sensible designs. The optimal design is very model-dependent, and the mathematics are invariably difficult. He argues that designs intended for this purpose might better be chosen on *a priori* grounds, avoiding clustering and with regard to topography and subregions of greater variability. Unfortunately, the situation is not always so simple. In impact detection, for example, the choice of the design is critical.

Kriging has its attendant theory of design, based on minimizing mean squared estimation error (Cressie *et al.*, 1990). For impact design, this criterion may not be the most natural. Rather, one wants to maximize the power of the test.

In general, the appropriate design criterion is as uncertain as the objective itself (see Rodriguez-Iturbe, 1974, for a discussion). Caselton and Zidek

(1984) argue that a reasonable design criterion will be based on an index of the information transmitted. A particular set of monitoring stations is good if it provides a lot of information (in the sense of Shannon) about unmonitored sites (see §4.3.2).

4.3.1 Impact Design

Suppose we need to assess the effect of a potential impact taking place at a known time. Typical examples are changes in environmental requirements, closure or startup of potential pollution sources, and environmental disasters. The null hypothesis is that of constant mean before and after the change. Suppose that it is feasible to make observations at any point on the grid of potential monitoring sites before and after the known time of potential change. According to an emerging body of evidence, it is very difficult to detect even fairly large changes in ambient levels with high probability. For example, Hirsch and Gilroy (1985) use a certain nonparametric testing procedure, a sulfate deposition model fitted to data from New York state, and simulated sulfate deposition experiments with step changes of various magnitudes, including 20%. They show that with one monitoring station, 90% power requires 15 years of post-change records with 5 years of pre-change records. Using 8 stations, one still needs 2 years of post-change records, and adding more stations does not yield appreciable reductions. Much of the difficulty is the result of the large component of meteorological variability in deposition. In the work of Vong *et al.* (1988), a design based on meteorological criteria was used to reduce this variability, which yielded unambiguous evidence of the local deposition effect of a copper smelter.

Regard a design D as a set of labels designating the sampling sites. The region of interest is overlain with an imaginary grid of potential sites from which D is to be chosen. An impact is regarded as a random field Z, covering the whole region. At site i, Z_i is the size of the change owing to development and other uncontrolled factors. Only Z_i with i in D will in fact be measured (with error) once D is specified.

Suppose that K replicate measurements of Z_i are taken at each site in D. Their variability is assumed constant over i, and indicates the precision of the process of measurement. Changes will be measured against this variance. A strategy (suggested, e.g., by Green, 1979) can be used to reduce the impact of temporal effects. Sites outside areas of likely impact are admitted as possible quasi-controls. These do increase the power of tests, even though they, strictly speaking, are not controls. The null hypothesis (again following

Green, 1979) is that of no time-space interaction. Assuming the standard two-way fixed effects ANOVA model, the F-statistic has power depending on the noncentrality parameter which can be estimated by

$$\delta^2 = K \sum \frac{(Z_i - \bar{Z}_D)^2}{2\sigma^2},$$

where $\bar{Z}_D$ is the average of the observations. In some special cases it is possible to maximize $\mathbf{E}(\delta^2)$. Suppose that the area of potential impact can be divided into a collection of homogeneous zones (this has to be done using expert knowledge). Then the problem of maximizing the expected non-centrality parameter is reduced to that of finding the optimal sampling fractions, which is a quadratic integer programming problem (Schumacher and Zidek, 1989). Simulated annealing is being explored as an alternative approach to the optimization (Sacks and Schiller, 1988).

4.3.2 Information Transmission Network Design

The future benefits that may be derived from a network cannot all be specified in advance. Even when a network is designed with a particular objective in mind, it is quite common that the answer to very different questions must be elicited from the data once the network is operational. Caselton and Zidek (1984) suggest circumventing these difficulties by an approach that may be suboptimal in specific cases but has overall merits for these types of networks.

We let $\mathbf{Z}$ denote a random field of measurable quantities indexed by potential site labels i. We decompose $\mathbf{Z}$ into the gauged sites $G = (Z_i,\ i \in \mathrm{D})$ and the ungauged sites U. The choice of D will be made to maximize the amount of information in G about U. Here the information measure is taken to be $I(U,G) = \mathbf{E}(\log(f(U|G)/f(U)))$, Shannon's index of information transmission, where $f(U|G)$ is the conditional density of U given G, and $f(U)$ the *a priori* density of U.

A simple special case is when the random field is multivariate normal, when $I(U,G) = -\frac{1}{2}\log|I - R|$, where I is the identity matrix, and R the diagonal matrix whose elements are the squared canonical correlation coefficients between U and G. These can be obtained from estimates of the spatial correlations, for example, using the method of Sampson and Guttorp mentioned in the previous section.

For particular patterns of the covariance matrix of $\mathbf{Z}$, derived from models of acidic deposition (such as that used by Vong *et al.*, 1988), it is possi-

FIGURE 4.4: The MAP3S monitoring network.

ble to develop workable approximations to the canonical correlations and to solve the design problem in terms of signal-to-noise ratios at gauged and ungauged sites, respectively. The analysis suggests the importance of replicate measurements at gauged sites (Guttorp *et al.*, 1987, section 3.3).

Example: Finding the Least Informative Station in a Network

The Multistate Atmospheric Power Product Pollution Study/Precipitation Chemistry Network (MAP3S/PCN) of nine monitoring stations (Figure 4.4) in the northeastern United States was initiated in 1976 with the objective of creating a long-term, high-quality data base for the development of regional transport and deposition models. There is substantial seasonal variability in the data, and we concentrate here on log deposition of H^+, using four-week totals for January through April. Guttorp *et al.* (1991) has further details. In order to decide which station carries the least information in the network, we need to compute the information in the network leaving out each station in turn. Thus the station left out is considered ungauged, and all the other

TABLE 4.1: Multiple Correlation Coefficients

Station(U)	$I(U,G)$	standard error
Lewes, Del.	.26	.08
Illinois, Ill.	.66	.10
Ithaca, N.Y.	.49	.09
Whiteface, N.Y.	.40	.09
Brookhaven, N.Y.	.42	.09
Oxford, Ohio	.58	.10
Penn State, Pa.	.57	.10
Virginia, Va.	.31	.08
Oak Ridge, Tenn.	.29	.08

stations are gauged. For each station left out, we compute $I(U,G)$ from the other stations in the network. The analysis of canonical correlations (which for one ungauged site simplifies to the multiple correlation coefficient) indicates that the three stations in Illinois, Ohio, and Pennsylvania each have significantly higher multiple correlations with the remainder of the network than have any other stations. The results are listed in Table 4.1, where it is seen that Illinois, Ill., is the least informative station in the network, in the sense of being best predicted by the other stations. In other words, the gauged stations have the highest information about the (presumed) ungauged station at Illinois.

It is worth noting that the stations at Oxford, Ohio, and Penn State, Pennsylvania, are not significantly different from the Illinois station. On the other hand, the geographically extreme stations in Delaware, Virginia, and Tennessee are all poorly predicted, and are therefore highly informative stations.

4.4 Modeling Precipitation Using Space-Time Point Processes

An environmental problem of enormous potential impact is the global warming due to increased CO_2 concentration in the atmosphere. Much effort has been extended to develop realistic models of global climate in order to be able to assess the potential impact of changes in atmospheric gasses on dif-

ferent aspects of weather patterns. In order to do this, hydrologists have found it useful to employ stochastic models of precipitation, which is an important factor in climate change, and also itself affected by climate change. Such models have also found important applications in assessing the risk of flash floods and in design of dams.

A realistic stochastic model of rainfall must take into account the physical structure and organization of storms, such as the description of cyclonic storms in Hobbs and Locatelli (1978). In essence, the storm system contains mesoscale rainbands, which contain smaller mesoscale regions, or precipitation cores, which are characterized by higher rainfall rates. These cores originate in generating cells aloft (in warm frontal bands) or within layers of potentially unstable air (in cold frontal bands). This description was used by Waymire *et al.* (1984) and by Kavvas and Herd (1985) to construct appropriate stochastic models, following the work of Le Cam (1961). In what follows, we essentially follow the Waymire *et al.* description.

The essence of the Waymire *et al.* (1984) model is the following stochastic representation of the rainfall intensity ξ at time t and location $\mathbf{z}$:

$$\xi(t,\mathbf{z}) = \int_{\mathbf{R}^2} \int_{(0,t]} g_\eta(t-\tau\,;\,\mathbf{z}-\mathbf{y})\, X(d\tau, d\mathbf{y})\,,$$

where g_η is a dispersion function, representing the rainfall intensity from a given cell born at $(\tau,\mathbf{y})$ depending on the random variable η, and $X(d\tau,d\mathbf{y})$ counts the rain cells alive in an infinitesimal neighborhood of $(\tau,\mathbf{y})$. Thus X is a point process that has the structure of a cluster process (see Daley and Vere-Jones, 1988, and the discussion in chapter 7 of this report). From this representation, it is easy to write down formulae for the mean and covariance of the random field ξ. In order to get useful results, one needs to make a few more assumptions. If it is reasonable to assume that the dispersion of a rain cell is independent of the occurrence of rain cells, then the expected value can be written

$$\mathbf{E}\xi(t,\mathbf{z}) = \iint \mathbf{E}[g_\eta(t-\tau\,;\,\mathbf{z}-\mathbf{y})]\, p_X^{(1)}(\tau,\mathbf{y})\, d\tau\, d\mathbf{y}\,,$$

where $p_X^{(k)}$ is the k^{th} order product moment density for the point process X, measuring the joint probability density of k events. It may be reasonable to assume that spatial and temporal features are separable, in the sense that

$$p_X^{(1)}(\tau,\mathbf{y}) = p_1^{(1)}(\tau)\, p_2^{(1)}(\mathbf{y})$$

and

$$g_\eta(\mathbf{u};\mathbf{v}) = \eta g_1(\mathbf{u}) g_2(\mathbf{v}).$$

With these assumptions, it is easy to see that

$$\mathbf{E}\xi(t,\mathbf{z}) = \mathbf{E}(\eta)[g_1 * p_1^{(1)}](t)\,[g_2 * p_2^{(1)}](\mathbf{z}),$$

where $[f_1 * f_2]$ is the convolution of f_1 with f_2.

Similar computations yield that

$$\begin{aligned}\mathbf{E}\xi(t_1,\mathbf{z}_1)\xi(t_2,\mathbf{z}_2) &= \mathbf{E}^2(\eta)\left[\iint g_1(t_1-\tau_1)g_1(t_2-\tau_2)p_1^{(2)}(\tau_1,\tau_2)\,d\tau_1\,d\tau_2\right] \\ &\quad\times\left[\iint g_2(\mathbf{z}_1-\mathbf{y}_1)g_2(\mathbf{z}_2-\mathbf{y}_2)p_2^{(2)}(\mathbf{y}_1,\mathbf{y}_2)\,d\mathbf{y}_1\,d\mathbf{y}_2\right] \\ &\quad+\mathbf{E}(\eta^2)\left[\int g_1(t_1-\tau)g_1(t_2-\tau)p_1^{(1)}(\tau)\,d\tau\right]\left[\int g_2(\mathbf{z}_1-\mathbf{y})g_2(\mathbf{z}_2-\mathbf{y})p_2^{(1)}(\mathbf{y})\,d\mathbf{y}\right] \\ &\equiv J_1 + J_2.\end{aligned}$$

If, in addition, $p_X^{(2)}(\tau_1,\mathbf{z}_1;\tau_2,\mathbf{z}_2) = p_X^{(1)}(\tau_1,\mathbf{z}_1)p_X^{(1)}(\tau_2,\mathbf{z}_2)$, the covariance simplifies to J_2.

Most processes of interest can be written as a function of the intensity process ξ. For example, the dry area in a region A during the time interval (t_1,t_2) can be expressed as

$$\int_{t_1}^{t_2}\int_A 1(\xi(t,\mathbf{z}) < \epsilon)\,d\mathbf{z}\,dt,$$

where $1(B)$ is the indicator function of the set B, and ϵ is the limit of detectability. Of course, the process ξ itself cannot be observed; we only observe time integrals of ξ at given points.

The detailed structure of the parameter functions discussed here is currently the emphasis of intense research in the hydrological community. A discussion of some of these features is given in Guttorp (1988). Recent advances in satellite and radar imagery enables the identification of some of the major features of the model, and thus can both suggest functional forms for some of the parameter functions and permit testing the goodness of fit of the model. The problem of parameter identification from time-averaged quantities is discussed in Guttorp (1986) for the nonspatial case when only presence or absence of precipitation at a single station in each time interval is recorded, and in Guttorp and Thompson (1990) for the case when counts of

the number of events in each time interval are recorded. Generally, because of the intractable nature of the likelihood function, estimation is usually based on the method of moments. Further discussion of problems involved in spatial and temporal averaging of precipitation data and the attendant problems of parameter estimation can be found, e.g., in Rodriguez-Iturbe *et al.* (1974), Valdes *et al.* (1985), Rodriguez-Iturbe and Eagleson (1987), Sivapalan and Wood (1987), and Phelan (1991).

Bibliography

[1] Bookstein, F. L., *The Measurement of Biological Shape and Shape Change*, Lecture Notes in Biomath. **24**, Springer, New York, 1978.

[2] Caselton, W. F., and J. V. Zidek, Optimal network monitoring design, *Stat. Prob. Lett.* **2** (1984), 223–227.

[3] Cressie, N., C. A. Gotway, and M. O. Grondona, Spatial prediction from networks, *Chemometrics Int. Lab. Syst.* **7** (1990), 251–271.

[4] Daley, D. J., and D. Vere-Jones, *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York, 1988.

[5] Green, R. H., *Sampling Designs and Statistical Methods for Environmental Biologists*, John Wiley and Sons, New York, 1979.

[6] Guttorp, P., On binary time series obtained from continuous time point process models describing rainfall, *Water Resour. Res.* **22** (1986), 897–904.

[7] Guttorp, P., Analysis of event based precipitation data with a view towards modeling, *Water Resour. Res.* **24** (1988), 35–44.

[8] Guttorp, P., and M. L. Thompson, Nonparametric estimation of intensities for sampled counting processes, to appear *J. R. Stat. Soc., B*, 1990.

[9] Guttorp, P., A. J. Petkau, P. D. Sampson, and J. V. Zidek, *Environmental Monitoring: Models, Network Design, and Data Analysis*, Department of Statistics, University of Washington, SIMS Technical Report **107** (1987).

[10] Guttorp, P., K. Newman, and P. D. Sampson, Nonparametric estimation of spatial covariance with an application to monitoring network design, to appear in *Statistics in Environmental and Earth Sciences*, P. Guttorp and A. Walden, eds., Griffin, London, 1991.

[11] Hay, J. E., Solar energy system design: The impact of mesoscale variations in solar radiation, *Atmosphere-Ocean* **21** (1983), 138–157.

[12] Hay, J. E., An assessment of the mesoscale variability of solar radiation at the earth's surface, *Solar Energy* **32** (1984), 425–434.

[13] Hirsch, R. M., and E. J. Gilroy, (1985), Detectability of step trends in rate of atmospheric deposition of sulfate, *Water Resour. Bull.* **21** (1985), 773–784.

[14] Hobbs, P. V., and J. D. Locatelli, Rainbands, precipitation cores and generating cells in a cyclonic storm, *J. Atmos. Sci.* **35** (1978), 230–241.

[15] Kavvas, M. L., and K. R. Herd, A radar-based stochastic model for short-time-increment rainfall, *Water Resour. Res.* **21** (1985), 1437–1455.

[16] Le Cam, L. M., A stochastic description of precipitation, pp. 165–186 in *Proceedings Fourth Berkeley Symposium on Mathematical Statistics and Probability* **3**, J. Neyman, ed., University of California Press, Berkeley, 1961.

[17] Ma, H. W., H. Joe, and J. V. Zidek, *A Bayesian Nonparametric Univariate Smoothing Method, with Applications to Acid Rain Data Analysis*, SIMS Technical Report, **104** (1986), University of British Columbia.

[18] National Research Council, *Acid Deposition: Long-Term Trends*, Committee on Monitoring and Assessment of Trends in Acid Deposition, National Academy Press, Washington, D.C., 1986.

[19] Phelan, M. J., Aging functions and their nonparametric estimation in point process models of rainfall, to appear in *Statistics in Environmental and Earth Sciences*, P. Guttorp and A. Walden, eds., Griffin, London, 1991.

[20] Pollack, A. K., A. B. Hudischewskyj, T. S. Stoeckenius, and P. Guttorp, Analysis of Variability of UAPSP Precipitation Chemistry Measurements, Draft Final Report SYSAPP-89/041, Systems Applications, Inc., San Rafael, 1989.

[21] Rodriguez-Iturbe, I., The design of rainfall networks in time and space, *Water Resour. Res.* **10** (1974), 713–728.

[22] Rodriguez-Iturbe, I., and P. S. Eagleson, Mathematical models of rainstorm events in space and time, *Water Resour. Res.* **23** (1987), 181–190.

[23] Sacks, J., and S. Schiller, Spatial designs, in S. S. Gupta and J. O. Berger (eds.), *Statistical Decision Theory and Related Topics IV*, vol. 2, Springer, New York, 1988.

[24] Sampson, P. D., and P. Guttorp, *Nonparametric Estimation of Non-Stationary Spatial Covariance Structure*, Department of Statistics, University of Washington, SIMS Technical Report **148** (1990).

[25] Schumacher, P., and J. V. Zidek, Using prior information in designing point impact detection networks, talk at ISI satellite meeting on Statistics, Earth and Space Sciences (1989), Leuven (to appear).

[26] Sivapalan, M., and E. F. Wood, A multidimensional model of non-stationary space-time rainfall at the catchment scale, *Water Resour. Res.* **23** (1987), 1289–1299.

[27] Sivkov, S. I., *Computation of Solar Radiation Characteristics*, Israel Program for Scientific Translations, Jerusalem, 1971.

[28] Switzer, P., Statistical considerations of network design, *Eos* (1979), 1712–1716.

[29] Switzer, P., and C. Loader, Spatial covariances, talk at ISI satellite meeting on Statistics, Earth and Space Sciences (1989), Leuven (to appear).

[30] Valdes, J. B., I. Rodriguez-Iturbe, and V. K. Gupta, Approximations of temporal rainfall from a multidimensional model, *Water Resour. Res.* **21** (1985), 1259–1270.

[31] Vong, R. J., L. Moseholm, D. S. Covert, P. D. Sampson, J. F. O'Loughlin, M. N. Stevenson, R. J. Charlson, W. H. Zoller, and T. V. Larson, Changes in rainwater acidity associated with closure of a copper smelter, *J. Geophys. Res., D* **93** (1988), 7169–7179.

[32] Waymire, E., V. K. Gupta, and I. Rodriguez-Iturbe, A spectral theory of rainfall intensity at the meso-β scale, *Water Resour. Res.* **20** (1984), 1453–1465.

[33] Weerahandi, S., and J. V. Zidek, Bayesian nonparametric smoothers. *Can. J. Stat.* **16** (1988), 61–74.

5

Geostatistical Analysis of Spatial Data

Noel Cressie
Iowa State University

5.1 Introduction

All data have (more or less) precise spatial and temporal labels associated with them. That is, a measurement is obtained from a particular location at a particular time, although that information may be lost by omission or made less precise by aggregation. For most of this chapter, it is assumed that only the data's spatial labels are important—hence the term *spatial data*.

As a discipline, spatial statistics has components of all the classical areas of statistics, such as design, statistical methods (including data analysis and diagnostics), stochastic modeling, and statistical inference. Importantly, the spatial labels form an integral part of a spatial statistical analysis. Geostatistics is the area of spatial statistics that is concerned mostly with prediction of unknown values at given locations (or of aggregations over given regions). Typically, the prediction is based on univariate and bivariate distributions of the spatial values, and these distributions (or appropriate moments of them) are estimated from an initial analysis of the data.

The prefix "geo" in geostatistics originally implied statistics pertaining to the earth (Matheron, 1963; see also Hart, 1954, who used the term differently from Matheron, in a geographical context). However, more recently, geostatistics has been used to solve problems in agricultural engineering, atmospheric science, ecology, forestry, hydrology, meteorology, remote sensing, etc. Although it is Matheron's development of the area within mining

that is best known, a Soviet meteorologist, L. S. Gandin, independently developed a framework for inference that is virtually identical (Gandin, 1963); he chose the term "objective analysis" instead of "geostatistics."

Section 5.2 presents the basic ideas behind a geostatistical analysis, including a brief discussion of splines and conditional simulation. The first part of §5.3 gives several applications of geostatistics, and the second part discusses recent advances and future directions.

5.2 Theory and Methods of Geostatistics

Geostatistics is mostly concerned with spatial prediction, but there are other important areas, such as model selection, effect of aggregation, and spatial sampling design, that offer fruitful open problems; see §5.3.2. The emphasis in this section will be on a spatial-prediction method known as *kriging*. Matheron (1963) coined the term in honor of D. G. Krige, a South African mining engineer (see Cressie, 1990, for an account of the origins of kriging).

5.2.1 The Variogram

First, a measure of the (second-order) spatial dependence exhibited by the spatial data is needed. A model-based parameter (which is a function) known as the variogram is defined here; its estimate provides such a measure. Statisticians are used to dealing with the autocovariance function. It is demonstrated here that the class of processes with a variogram contains the class of processes with an autocovariance function, and that kriging can be carried out on a wider class of processes than the one traditionally used in statistics.

Let $\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbf{R}^d\}$ be a real-valued stochastic process defined on a domain D of the d-dimensional space $\mathbf{R}^d$, and suppose that differences of variables lagged $\mathbf{h}$-apart vary in a way that depends only on $\mathbf{h}$. Specifically, suppose

$$\text{var}(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = 2\gamma(\mathbf{h}) \quad \text{for all } \mathbf{s}, \mathbf{s}+\mathbf{h} \in D\,; \tag{5.1}$$

typically the spatial index $\mathbf{s}$ is two- or three-dimensional (i.e., $d = 2$ or 3). The quantity $2\gamma(\cdot)$, which is a function only of the *difference* between the spatial locations $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$, has been called the *variogram* by Matheron (1963), although earlier appearances in the scientific literature can be found. It has been called a *structure function* by Yaglom (1957) in

probability and by Gandin (1963) in meteorology, and a *mean-squared difference* by Jowett (1952) in time series. Kolmogorov (1941) introduced it in physics to study the local structure of turbulence in a fluid. Nevertheless, it has been Matheron's mining terminology that has persisted. The variogram must satisfy the conditional negative semi-definiteness condition, $\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j 2\gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0$, for any finite number of spatial locations $\{\mathbf{s}_i : i = 1, \ldots, k\}$, and real numbers $\{a_i : i = 1, \ldots, k\}$ satisfying $\sum_{i=1}^{k} a_i = 0$. When $2\gamma(\mathbf{h})$ can be written as $2\gamma^0(\|\mathbf{h}\|)$, for $\mathbf{h} \in \mathbf{R}^d$, the variogram is said to be *isotropic*.

Variogram models that depend only on a few parameters $\boldsymbol{\theta}$ can be used as summaries of the spatial dependence and as an important component of optimal linear prediction (kriging). Three basic isotropic models, given here in terms of the semivariogram (half the variogram), are:

Linear model (valid in $\mathbf{R}^d$, $d \geq 1$)

$$\gamma(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} 0 & \mathbf{h} = \mathbf{0} \\ c_0 + b_\ell \|\mathbf{h}\| & \mathbf{h} \neq \mathbf{0} \end{cases}$$
$$\text{where } \boldsymbol{\theta} = (c_0, b_\ell),\ c_0 \geq 0,\ b_\ell \geq 0;$$

Spherical model (valid in $\mathbf{R}^1$, $\mathbf{R}^2$, and $\mathbf{R}^3$)

$$\gamma(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} 0 & \mathbf{h} = \mathbf{0} \\ c_0 + c_s[\frac{3}{2}(\|\mathbf{h}\|/a_s) - \frac{1}{2}(\|\mathbf{h}\|/a_s)^3] & 0 < \|\mathbf{h}\| \leq a_s \\ c_0 + c_s & \|\mathbf{h}\| \geq a_s \end{cases}$$
$$\text{where } \boldsymbol{\theta} = (c_0, c_s, a_s),\ c_0 \geq 0,\ c_s \geq 0,\ a_s \geq 0;$$

Exponential model (valid in $\mathbf{R}^d$, $d \geq 1$)

$$\gamma(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} 0 & \mathbf{h} = \mathbf{0} \\ c_0 + c_e[1 - \exp(-\|\mathbf{h}\|/a_e)] & \mathbf{h} \neq \mathbf{0} \end{cases}$$
$$\text{where } \boldsymbol{\theta} = (c_0, c_e, a_e),\ c_0 \geq 0,\ c_e \geq 0,\ a_e \geq 0.$$

Another semivariogram model is the *rational quadratic model* (valid in $\mathbf{R}^d$, $d \geq 1$):

$$\gamma(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} 0 & \mathbf{h} = \mathbf{0} \\ c_0 + \frac{c_r\|\mathbf{h}\|^2}{1+\|\mathbf{h}\|^2/a_r} & \mathbf{h} \neq \mathbf{0} \end{cases}$$
$$\text{where } \boldsymbol{\theta} = (c_0, c_r, a_r),\ c_0 \geq 0,\ c_r \geq 0,\ a_r \geq 0.$$

A semivariogram model that exhibits negative correlations caused by periodicity of the process is the *wave* (or *hole-effect*) model (valid in $\mathbf{R}^1$, $\mathbf{R}^2$, and $\mathbf{R}^3$):

$$\gamma(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} 0 & \mathbf{h} = \mathbf{0} \\ c_0 + c_w \frac{1 - a_w \sin(\|\mathbf{h}\|/a_w)}{\|\mathbf{h}\|} & \mathbf{h} \neq \mathbf{0} \end{cases}$$
$$\text{where } \boldsymbol{\theta} = (c_0, c_w, a_w),\ c_0 \geq 0,\ c_w \geq 0,\ a_w \geq 0\,.$$

A further condition that a variogram model must satisfy is (Matheron, 1971)

$$2\gamma(\mathbf{h})/\|\mathbf{h}\|^2 \to 0 \text{ as } \|\mathbf{h}\| \to \infty\,.$$

In fact, the *power* semivariogram model,

$$\gamma(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} 0 & \mathbf{h} = \mathbf{0} \\ c_0 + b_p\|\mathbf{h}\|^\lambda & \mathbf{h} \neq \mathbf{0} \end{cases}$$
$$\text{where } \boldsymbol{\theta} = (c_0, b_p, \lambda),\ c_0 \geq 0,\ b_p \geq 0,\ 0 \leq \lambda < 2\,,$$

is a valid semivariogram model in $\mathbf{R}^d$, $d \geq 1$.

When the process Z is anisotropic (i.e., dependence between $Z(\mathbf{s})$ and $Z(\mathbf{s}+\mathbf{h})$ is a function of both the magnitude *and* the direction of $\mathbf{h}$), the variogram is no longer purely a function of distance between two spatial locations. Anisotropies are caused by the underlying physical process evolving differentially in space. Sometimes the anisotropy can be corrected by a linear transformation of the lag vector $\mathbf{h}$. That is,

$$2\gamma(\mathbf{h}) = 2\gamma^0(\|A\mathbf{h}\|),\ \mathbf{h} \in \mathbf{R}^d\,,$$

where A is a $d \times d$ matrix and $2\gamma^0$ is a function of only one variable.

Replacing (5.1) with the stronger assumption

$$\text{cov}(Z(\mathbf{s}+\mathbf{h}), Z(\mathbf{s})) = C(\mathbf{h}) \quad \text{for all } \mathbf{s},\ \mathbf{s}+\mathbf{h} \in D \tag{5.2}$$

and specifying the mean function to be constant, i.e.,

$$\mathrm{E}(Z(\mathbf{s})) = \mu \quad \text{for all } \mathbf{s} \in D\,, \tag{5.3}$$

defines the class of *second-order* (or wide-sense) *stationary* processes in D, with covariance function $C(\cdot)$. Time series analysts often assume (5.2) and work with the quantity $\rho(\cdot) \equiv C(\cdot)/C(\mathbf{0})$. Conditions (5.1) and (5.3) define

the class of *intrinsically stationary* processes, which is now shown to contain the class of second-order stationary processes.

Assuming only (5.2),

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}), \tag{5.4}$$

that is, the semivariogram is related very simply to the covariance function. An example of a process for which $2\gamma(\cdot)$ exists but $C(\cdot)$ does not is a one-dimensional standard Wiener process $\{W(t) : t \geq 0\}$. Here, $2\gamma(h) = |h|$ $(-\infty < h < \infty)$, but $\text{cov}(W(t), W(u)) = \min(t, u)$, which is not a function of $|t-u|$. Thus, the class of intrinsically stationary processes *strictly* contains the class of second-order stationary processes.

Now consider estimation of the variogram from data $\{Z(\mathbf{s}_i) : i = 1, \ldots, n\}$. Suppose these are observations on an intrinsically stationary process (i.e., a process that satisfies (5.1) and (5.3)), taken at the n spatial locations $\{\mathbf{s}_i : i = 1, \ldots, n\}$. Because of (5.3), $\text{var}(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = \text{E}(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2$. Hence, the method-of-moments estimator of the variogram $2\gamma(\mathbf{h})$ is

$$2\hat{\gamma}(\mathbf{h}) \equiv \sum_{N(\mathbf{h})} [Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2 / |N(\mathbf{h})|, \quad \mathbf{h} \in \mathbf{R}^d, \tag{5.5}$$

where the average in (5.5) is taken over $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}\}$, and $|N(\mathbf{h})|$ is the number of distinct elements in $N(\mathbf{h})$. For irregularly spaced data, $N(\mathbf{h})$ is usually modified to $\{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j \in T(\mathbf{h})\}$, where $T(\mathbf{h})$ is a tolerance region of $\mathbf{R}^d$ surrounding $\mathbf{h}$. Other estimators, more robust than (5.5), are given in Cressie and Hawkins (1980) and Cressie (1991, sec. 2.4). Parametric models, $2\gamma(\cdot;\boldsymbol{\theta})$, can be fit to the estimator (5.5) by various means; as a compromise between efficiency and simplicity, Cressie (1985) advocates minimizing a weighted sum of squares

$$\sum_{k=1}^{K} \left\{ \frac{2\hat{\gamma}(\mathbf{h}(k))}{2\gamma(\mathbf{h}(k);\boldsymbol{\theta})} - 1 \right\}^2 |N(\mathbf{h}(k))|$$

with respect to variogram model parameters $\boldsymbol{\theta}$. The sequence $\mathbf{h}(1), \ldots, \mathbf{h}(K)$ denotes the "lags" at which an estimator (5.5) was obtained, and which satisfy range and replication conditions such as those given by Journel and Huijbregts (1978, p. 194, eq. III.42). Zimmerman and Zimmerman (1991) summarize and compare several methods of variogram-parameter estimation based on simulated Gaussian data. They find that the weighted-least-squares approach usually performs well, and never does poorly, against such competitors as maximum likelihood estimation (both ordinary and restricted) and minimum norm quadratic unbiased estimation.

5.2.2 Kriging

For the purposes of this section, assume that the variogram is known; in practice, variogram parameters are estimated from the spatial data. Suppose it is desired to predict $Z(\mathbf{s}_0)$ at some unsampled spatial location $\mathbf{s}_0$ using a linear function of the data $\mathbf{Z} \equiv (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$:

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i)\,. \tag{5.6}$$

It is sensible to look for coefficients $\{\lambda_i : i = 1, \ldots, n\}$ for which (5.6) is uniformly unbiased and which minimize the mean-squared prediction error $\mathrm{E}(Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0))^2$. More generally, one could try to minimize $\mathrm{E}(L[Z(\mathbf{s}_0), p(\mathbf{Z})])$ with respect to predictor $p(\mathbf{Z})$, where L is a loss function. For example, the loss function proposed by Zellner (1986),

$$L[Z(\mathbf{s}_0), p(\mathbf{Z})] = b\,\{\exp[a(Z(\mathbf{s}_0) - p(\mathbf{Z}))] - a(Z(\mathbf{s}_0) - p(\mathbf{Z})) - 1\},\ b > 0\,,$$

allows overprediction to incur a different loss than underprediction. Minimizing mean-squared prediction error results from using

$$L[Z(\mathbf{s}_0), p(\mathbf{Z})] = b[Z(\mathbf{s}_0) - p(\mathbf{Z})]^2,\ b > 0\,,$$

which is the squared-error loss function. In all that is to follow, squared-error loss is used.

The uniform unbiasedness condition imposed on (5.6) is simply $\mathrm{E}(\hat{Z}(\mathbf{s}_0)) = \mu = \mathrm{E}(Z(\mathbf{s}_0))$, for all $\mu \in \mathbf{R}$, which is equivalent to

$$\sum_{i=1}^{n} \lambda_i = 1\,. \tag{5.7}$$

Now, assuming (5.7), the mean-squared prediction error can be written in two ways. If the process is second-order stationary,

$$\mathrm{E}(Z(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i))^2 = C(0) - 2\sum_{i=1}^{n} \lambda_i C(\mathbf{s}_i - \mathbf{s}_0) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(\mathbf{s}_i - \mathbf{s}_j)\,, \tag{5.8}$$

or, if the process is intrinsically stationary (a weaker assumption),

$$\mathrm{E}(Z(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i))^2 = 2\sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_i - \mathbf{s}_0) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{s}_i - \mathbf{s}_j)\,. \tag{5.9}$$

Using differential calculus and the method of Lagrange multipliers, optimal coefficients $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ can be found that minimize (5.9) subject to (5.7); they are

$$\boldsymbol{\lambda} = \Gamma^{-1}\left[\boldsymbol{\gamma} + \frac{(1 - \mathbf{1}'\Gamma^{-1}\boldsymbol{\gamma})\mathbf{1}}{\mathbf{1}'\Gamma^{-1}\mathbf{1}}\right], \tag{5.10}$$

and the minimized value of (5.9) (kriging variance) is

$$\sigma_k^2(\mathbf{s}_0) = \boldsymbol{\gamma}'\Gamma^{-1}\boldsymbol{\gamma} - \frac{(1 - \mathbf{1}'\Gamma^{-1}\boldsymbol{\gamma})^2}{\mathbf{1}'\Gamma^{-1}\mathbf{1}}. \tag{5.11}$$

In (5.10) and (5.11), $\boldsymbol{\gamma} = [\gamma(\mathbf{s}_1 - \mathbf{s}_0), \ldots, \gamma(\mathbf{s}_n - \mathbf{s}_0)]'$, $\mathbf{1} = (1, \ldots, 1)'$, and Γ is the $n \times n$ symmetric matrix with $(i,j)^{\text{th}}$ element $\gamma(\mathbf{s}_i - \mathbf{s}_j)$.

The kriging predictor given by (5.6) and (5.10) is appropriate if the process Z contains no measurement error. If measurement error is present, then a "noiseless version" of Z should be predicted; Cressie (1988) has details on when and how this should be implemented.

Thus far, kriging has been derived under the assumption of a constant mean. More realistically, assume

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \quad \mathbf{s} \in D, \tag{5.12}$$

where $\mathrm{E}(Z(\mathbf{s})) = \mu(\mathbf{s})$ for $\mathbf{s} \in D$ and $\delta(\cdot)$ is a zero-mean, intrinsically stationary stochastic process with $\mathrm{var}(\delta(\mathbf{s} + \mathbf{h}) - \delta(\mathbf{s})) = \mathrm{var}(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})) = 2\gamma(\mathbf{h})$, $\mathbf{h} \in \mathbf{R}^d$. In (5.12) the "large-scale variation" $\mu(\cdot)$ and the "small-scale variation" $\delta(\cdot)$ are modeled as deterministic and stochastic processes, respectively, but with no unique way of identifying either of them. What is one person's mean structure could be another person's correlation structure. Often this problem is resolved in a substantive application by relying on scientific or habitual reasons for determining the mean structure.

Suppose $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$, a linear combination of variables that could include trend-surface terms or other explanatory variables thought to influence the behavior of the large-scale variation. Thus,

$$Z(\mathbf{s}) = \sum_{j=0}^{p} x_j(\mathbf{s})\beta_j + \delta(\mathbf{s}), \quad \mathbf{s} \in D, \tag{5.13}$$

where $\boldsymbol{\beta} \equiv (\beta_0, \ldots, \beta_p)'$ are unknown parameters and $\delta(\cdot)$ satisfies (5.1) and (5.3) with zero mean. Although the model has changed, the problem of predicting $Z(\mathbf{s}_0)$ using an unbiased linear predictor (5.6) remains. The uniform unbiasedness condition is now equivalent to the condition

$$\boldsymbol{\lambda}'X = \mathbf{x}_0', \tag{5.14}$$

where $\mathbf{x}_0 \equiv (x_0(\mathbf{s}_0), \ldots, x_p(\mathbf{s}_0))'$ and X is an $n \times (p+1)$ matrix whose $(i,j)^{\text{th}}$ element is $x_{j-1}(\mathbf{s}_i)$. Then, provided (5.7) is implied by (5.14), minimizing the mean-squared prediction error subject to (5.14) yields the *universal kriging* predictor

$$\hat{Z}_U(\mathbf{s}_0) = \boldsymbol{\lambda}_U' \boldsymbol{Z}\,, \tag{5.15}$$

where

$$\boldsymbol{\lambda}_U = \Gamma^{-1}[\boldsymbol{\gamma} + X(X'\Gamma^{-1}X)^{-1}(\mathbf{x}_0 - X'\Gamma^{-1}\boldsymbol{\gamma})]\,; \tag{5.16}$$

the (universal) kriging variance is

$$\sigma_k^2(\mathbf{s}_0) = \boldsymbol{\gamma}'\Gamma^{-1}\boldsymbol{\gamma} - (X'\Gamma^{-1}\boldsymbol{\gamma} - \mathbf{x}_0)'(X'\Gamma^{-1}X)^{-1}(X'\Gamma^{-1}\boldsymbol{\gamma} - \mathbf{x}_0)\,. \tag{5.17}$$

Another way to write the equations (5.14) and (5.15) is

$$\hat{Z}(\mathbf{s}_0) = \mathbf{v}_1'\boldsymbol{\gamma} + \mathbf{v}_2'\mathbf{x}_0\,, \tag{5.18}$$

where $\mathbf{v}_1$ (an $n \times 1$ vector) and $\mathbf{v}_2$ (a $(p+1) \times 1$ vector) solve

$$\begin{aligned} \Gamma\mathbf{v}_1 + X\mathbf{v}_2 &= \mathbf{Z} \\ X'\mathbf{v}_1 &= \mathbf{0}\,. \end{aligned} \tag{5.19}$$

Equations (5.18) and (5.19) are known as the dual-kriging equations, since the predictor is now expressed as a linear combination of the elements of $(\boldsymbol{\gamma}', \mathbf{x}_0')$. From (5.19), it is clear that *spline smoothing* is equivalent in form to universal kriging (see Watson, 1984, where the relationship between the two prediction techniques is reviewed). Kriging has the advantage that in practice the data are first used to estimate the variogram, so adapting to the quality and quantity of spatial dependence in the data. Furthermore, kriging produces a mean-squared prediction error, given by (5.17), that quantifies the degree of uncertainty in the predictor. Cressie (1989b) presents these two faces of spatial prediction along with 12 others, including disjunctive kriging and inverse-distance-squared weighting.

5.2.3 Conditional Simulation of Spatial Data

Simulation of spatial data $\{Z(\mathbf{s}_i) : i = 1, \ldots, N\}$ with given means $\{\mu(\mathbf{s}_i) : i = 1, \ldots, N\}$ and covariances $\{C(\mathbf{s}_i, \mathbf{s}_j) : 1 \le i \le j \le N\}$ can be carried out in a number of ways, depending on the size of N and the sparseness of Σ_N, the $N \times N$ symmetric matrix whose $(i,j)^{\text{th}}$ element is $C(\mathbf{s}_i, \mathbf{s}_j)$. One way

is to use the Cholesky decomposition $\Sigma_N = L_N L_N'$, where L_N is a lower-triangular $N \times N$ matrix (e.g., Golub and Van Loan, 1983, pp. 86–90). Then $\mathbf{Z}_N \equiv (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_N))'$ can be simulated by

$$\mathbf{Z}_N = \boldsymbol{\mu}_N + L_N \boldsymbol{\epsilon}_N , \tag{5.20}$$

where $\boldsymbol{\mu}_N \equiv (\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_N))'$, and $\boldsymbol{\epsilon}_N$ is an $N \times 1$ vector of simulated independent and identically distributed random variables, each with zero mean and unit variance. Other methods, including polynomial approximations, Fourier transforms, and turning bands, are presented and compared in Cressie (1991, sec. 3.6).

Now consider the simulation of values of $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ *conditional* on observed values $\mathbf{Z}_n$. Call this conditionally simulated process $\{W(\mathbf{s}) : \mathbf{s} \in D\}$, and suppose $\{V(\mathbf{s}) : \mathbf{s} \in D\}$ is an unconditionally simulated process with the same first and second moments as $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$. For example, (5.20) might be used to simulate $\mathbf{V}_N \equiv (V(\mathbf{s}_1), \ldots, V(\mathbf{s}_N))'$, where $N \geq n$.

Consider conditional simulation at an arbitrary location $\mathbf{s}_{n+1}$ in D. Now write

$$\Sigma_{n+1} = \begin{bmatrix} \Sigma_n & \mathbf{c}_n \\ \mathbf{c}_n' & C(\mathbf{s}_{n+1}, \mathbf{s}_{n+1}) \end{bmatrix}$$

and notice that the two terms of the decomposition

$$Z(\mathbf{s}_{n+1}) = \mathbf{c}_n' \Sigma_n^{-1} \mathbf{Z}_n + [Z(\mathbf{s}_{n+1}) - \mathbf{c}_n' \Sigma_n^{-1} \mathbf{Z}_n] \tag{5.21}$$

are uncorrelated. Hence, the conditional simulation

$$W(\mathbf{s}_{n+1}) = \mathbf{c}_n' \Sigma_n^{-1} \mathbf{Z}_n + [V(\mathbf{s}_{n+1}) - \mathbf{c}_n' \Sigma_n^{-1} \mathbf{V}_n], \quad \mathbf{s}_{n+1} \in D , \tag{5.22}$$

has the same first two moments, unconditionally, as the process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ and $W(\mathbf{s}_i) = Z(\mathbf{s}_i)$, $i = 1, \ldots, n$. That is, unconditional simulation of sample paths of V yields, through (5.22), conditionally simulated sample paths of W.

It is apparent from (5.20) and (5.21) that when the ϵ_i's are Gaussian, so too is the process $\{W(\mathbf{s}) : \mathbf{s} \in D\}$. However, this may not reflect the reality of the conditional process when the original process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ is "far from" Gaussian, even though the first two moments match and the two processes agree at the data locations. There is clearly a danger in using conditional simulation uncritically.

5.3 Applications and Research Frontiers

A geostatistical analysis of spatial data has a "nonparametric" flavor to it, in that inferences are based on properties of univariate and bivariate distributions of $Z(\mathbf{s})$ and $Z(\mathbf{u})$, which are estimated from the data. In other words, assumptions are few, although often it helps to transform the data so that they are Gaussian-like. In contrast, Markov-random field models, or simultaneous spatial autoregressive models, have a very rigid structure that is not so well adapted to problems of spatial prediction (kriging). Section 5.3.1 shows how geostatistics has considerable flexibility in applications across diverse scientific disciplines.

5.3.1 Applications

The strength of geostatistics over more classical statistical approaches is that it recognizes spatial variability at both the "large scale" *and* the "small scale," or in statistical parlance, it models both spatial trend *and* spatial correlation. Trend-surface methods include only large-scale variation by assuming independent errors. Watson (1972) eloquently compares the two approaches and points out that most geological problems have a small-scale variation, typically exhibiting strong positive correlation between data at nearby spatial locations. The books by David (1977), Journel and Huijbregts (1978), and Clark (1979) are all aimed at applications of geostatistics in the mining industry.

The geostatistical method has also found favor among soil scientists who seek to map soil properties of a field from a small number of soil samples at known locations throughout the field; soil pH in water, soil electrical conductivity, exchangeable potassium in the soil, and soil-water infiltration are some of the variables that could be sampled and mapped.

Water erosion is of great concern to agriculturalists, since rich topsoil can be carried away in runoff water. The greater the soil-water infiltration, the less the runoff, resulting in less soil erosion and less stream pollution by pesticides and fertilizers. Also, greater infiltration implies better soil structure, which is more conducive to crop growth. Cressie and Horton (1987) describe how double-ring infiltrometers were placed at regular locations in a field that had received four tillage treatments, moldboard, paraplow, chisel, and no-till. From these data, the spatial relationships were characterized; Gotway and Cressie (1990) used the resulting stochastic models to estimate efficiently the tillage effects and to build a spatial analysis of variance table,

from which tillage differences can be tested.

Kriging can be applied in geophysical problems that require accurate mapping of the ocean floor. Data are slopes or depths and a variety of assumptions are made about the large-scale and small-scale variations defined by (5.12) (e.g., Shaw and Smith, 1987; Smith and Jordan, 1988; Gilbert, 1989; Malinverno, 1989). This area of investigation would benefit from geostatistical analyses that use the data initially to fit an appropriate variogram model and then draw kriging maps based on the fitted variogram.

Applications of geostatistics abound in other areas, such as rainfall precipitation (e.g., Ord and Rees, 1979), atmospheric science (e.g., Thiebaux and Pedder, 1987), acid deposition (e.g., Bilonick, 1985), and groundwater flow (e.g., Clark *et al.*, 1989). Examples from groundwater flow and acid deposition will now be used to illustrate the geostatistical method described in §5.2.

Flow of Groundwater from a Proposed Nuclear Waste Site

In 1986 three high-level nuclear waste sites were proposed in the United States (in Nevada, Texas, and Washington), thus prompting study of the soil and water-bearing properties of their surrounding regions. The chosen site will probably contain more than 68,000 high-level waste canisters placed about 30 feet apart in holes or trenches surrounded by salt, at a depth of 2,000 feet. However, leaks could occur, or the radioactive heat could cause the tiny quantities of water in the salt to migrate toward the heat until eventually each canister would be surrounded by about 6 gallons of water. The chemical reaction of salt and water would create hydrochloric acid that could slowly corrode the canisters. Eventually, the nuclear wastes could reach the aquifer and sometime later contaminate the drinking water.

Therefore, the types of questions one might ask are: If a nuclear waste site were to be designated for, say, Deaf Smith County, Texas, what are the risk parameters for radionuclides contaminating the groundwater? Where would they flow? How long would they take to get there? Here the direction-of-flow question will be addressed; kriging will be used to draw a spatial map of potentiometric heads throughout the area of interest.

Potentiometric heads in the West Texas/New Mexico region are shown in Figure 5.1, and are given by Harper and Furr (1986). They were measured by drilling a narrow pipe into the aquifer and letting the water find its own level in the pipe. Measurements are given in feet above sea level.

An anisotropic variogram model was fit to the data; in each of two

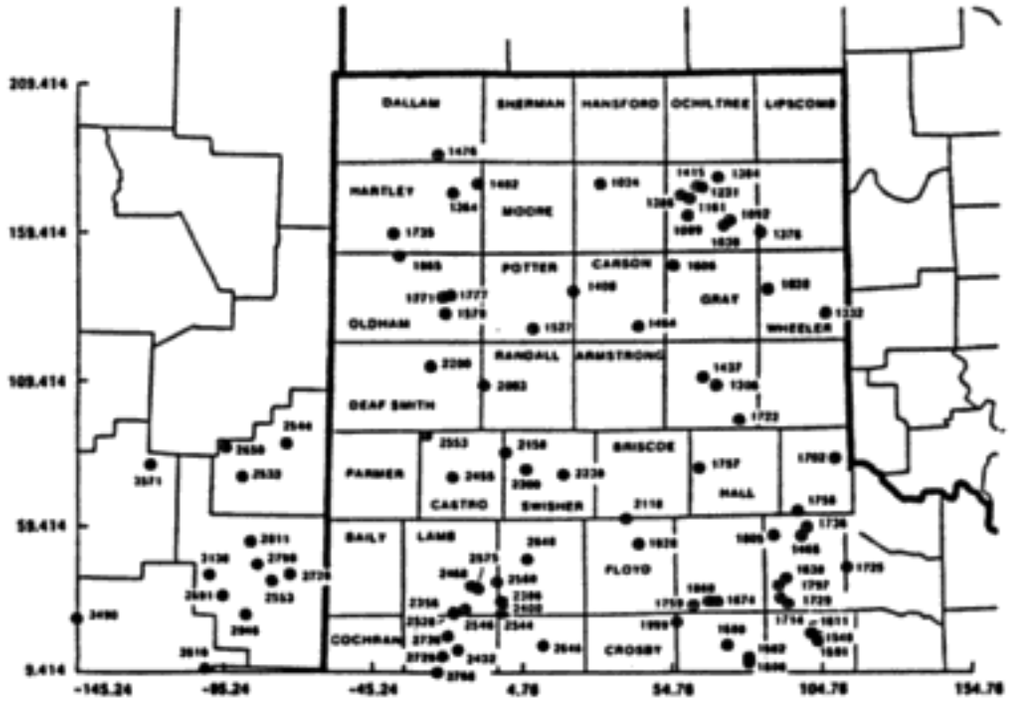


FIGURE 5.1: Locations and levels of piezometric-head data in West Texas/New Mexico. Amarillo is located close to the "1527" well in Potter Country. Reprinted, by permission, from Cressie (1989a). Copyright © 1989 by American Statistical Association.

orthogonal directions, values of $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ in

$$2\gamma(h;\boldsymbol{\theta}) = \begin{cases} 2\theta_1 + 2\theta_2|h|^{\theta_3} & h \neq 0 \\ 0 & h = 0 \end{cases}$$

were estimated. Restrictions on the parameters, in order that $2\gamma(\cdot\,;\boldsymbol{\theta})$ be a valid variogram, are $\theta_1 \geq 0$, $\theta_2 \geq 0$, and $0 \leq \theta_3 < 2$.

From the fitted variogram, kriging predictors $\{\hat{Z}(\mathbf{s}_0) : \mathbf{s}_0 \in D\}$ given by (5.6) and (5.10) and kriging standard errors $\{\sigma_k(\mathbf{s}_0) : \mathbf{s}_0 \in D\}$ given by (5.11) were obtained; see Cressie (1989a). Figure 5.2 shows the predicted surface, from which it can be concluded that contaminated groundwater from Deaf Smith County, Texas, would flow directly "downhill" toward

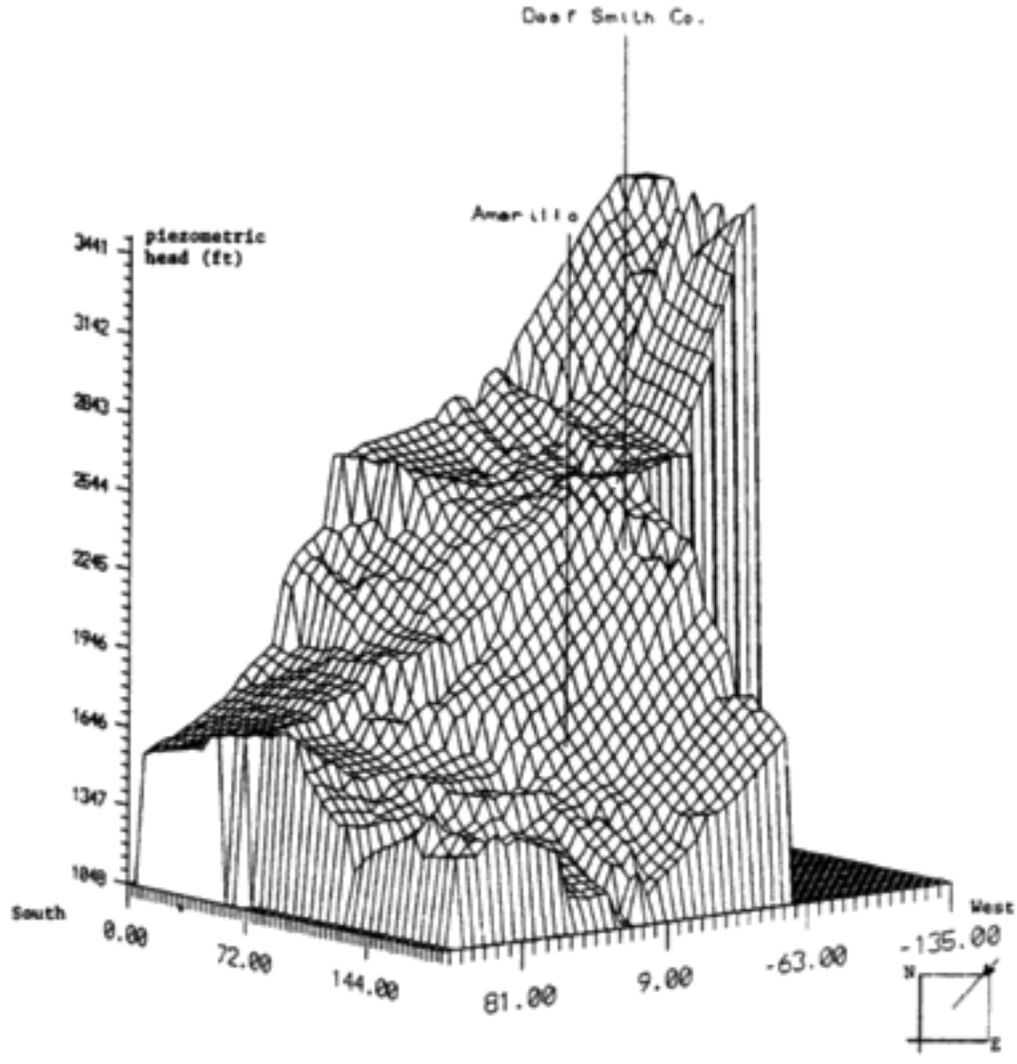


FIGURE 5.2: Three dimensional view of kriging surface $\{\hat{Z}(\mathbf{s}_0) : \mathbf{s}_0 \in D\}$, from the northeast corner of D. Reprinted, by permission, from Cressie (1989a). Copyright © 1989 by American Statistical Association.

Amarillo, Texas. However, Amarillans need not be concerned; a decision was made in 1987 by the U. S. Congress to locate the nation's high-level nuclear waste dump site in Nevada, probably at Yucca Mountain.

Acid Deposition and Network Design

It is generally accepted that an important factor in the relatively recent increase of acid deposition is the emission of industrial by-products into the atmosphere; the consequences for aquatic and terrestrial ecosystems are potentially disastrous. Most fish populations in freshwater lakes are very sensitive to changes in pH (EIFAC, 1969). More fundamentally, such changes could also adversely affect most other aquatic organisms and plants, resulting in a disruption of the food chain. Acid deposition has also been closely connected with forest decline (Pitelka and Raynal, 1989) in both Europe and the United States.

In the United States, acid deposition results mainly from the atmospheric alteration of sulfur and nitrogen air pollutants produced by industrial processes, combustion, and transportation sources. Total acid deposition includes acid compounds in both wet and dry form. Dry deposition is the removal of gaseous pollutants, aerosols, and large particles from the air by direct contact with the earth (NAPAP, 1988). Since dry deposition is difficult to monitor, and attempts at any such monitoring are relatively new, we focus on wet deposition here.

Wet deposition, or acid precipitation as it is commonly called, is defined as the hydrogen ion concentration in all forms of water that condenses from the atmosphere and falls to the ground. Measurement of the total annual amount of hydrogen ion is the end result of a very complicated process beginning with the release of pollutants into the atmosphere. They might remain there for up to several days and, depending on a variety of meteorological conditions (e.g., cold fronts or wind currents), they may be transported large distances. While in the atmosphere, the pollutants are chemically altered, then redeposited on the ground via rain, snow, or fog.

A model for the spatial distribution of total yearly hydrogen ion (H^+), measured on the Utility Acid Precipitation Study Program (UAPSP) network in 1982 and 1983, was developed by Cressie *et al.* (1990). We present their results for the 1982 data, including implications of the fitted model for network design.

Figure 5.3 is a map of the eastern half of the United States, showing the 19 UAPSP monitoring sites. Their latitudes, longitudes, and annual acid

FIGURE 5.3: Monitoring sites of the UAPSP network for the years 1982 and 1983. The square denotes an optimally located additional site. Reprinted, by permission, from Cressie *et al.* (1990). Copyright © 1990 by Elsevier Science Publishers, Physical Sciences and Engineering Division.

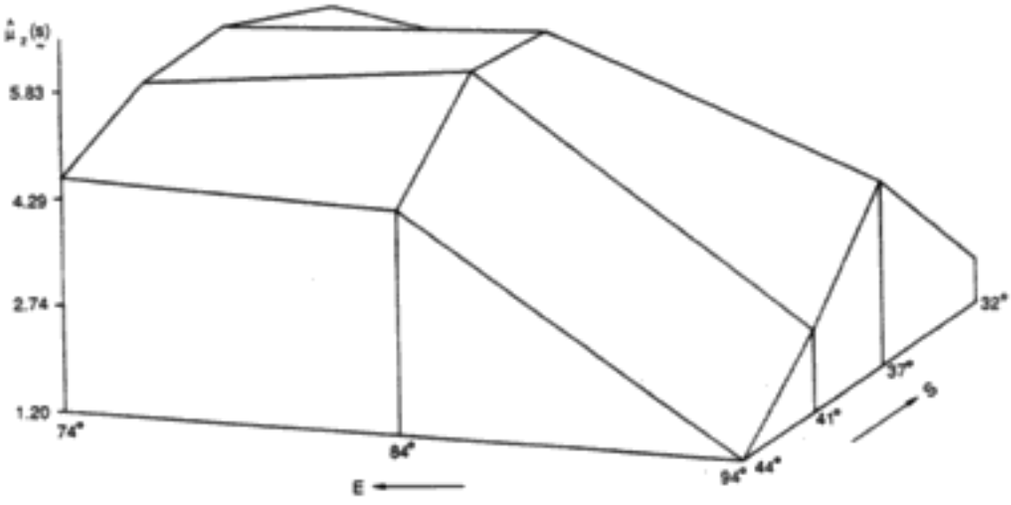


FIGURE 5.4: Median-polish surface obtained from the 1982 data. Units on the vertical axes are in μmoles H^+/cm^2. Reprinted, by permission, from Cressie *et al.* (1990). Copyright © 1990 by Elsevier Science Publishers, Physical Sciences and Engineering Division.

depositions (in μmole H^+/cm^2) for 1982 (and 1983) are presented in Cressie *et al.* (1990). By grouping nearby sites, a 4×3 table of acid-deposition data was constructed. The table was then median-polished (e.g., Emerson and Hoaglin, 1983), from which a crude picture of the large-scale variation was obtained; see Figure 5.4.

In the east-west direction there appears to be a positive linear trend, reflecting higher acid-deposition levels in the east. However, in the north-south direction, the trend is quadratic, with higher levels in the central region and lower levels in the extreme north and extreme south.

The surface in Figure 5.4 was subtracted from the original data to obtain residuals, $\{\hat{\delta}(\mathbf{s}_i) : i = 1, \ldots, 19\}$. Using great-arc distances to define distances between sites, an isotropic (robust) variogram estimator was computed, to which a spherical variogram model was fit by weighted least squares (§5.2.1). The fitted parameters were $\hat{c}_0 = 0.608(\mu\text{moles } H^+/cm^2)^2$, $\hat{c}_s = 2.041(\mu\text{moles } H^+/cm^2)^2$, and $\hat{a}_s = 361.210$ miles. Figure 5.5 gives a graphical representation of the results.

Optimal spatial prediction (ordinary kriging) can be implemented on the residual process $\hat{\delta}(\cdot)$ through equations (5.6), (5.10), and (5.11). A predicted surface of acid-deposition levels can then be obtained by adding back this kriging surface (of the residual process) to the median-polish surface shown in Figure 5.4. This is called *median-polish kriging* by Cressie (1986). The mean-squared errors of prediction (median-polish-kriging variances) $\{\sigma_k^2(\mathbf{s}_0) : \mathbf{s}_0 \in \text{eastern United States}\}$ are given by (5.11) and will now be used to choose the optimal location of a new site.

Let $S \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ denote the existing network and let $S_P \equiv \{\mathbf{s}_{n+1}, \ldots, \mathbf{s}_{n+m}\}$ denote $m \geq 2$ potential new sites from which one will be chosen. Define $S_{+i} \equiv S \cup \{\mathbf{s}_i\}$, $i = n+1, \ldots, n+m$ to be augmented networks. Then S_{+j} is preferred if it predicts best the remaining $m-1$ sites in S_P (on the average).

Specifically, let $\sigma_k^2(\mathbf{s}_0; S_{+i})$ denote the kriging variance for predicting the acid-deposition level at $\mathbf{s}_0$ using the augmented network S_{+i}, where $i = n+1, \ldots, n+m$. For illustration, define the objective function

$$V(\mathbf{s}_j) = \sum_{\substack{i=n+1 \\ i \neq j}}^{n+m} \sigma_k^2(\mathbf{s}_i; S_{+j})/(m-1), \quad j = n+1, \ldots, n+m. \tag{5.23}$$

Then the site in S_P that achieves $\min\{V(\mathbf{s}_j) : j = n+1, \ldots, n+m\}$ will be declared the optimal site to add. (Other criteria are considered in Cressie *et al.*, 1990.)

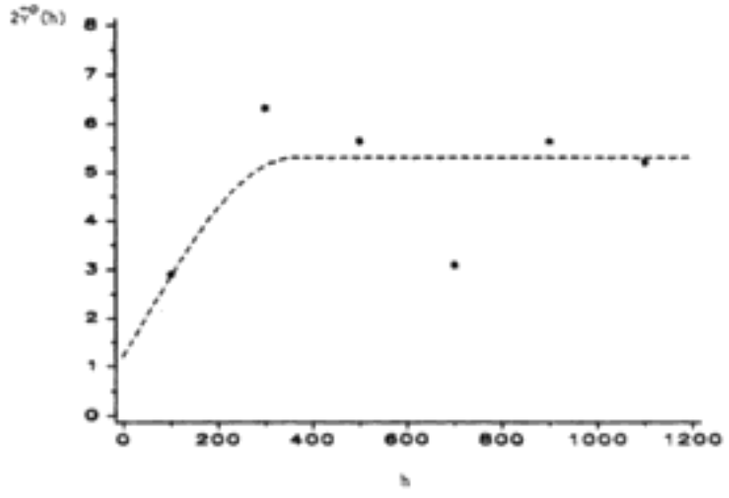


FIGURE 5.5: Empirical variograms (robust) for median-polished residuals. The superimposed dashed line indicates the weighted-least-squares fit. Units on the vertical axes are in $(\mu\text{moles } H^+/\text{cm}^2)^2$; units on the horizontal axes are in miles. Reprinted, by permission, from Cressie *et al.* (1990). Copyright © 1990 by Elsevier Science Publishers, Physical Sciences and Engineering Division.

Eleven potential sites (Minneapolis, Minnesota; Des Moines, Iowa; Jefferson City, Missouri; Madison, Wisconsin; Springfield, Illinois; Altoona, Pennsylvania; Charlottesville, Virginia; Charleston, West Virginia; Baltimore, Maryland; Trenton, New Jersey; and Knoxville, Tennessee) were chosen to improve geographic coverage of the existing network (of 19 sites). From among these eleven sites, Baltimore (marked with a square on Figure 5.3) was chosen as the optimal site to add. Its associated average kriging variance, given by (5.23), was $2.56\,(\mu\text{moles}\,H^+/\text{cm}^2)^2$, compared to Minneapolis's 2.59 (the second smallest value); Charlottesville had the largest value of 2.77.

5.3.2 Research Frontiers

Change of Support

The change-of-support problem remains a major challenge to geostatisticians. Although data come as $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$, inference may be required for $Z(B) \equiv \frac{1}{|B|}\int_B Z(\mathbf{u})\,d\mathbf{u}$. Kriging adapts very easily to ac-

commodate the change from point support $\mathbf{s}_0$ to block support B. For example, in (5.10) and (5.11), $\boldsymbol{\gamma}$ is modified to $\boldsymbol{\gamma}(B) \equiv [\frac{1}{|B|}\int_B \gamma(\mathbf{s}_1 - \mathbf{u})\,d\mathbf{u}, \ldots, \frac{1}{|B|}\int_B \gamma(\mathbf{s}_n - \mathbf{u})\,d\mathbf{u}]'$, and in (5.11) $\sigma_k^2(B)$ has the extra term $\frac{-1}{|B|^2}\int_B\int_B \gamma(\mathbf{u}-\mathbf{v})\,d\mathbf{u}\,d\mathbf{v}$. But in mining applications and emission compliance, for example, the quantity of greatest interest is the conditional distribution $\Pr(Z(B) > z \,|\, \mathbf{Z})$. Both disjunctive kriging (Matheron, 1976) and indicator kriging (Journel, 1983) attempt to answer this question based on bivariate distributional properties of the (possibly transformed) process. The problem is important enough to pursue beyond these initial approaches.

Multivariate Spatial Data

Prediction of a value $Z(\mathbf{s}_0)$ based on data $\mathbf{Z}$ and observations on other processes is known as *cokriging*. The appropriate generalization of the variogram (5.1) is the cross variogram

$$\operatorname{var}(Y(\mathbf{u}) - Z(\mathbf{v})) = 2\gamma_{YZ}(\mathbf{u},\mathbf{v}), \tag{5.24}$$

where $Y(\mathbf{u})$ and $Z(\mathbf{v})$ are normalized to have the same units. Cokriging equations for predicting $Z(\mathbf{s}_0)$ from $\mathbf{Z}$ and $\mathbf{Y}$ can be obtained in terms of γ_{ZZ}, γ_{YY}, and γ_{YZ} (Clark *et al.*, 1989). However, there is a dearth of models for (5.24); the basic requirement for a valid model is that its parameters can be estimated from the partial realization $(\mathbf{Z}', \mathbf{Y}')$ of the bivariate process.

Variogram Model Fitting and its Effect on Inferences

The variogram (5.1) has the property of conditional negative-definiteness. Based on a nonparametric estimator $2\hat{\gamma}(\cdot)$, say, current practice is to fit a parametric model $2\gamma(\cdot\,;\boldsymbol{\theta})$, which is guaranteed to be conditionally negative-definite. Is there a way to find a nonparametric fit to $2\hat{\gamma}(\cdot)$ from the set of all conditionally negative-definite functions? If it can be found, its description is not likely to be very parsimonious. Variogram-model choice should probably balance the closeness of its fit to the data, with its predictive power. For temporal data, Rissanen (1984, 1987) takes such an approach; however, his being able to sequence the observations is important, since the accumulated prediction errors form an integral part of his method. Development of a spatial version is an area worth investigating. Now, having chosen a model $2\gamma(\cdot\,;\hat{\boldsymbol{\theta}})$, what effect does the estimation of $\boldsymbol{\theta}$ have on inferences for $Z(\mathbf{s}_0)$? Zimmerman and Cressie (1991) have some initial results, but considerable further research is needed to resolve this important problem.

Bibliography

[1] Bilonick, R. A., The space-time distribution of sulfate deposition in the northeastern United States, *Atmos. Environ.* **17** (1985), 2513–2524.

[2] Clark, I., *Practical Geostatistics*, Applied Science Publishers, Essex, England, 1979.

[3] Clark, I., K. L. Basinger, and W. V. Harper, MUCK: A novel approach to co-kriging, pp. 473–493 in *Proceedings of the Conference on Geostatistical, Sensitivity, and Uncertainty Methods for Groundwater Flow and Radionuclide Transport Modeling*, B. E. Buxton, ed., Battelle Press, Columbus, Ohio, 1989.

[4] Cressie, N., Fitting variogram models by weighted least squares, *J. Int. Assoc. Math. Geol.* **17** (1985), 563–586.

[5] Cressie, N., Kriging nonstationary data, *J. Amer. Stat. Assoc.* **81** (1986), 625–634.

[6] Cressie, N., Spatial prediction and ordinary kriging, *Mathematical Geology* **20** (1988), 405–421.

[7] Cressie, N., Geostatistics, *Am. Stat.* **43** (1989a), 197–202.

[8] Cressie, N., The many faces of spatial prediction, pp. 163–174, in *Geostatistics, Vol. 1*, M. Armstrong, ed., Kluwer, Dordrecht, 1989b.

[9] Cressie, N., The origins of kriging, *Mathematical Geology* **22** (1990), forthcoming.

[10] Cressie, N., *Statistics for Spatial Data*, John Wiley and Sons, New York, forthcoming (1991).

[11] Cressie, N., and D. M. Hawkins, Robust estimation of the variogram, I, *J. Int. Assoc. Math. Geol.* **12** (1980), 115–125.

[12] Cressie, N. A. C., and R. Horton, A robust/resistant spatial analysis of soil-water infiltration, *Water Resour. Res.* **23** (1987), 911–917.

[13] Cressie, N., C. A. Gotway, and M. O. Grondona, Spatial prediction from networks, *Chemometrics and Intelligent Laboratory Systems* **7** (1990), 251–271.

[14] David, M., *Geostatistical Ore Reserve Estimation*, Elsevier, Amsterdam, 1977.

[15] Emerson, J. D., and D. C. Hoaglin, Analysis of two-way tables by medians, pp. 166–210, in *Understanding Robust and Exploratory Data Analysis*, D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds., John Wiley and Sons, New York, 1983.

[16] European Inland Fisheries Advisory Committee (EIFAC), Water quality criteria for European freshwater fish, *Water Res.* **3** (1969), 593–611.

[17] Gandin, L. S., *Objective Analysis of Meterological Fields*, Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), Leningrad, 1963 (translated by Israel Program for Scientific Translations, Jerusalem, 1965).

[18] Gilbert, L. E., Are topographic data sets fractal?, *Pure and Appl. Geophys.* **131** (1989), 241–254.

[19] Golub, G. H., and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 1983.

[20] Gotway, C. A., and N. Cressie, A spatial analysis of variance applied to soil-water infiltration, *Water Resour. Res.* **26** (1990), forthcoming.

[21] Harper, W. V., and J. M. Furr, *Geostatistical Analysis of Potentiometric Data in the Wolfcamp Aquifer of the Palo Duro Basin, Texas*, Technical report EMI/ONWI–587, Battelle Memorial Institute, Columbus, Ohio, 1986.

[22] Hart, J. F., Central tendency in areal distributions, *Economic Geography* **30** (1954), 48–59.

[23] Journel, A. G., Nonparametric estimation of spatial distributions, *J. Int. Assoc. Math. Geol.* **15** (1983), 445–468.

[24] Journel, A. G., and C. J. Huijbregts, *Mining Geostatistics*, Academic Press, London, 1978.

[25] Jowett, G. H., The accuracy of systematic sampling from conveyer belts, *Appl. Stat.* **1** (1952), 50–59.

[26] Kolmogorov, A. N., The local structure of turbulence in an incompressible fluid at very large Reynolds numbers, *Doklady Akademii Nauk SSSR* **30** (1941), 301–305. Reprinted in *Turbulence: Classic Papers on Statistical Theory*, S. K. Friedlander and L. Topping, eds., Interscience, New York, 1961, 151–155.

[27] Malinverno, A., Testing linear models of sea-floor topography, *Pure and Appl. Geophys.* **131** (1989), 139–155.

[28] Matheron, G., Principles of geostatistics, *Econ. Geol.* **58** (1963), 1246–1266.

[29] Matheron, G., *The Theory of Regionalized Variables and its Applications*, Cahiers du Centre de Morphologie Mathematique **5**, Fontainebleau, France, 1971.

[30] Matheron, G., A simple substitute for conditional expectation: The disjunctive kriging, pp. 221–236, in *Advanced Geostatistics in the Mining Industry*, M. Guarascio, M. David, and C. Huijbregts, eds., Reidel, Dordrecht, 1976.

[31] National Acid Precipitation Assessment Program (NAPAP), *Interim Assessment, the Causes and Effects of Acidic Deposition*, vols. I, II, III, IV, U. S. Government Printing Office, Washington, D.C., 1988.

[32] Ord, J. K., and M. Rees, Spatial processes: Recent developments with applications to hydrology, pp. 95–118 in *The Mathematics of Hydrology and Water Resources*, E. H. Lloyd, T. O'Donnell, and J. C. Wilkinson, eds., Academic Press, London, 1979.

[33] Pitelka, L. F., and D. J. Raynal, Forest decline and acid deposition, *Ecology* **70** (1989), 2–10.

[34] Rissanen, J., Universal coding, information, prediction, and estimation, *IEEE Trans. Inf. Theory* **30** (1984), 629–636.

[35] Rissanen, J., Stochastic complexity, *J. R. Stat. Soc. B* **49** (1987), 223–239.

[36] Shaw, P. R., and D. K. Smith, Statistical methods for describing seafloor topography, *Geophys. Res. Lett.* **14** (1987), 1061–1064.

[37] Smith, D. K., and T. H. Jordan, Seamount statistics in the Pacific Ocean, *J. Geophys. Res.* **93** (1988), 2899–2918.

[38] Thiebaux, H. J., and M. A. Pedder, *Spatial Objective Analysis with Applications in Atmospheric Science*, Academic Press, London, 1987.

[39] Watson, G. S., Trend surface analysis and spatial correlation, *Geol. Soc. Amer., Special Paper* **146** (1972), 39–46.

[40] Watson, G. S., Smoothing and interpolation by kriging and with splines, *J. Int. Assoc. Math. Geol.* **16** (1984), 601–615.

[41] Yaglom, A. M., Some classes of random fields in n-dimensional space, related to stationary random processes, *Theory Probab. Its Appl.* **2** (1957), 273–320.

[42] Zellner, A., Bayesian estimation and prediction using asymmetric loss functions, *J. Amer. Stat. Assoc.* **81** (1986), 446–451.

[43] Zimmerman, D. L., and N. Cressie, Mean squared prediction error in the spatial linear model with estimated covariance parameters, *Ann. Inst. Stat. Math.* **43** (1991), forthcoming.

[44] Zimmerman, D. L., and M. B. Zimmerman, A Monte Carlo comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors, *Technometrics* **33** (1991), forthcoming.

6

Spatial Statistics in the Analysis of Agricultural Field Experiments

Julian Besag
University of Washington[1]

6.1 Introduction

The main purpose of agricultural field experiments is to compare the effectiveness of different treatments (e.g., fertilizers) on a particular crop variety or to make comparisons between different varieties of the same crop. Accuracy is paramount, but valid assessment of error is also important. A typical experimental layout consists of a linear or two-dimensional array of contiguous rectangular strips of land, called plots, each of which is devoted to a single treatment or variety. Plots are usually long and narrow (e.g., 20 m × 2 m), partly as a trade-off between ease of management and compactness of the experiment.

In a linear layout, the longer sides are chosen to abut one another, so as to minimize the impact of fertility gradients across plots (see Figure 6.1). The most common measurement is that of plot yield at harvest, which in an ideal world would provide a direct assessment of the corresponding treatment or variety effect. However, yield is influenced by external factors such as weather and plot fertility. It can often be assumed that weather has a uniform effect, in which case comparisons remain valid under more general conditions; otherwise, several experiments may be required. On the other hand, variation in fertility over the experimental region is usually substantial

[1] Now at University of Newcastle upon Tyne.

and if ignored can lead to quite erroneous comparisons. Proper design and analysis of field experiments aims to minimize such problems. By the term design, here we mean, somewhat narrowly, the rule by which treatments are allocated to plots (henceforth we include the possibility that treatments are in fact different varieties).

Clearly, the task of controlling for variation in fertility is inherently spatial and is the focus of this chapter. In §6.2, we discuss general background, largely from a historical perspective, and in §6.3, we consider some recent progress and possible future directions at a more technical level. First, however, it should be noted that there are other types of field experiments that are not strictly covered by the above description. For example, interest may center on a measurement other than yield, such as resistance to disease or quality of product. Also there are experiments that are multisite or multistage or that involve mixtures, spacings, intercropping, competition, interference, and so on. Thus some experiments involve spatial considerations rather different from those on which we concentrate in this chapter; for example, in assessing resistance of different treatments to a particular pest, the main problem may concern patchiness of infestation over the experimental region. Nevertheless, we hope that in focusing on a single important topic, the richness of the subject as a whole will not be lost.

6.2 General Background

Methods of controlling for variation in fertility across the experimental region have a long history. Perhaps the first was the use of check plots (e.g., Wiancko, 1914); that is, plots interspersed regularly at frequent intervals throughout the experiment and containing a standard treatment. The yields from these plots can be used to calculate a local fertility index for each experimental plot and to adjust its yield accordingly; the assumption is that variations in adjacent plots are relatively small. Check plots are still employed in early generation selection trials where it is required to choose say 10% from many hundreds of varieties for further assessment. This selection is done at a stage where only a single plot is available to each variety because of restrictions of management, space, and the quantity of seed available. Besag and Kempton (1986) give an example involving 1560 different lines of winter wheat. However, for general use, check plots are rather crude, demand additional space, labor, and expense, and are somewhat self-defeating in that experimental plots become even more widely dispersed

over the field. Furthermore, there is no obvious means of judging the reliability of estimates. This is also true of systematic designs, such as the Knut Vik square, in which several plots, regularly dispersed over the experimental region, are devoted to each treatment. In this case, control for fertility is implicit rather than explicit; each treatment set of plots should be subject to approximately the same variation. Systematic designs also have a long history (e.g., Beaven, 1909) and still find favor in some Nordic countries. Two other possibilities are (1) the use of soil analysis in each plot to construct a fertility index, although this appears to have little if any support in practice, and (2) the incorporation of data from previous experiments, although this may be awkward operationally and requires the generally dubious assumption that fertility gradients remain approximately static from year to year.

The method of control that has now become standard in most countries was first proposed by R. A. Fisher in the 1920s. Fisher's triumph was to construct an entirely self-contained inferential framework that is valid whatever the pattern of fertility might be, subject only to an assumption of treatment additivity; that is, it is assumed that the relative effect of any particular treatment would be the same on any plot. The methodology relies on three key ingredients: *replication*, *blocking*, and *randomization*. Replication means that each treatment appears several times, usually with equal frequency, in the experiment; this generally improves accuracy and provides a basis for its assessment. Blocking implies that there are restrictions on the allocation of treatments to plots, which are imposed to counteract suspected fertility gradients. The intention is that fertility should be approximately constant within blocks, so that corresponding differences in plot yields are meaningful; this further improves the accuracy of treatment comparisons. Finally, randomization requires that treatments are allocated to plots entirely at random within the constraints of the blocking structure; it is this step that ensures the unbiasedness of treatment comparisons (contrasts) and the validity of the associated standard errors within the inferential framework.

Since this framework is not at all obvious, we provide some brief discussion in the particular context of the simplest common example, namely, the randomized complete blocks design, for which blocks and replicates coincide. Each plot yield is assumed to be the sum of three components, a fixed block effect, a treatment effect, and a plot effect. The natural means of introducing randomness into the model is to suppose that the plot effects represent a realization of a spatial stochastic process. The question, which is addressed

in §6.3, is, what process? Fisher brilliantly circumvented this problem by assuming plot effects to be fixed and by introducing randomness solely through the act of allocating treatments to plots. In other words, it is the very act of randomization that alone induces a probability distribution in the yields, and hence a basis for inference. For details see Kempthorne (1952, Ch. 8) but it turns out that the calculation of treatment estimates and associated standard errors coincides exactly with an ordinary least-squares analysis of the corresponding linear model assuming a fixed layout and uncorrelated random plot effects. For this reason, it is sometimes assumed that the latter formulation underpins the Fisherian analysis, whereas the two models are quite distinct, with the first addressing a fixed field with a randomly chosen layout and the second a random field with a fixed layout. Of course, in the second, an assumption of uncorrelated or equi-correlated plot effects within blocks is untenable if fertility gradients exist, as is generally the case unless the number of treatments (i.e., blocksize) is very small. Finally, note that the "usual" construction of confidence intervals based on the t-distribution can be shown to be approximately valid under randomization.

Thus, the main problem in adopting a randomized complete blocks design is not in the validity of the analysis (although one might challenge the relevance of the conceptual framework) but in its lack of sensitivity. A graph of estimated plot effects against their locations in the experiment will almost inevitably display substantial spatial structure; to ignore this is extremely wasteful. The classical remedy has been to develop much more sophisticated designs that employ a local blocking structure within which it can be more reasonably assumed that fertility is effectively constant. The most efficient designs, such as completely balanced lattice squares, are rarely practicable because of restrictions on the number of treatments and replicates. This difficulty has been met by the introduction of compromise designs, based on partially balanced incomplete blocks, and these are now used quite widely, especially in variety trials (Patterson and Hunter, 1983). However, a new problem arises with sophisticated designs, for there no longer exists a proper justification for the use of Gaussian-theory confidence intervals, unless additional or different assumptions are made in the statistical formulation. Furthermore, despite the obvious merits of sophisticated designs, a large proportion of experiments in the world employs nothing more complicated than randomized complete blocks, whether for reasons of tradition or ease of management. Thus it is important that methods of analysis be available that adopt explicit spatial models for fertility, if only as a means of salvaging badly designed experiments (Bartlett, 1978, 1988). Of course, it

would be unrealistic to expect a spatial model to provide more than a crude representation of a true fertility process, but this is probably unimportant, since (1) it is the replicated treatment effects rather than the individual plot fertilities that are of primary concern and (2) the purpose of the model is essentially one of interpolation rather than extrapolation. There is an interesting contrast here with the usual requirements in time series analysis.

In fact, the idea of extracting information from neighboring experimental plots as a means of controlling for variation is not at all new and was first proposed in the 1920s by J. S. Papadakis, a distinguished Greek agronomist. Unfortunately, his entirely empirical approach received very little attention by others; for some historical reflections, see Bartlett (1988). Nevertheless, Papadakis himself continued to use and develop his method over several decades (see Papadakis, 1984), and it is instructive to consider one particular version below.

Thus, let y denote the vector of observed yields, with plots indexed in some convenient manner, and suppose

$$y = T\tau + x + z\,, \tag{6.1}$$

where τ denotes treatment effects, T is the corresponding full-rank treatment design matrix, x represents the (fixed) fertility effects measured about zero, and z is residual error. If x^* denotes a current assessment of x and is presumed to be correct, the corresponding ordinary least-squares estimate of τ is

$$\tau^*(x^*) = (T^TT)^{-1}T^T(y - x^*)\,. \tag{6.2}$$

This provides a reassessment $y - T\tau^*$ of x but leads to circularity in the absence of some form of constraints on the parameter space. Papadakis resolved this difficulty by using as the new estimate of x

$$x^*(\tau^*) = H(y - T\tau^*)\,, \tag{6.3}$$

where H is a matrix that reflects anticipated similarity in fertility between neighboring plots. For example, in a two-dimensional layout, the fertility in any particular plot might be estimated by the average of the residuals in the four adjacent plots, with an appropriate modification at the boundary of the experiment. Papadakis initiated (6.2) with $x^* = 0$ and then iterated between (6.2) and (6.3), either for a prescribed number of cycles or until convergence. Here we concentrate exclusively on the latter option, referred to as the *iterated Papadakis procedure*. It follows that the final estimate τ^*

satisfies

$$\tau^* = (T^T T)^{-1} T^T [y - H(y - T\tau^*)],$$

so that

$$T^T(I - H)T\tau^* = T^T(I - H)y. \tag{6.4}$$

Thus an alternative viewpoint is that τ^* is the generalized least-squares estimate of τ when z is negligible and x is interpreted as a realization from a spatial stochastic process with (possibly generalized) inverse covariance matrix proportional to $I - H$, assuming this is symmetric positive (semi-) definite.

The implication is that Papadakis' empirical procedure may have a separate interpretation as a formal model-based approach to fertility adjustment. We investigate this and consider generalizations in §6.3. Of course, at this stage, there is no guarantee that τ^* has any particular merit, or that the $I - H$ induced by typical Papadakis adjustment holds appeal as inverse covariance matrices in a random field formulation. Finally, note that, where in our discussion $I - H$ is singular, estimates of treatment contrasts rather than τ^* itself will be uniquely determined.

6.3 Some Recent Progress and Future Directions

6.3.1 Aims

It has already been noted that, in most field experiments, plots are long and narrow. It follows that, even when the layout itself is two-dimensional, internal control for fertility variation is usually profitable only in the direction of the shorter plot axis. Thus, in §6.3.2, we concentrate on one-dimensional adjustment. In particular, we first discuss the role of simple stochastic models in experiments that only involve a single linear array (column) of plots; the results extend immediately to trials that in effect employ several separate columns. Then, in §6.3.3, we tackle the less common but nevertheless important situation in which genuine two-dimensional adjustment is necessary; this is a topic that requires considerable further research. Finally, in §6.3.4, we briefly discuss some other approaches and some outstanding problems.

In §§6.3.2 and 6.3.3, we assume a formulation that is in accordance with equation (6.1), where now y, x, and z are interpreted as realizations of spatial stochastic processes Y, X, and Z; thus,

$$Y = T\tau + X + Z, \tag{6.5}$$

where Y is the vector of random plot yields, τ represents fixed treatment effects, and T is the design matrix for the experiment. We further suppose that the components of Z are uncorrelated, with zero means and common variance ω, and that Z and X are uncorrelated; Z takes account of residual errors and is often negligible in practice when compared with the variation in the fertility process X.

6.3.2 One-Dimensional Adjustment

In discussing specification of the fertility process X for layouts that consist of a single column of n plots, it proves convenient initially to consider an ostensibly infinite column, with plots labelled $i = 0, \pm 1, \ldots$, according to their positions with respect to a reference plot 0. Let X_i denote the random fertility in plot i, measured about zero. The simplest specification of the X_i's that departs from independence is the classical first-order stationary autoregression in which the lag k autocorrelation is $\rho_k = \lambda^{|k|}, k = 0, \pm 1, \ldots$, where $|\lambda| < 1$. The corresponding autocorrelation generating function is

$$C(u) \equiv \sum_{k=-\infty}^{\infty} \rho_k u^k = \frac{1 - \lambda^2}{1 + \lambda^2 - \lambda(u + u^{-1})}. \tag{6.6}$$

In the present context, the above unilateral model is more naturally formulated as a bilateral autoregression, with

$$\begin{aligned} E(X_i | \text{ all } x_j, j \neq i) &= \beta(x_{i-1} + x_{i+1}), \\ \text{var}(X_i | \text{ all } x_j, j \neq i) &= \kappa, \end{aligned} \tag{6.7}$$

where $\beta = \lambda/(1 + \lambda^2)$ and $|\beta| < \frac{1}{2}$. The equivalence is confirmed by noting that (6.7) implies that the corresponding autocorrelations ρ_k satisfy

$$\rho_k = \beta(\rho_{k-1} + \rho_{k+1}), \quad k = \pm 1, \pm 2, \ldots, \tag{6.8}$$

and hence that $\rho_k = \lambda^{|k|}$. The duality between the unilateral and bilateral formulations does not generally extend to higher dimensions and rests on the factorization $h(u)h(u^{-1})$ of the denominator in (6.6), where $h(u) = 1 - \lambda u$; see §6.3.4 for some further comments.

Of course, in reality, neither X nor Z is observed but only Y over a finite column of plots, $i = 1, 2, \ldots, n$, say. If $\alpha = \omega/2\kappa$ and β were known, then estimation of τ could proceed by generalized least squares. Otherwise,

several methods of estimating ω, κ, and β are available. For example, one might assume additionally that X and Z are Gaussian and apply standard maximum-likelihood estimation (*cf.* Tiao and Ali, 1971, in a different context) or the residual maximum likelihood (REML) variation (Patterson and Thompson, 1971). What is important here is that in practice it is common that the estimate of α is zero or close to zero and that of β is very close to its maximum possible value, $\frac{1}{2}$. It is therefore instructive to consider both $\alpha = 0$ and $\beta \uparrow \frac{1}{2}$ in more detail. In each case, we again begin with an infinite line of plots.

First suppose that $\alpha = 0$ with β known and let H denote the doubly-infinite matrix with (i, j) element

$$H_{ij} = \begin{cases} \beta & j = i \pm 1 \\ 0 & \text{otherwise} \end{cases} .$$

Then it is easily checked that the inverse covariance matrix (i.e., precision matrix) for Y is proportional to $I - H$. It follows that the generalized least squares estimate τ^* of τ agrees with the iterated Papadakis estimate in (6.4) for which fertility in plot i is estimated at each stage by $\beta \times$ {sum of the residuals in the two adjacent plots}. This provides a useful connection between the present model-based approach and that espoused empirically by Papadakis. Furthermore, the duality between bilateral modeling and Papadakis' method is perfectly general, provided H in (6.3) is symmetric positive (semi-) definite, and extends not only to more complicated one-dimensional adjustment but also to higher dimensions (*cf.* §6.3.3). Of course, for real experiments with finite numbers of plots, it holds only as an approximation because of edge effects.

Second, suppose that $\beta \uparrow \frac{1}{2}$ for fixed α. Though the distribution of X itself degenerates, the differences $X_i - X_j$ remain well behaved with zero means and with variances

$$\operatorname{var}(X_i - X_j) = \frac{2\kappa(1 + \lambda^2)(1 - \lambda^{|i-j|})}{1 - \lambda^2} \to 2\kappa|i - j|$$

as $\beta \uparrow \frac{1}{2}$. Equivalently, first differences $X_i - X_{i+1}$ are, in the limit, uncorrelated and have equal variance 2κ, so that X can be thought of as a random walk. It is also of interest that, in the limit $\beta = \frac{1}{2}$, the conditional expectation formulation (6.7) remains valid, even though the marginal mean of X_i is undefined and the marginal variance is infinite. This is the most

basic example of an intrinsic autoregression (Künsch, 1987) and provides a stochastic version of simple linear interpolation; see also §6.3.3. The degeneracy is unimportant in practice because it is only comparisons of treatment effects that are being assessed. This becomes apparent algebraically if we return to the actual experiment. Thus, let $U_i = Y_i - Y_{i+1}$, $i = 1, \ldots, n-1$, or, in vector notation, $U = \Delta Y$, where Δ is the $n-1$ by n matrix taking first differences of the Y_i's. It is convenient at this point to single out the overall level γ of the experiment. Suppose τ is a p-vector; then perhaps after reparameterization, we can write

$$T\tau = \gamma 1 + D\delta\,, \tag{6.9}$$

where 1 is an n-vector of 1's, δ is a $p-1$ vector of (relative) treatment effects, and D is an n by $p-1$ design matrix of rank $p-1$. Since any treatment contrast $\phi = a^T\tau$, where $a^T 1 = 0$, can be written in terms of δ, it is sufficient to concentrate on the estimation of δ. The mean and variance-covariance structures of U are given by

$$E(U) = F\delta\,, \qquad D(U) = 2\kappa Q\,, \tag{6.10}$$

where $F = \Delta D, Q = I + \alpha\Delta\Delta^T$, and I is now the $(n-1)\times(n-1)$ identity. Note the absence of end-plot problems and the retention of information on treatment contrasts in the reduction of the data to $n-1$ first differences.

Finally, let δ^*_α denote the generalized least-squares estimate of δ, so that

$$\delta^*_\alpha = (F^T Q^{-1} F)^{-1} F^T Q^{-1} u\,, \tag{6.11}$$

where u is the observed value of U.

Two special cases of (6.11) merit attention. At one extreme, δ^*_0 is the ordinary least-squares estimate of δ based on u; at the other, δ^*_∞ is the ordinary least-squares estimate of δ based on y (*cf.* Besag and Kempton, 1986). Thus, for any intermediate value of α, δ^*_α provides a compromise between the ordinary least-squares estimates of δ based on u and on y, respectively; this resembles the combination of intra- and interblock information in the classical analysis of incomplete block designs but here using the notion of a moving block of size two.

The above discussion indicates that there is little point in retaining a flexible value of β and that $\beta = \frac{1}{2}$ should be adopted from the outset. In practice, it is still often found that the estimate of α is essentially zero,

TABLE 6.1: Layout and Yields (t/ha) for 62 Varieties of Winter Wheat

Replicate 1		Replicate 2		Replicate 3	
Variety	Yield	Variety	Yield	Variety	Yield
10	7.32	58	7.52	32	7.29
50	6.34	31	7.50	14	6.70
18	7.44	40	6.61	23	7.57
58	8.54	41	7.00	51	7.38
42	7.26	4	5.59	33	7.71
26	7.50	22	6.01	2	6.78
2	6.92	13	7.52	60	8.44
34	8.46	51	7.07	45	7.61
56	7.22	27	7.63	48	6.93
32	8.22	18	7.28	9	7.23
16	7.15	8	6.32	36	6.76
8	6.90	45	7.43	5	6.19
24	7.48	55	7.31	27	7.86
48	7.02	62	8.36	18	7.82
40	8.16	36	7.04	54	6.69
41	7.92	9	7.23	34	8.35
9	7.56	56	8.16	25	8.35
33	8.57	28	6.86	24	7.69
49	8.57	46	8.21	52	8.25
1	7.35	19	8.54	3	7.77
25	8.85	10	7.56	15	4.38
57	8.06	37	7.41	61	8.31
17	8.10	1	7.35	46	8.45
45	8.64	44	8.50	11	8.30
61	8.41	26	9.01	42	7.43
13	8.95	17	8.60	7	9.51
37	7.39	54	7.25	38	6.87
5	7.30	61	8.39	20	8.69
29	8.31	35	8.45	57	7.69
53	8.87	16	7.70	56	7.84
21	7.83	7	9.46	29	7.96
20	8.45	48	7.28	6	7.75
36	7.07	57	7.71	19	7.98
52	8.00	3	7.46	55	7.82
44	8.28	30	7.63	28	6.58
4	7.48	50	6.32	37	7.51
60	8.69	21	7.29	10	7.31
28	6.88	12	8.29	41	7.90
12	8.17	39	7.09	35	8.60
62	7.99	49	8.30	62	8.22
14	7.48	2	7.49	16	7.55
54	7.17	20	8.94	53	8.52
22	7.67	11	8.09	26	8.58
6	7.60	29	7.99	47	7.15
38	7.40	47	8.05	4	8.37
46	8.55	38	7.38	17	7.89
30	7.50	32	9.00	59	9.16
39	7.75	52	9.24	13	9.04
15	4.82	59	9.60	1	7.72
55	8.40	33	9.13	40	8.15
31	9.02	42	8.20	31	8.98
47	7.57	14	7.90	50	7.10
23	9.12	23	9.26	44	8.20
7	9.96	5	7.80	22	8.20
11	8.51	15	6.28	58	8.92
19	8.55	43	8.95	39	7.68
59	9.14	53	8.96	8	6.78
3	7.70	25	9.32	21	7.67
43	8.43	24	8.24	49	8.57
27	7.98	34	9.15	12	8.14
51	7.66	60	9.40	43	7.95
35	8.24	6	8.17	30	7.67

especially if due allowance for outliers and jumps in fertility is made (*cf.* §6.3.4); this leads to a very simple estimation procedure. Although design is not a consideration here, it is worth noting that if plots 1 and n contain the same treatment, δ_0^* is a linear combination of the $n-2$ *second* differences $y_{i-1} - 2y_i + y_{i+1}$ and is therefore invariant to linear trends in fertility and approximately so to *locally* linear trends (which are of greater concern). Incidentally, it also turns out that δ_0^* is featured in several other proposals for fertility adjustment and provides an agreeable unity between ostensibly different approaches; again see §6.3.4.

The general analysis with $\beta = \frac{1}{2}$ extends immediately to several (effectively) independent columns of plots. The single parameter γ is replaced by a vector of separate column effects and, subject to the usual necessity of such terms, there is again no loss of information on treatment contrasts by the reduction of the data to first differences within columns. For further details, including determination of standard errors, the analysis of a factorial experiment on triticale, and an assessment of accuracy and precision, see Besag and Kempton (1986).

Here we illustrate first-differences analysis on an official United Kingdom trial for winter wheat, carried out by the East of Scotland College of Agriculture and involving final assessment of 62 different varieties. The layout of the experiment, in three physically separated complete replicates, and the corresponding yields (t/ha) are listed in Table 6.1. The yields are graphed against plot position in Figure 6.1 (top); there is clear evidence of modest fertility gradients within replicates. First-differences analysis (i.e., with $\beta = \frac{1}{2}$) produces the estimate $\alpha = 1.76$ and the decomposition of yields into relative variety effects, fertility effects, and residuals shown in the bottom three panels of Figure 6.1. The standard errors of pairwise differences between varieties range from .202 to .238 with a mean of .223, compared with the value .418 for a complete block analysis. Thus, there is substantial improvement in precision and presumably in accuracy of variety estimates; for objective methods of assessment, see the penultimate paragraph of §6.3.4. Moreover, fertility gradients are often more pronounced than in this particular experiment (see, for example, the associated trials analyzed by Green *et al.*, 1985), particularly in countries that do not have the temperate climate of the United Kingdom.

In fact, because of the importance of the above trial, the within-replicate layout conformed to a generalized lattice design. This confers no particular benefit to the first-differences analysis, except perhaps to reduce the range

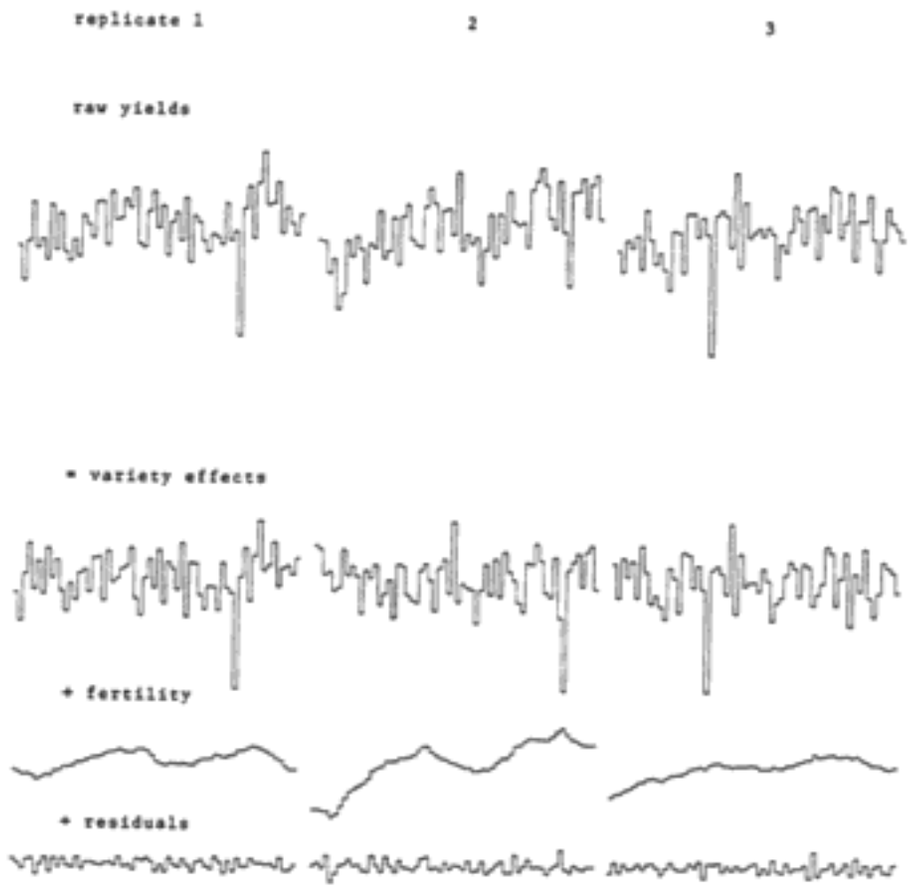


FIGURE 6.1: Raw yields and subsequent first-differences decomposition for 62 varieties of winter wheat.

of the standard errors, but enables classical incomplete-blocks analysis to be carried out. The corresponding standard error of a varietal difference, kindly supplied by Professor H. D. Patterson, is .235. The fairly close agreement with spatial statistical analysis seems typical but of course the latter does not require a sophisticated design and applies equally to the simple layouts encountered more commonly in practice.

6.3.3 Two-Dimensional Formulation

In this section, we seek to generalize the previous one-dimensional formulation. Thus, we again adopt equation (6.5) but now identify plots by integer

pairs of Cartesian coordinates $i = (i_1, i_2)$. Practical aspects are not developed to the same extent in two dimensions, the main problem being that edge effects are much more important than in one dimension (Guyon, 1982; Dahlhaus and Künsch, 1987) and must not be ignored, although their effects are sometimes overemphasized.

The generalization of equation (6.7) to a two-dimensional model is given by

$$\begin{aligned} E(X_i|\text{ all } x_j, j \neq i) &= \beta_1(x_{i_1-1,i_2} + x_{i_1+1,i_2}) + \beta_2(x_{i_1,i_2-1} + x_{i_1,i_2+1}), \\ \text{var}(X_i|\text{ all } x_j, j \neq i) &= \kappa, \end{aligned} \tag{6.12}$$

where $|\beta_1| + |\beta_2| < \frac{1}{2}$; we assume neither β_1 nor β_2 is zero. It follows that the autocorrelations ρ_k satisfy

$$\rho_k = \beta_1(\rho_{k_1-1,k_2} + \rho_{k_1+1,k_2}) + \beta_2(\rho_{k_1,k_2-1} + \rho_{k_1,k_2+1}) \tag{6.13}$$

for $k \neq (0,0)$ and that the corresponding generating function is

$$C(u) \equiv \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \rho_k u_1^{k_1} u_2^{k_2} \propto [1-\beta_1(u_1+u_1^{-1})-\beta_2(u_2+u_2^{-1})]^{-1}, \tag{6.14}$$

which cannot be reproduced by any finite unilateral autoregression. Formulae exist for the low-order autocorrelations but generally the ρ_k's are best calculated by recursive algorithms (these can be quite delicate) or approximated using Bessel functions; for details, see Besag (1981). Equation (6.12) can be easily extended to include more distant plot fertilities, with appropriate modification to (6.13) and (6.14); see Rosanov (1967) or Besag (1974), though (6.12) itself dates back to Lévy (1948). If Z is negligible (i.e., $\alpha = 0$), the equivalent iterated Papadakis adjustment (6.3) for any particular bilateral autoregression can be written down immediately.

Given a partial realization of X in (6.12) over a finite region, the parameters β_1, β_2, and κ can be estimated by matching the theoretical variance and first-order autocovariances with the edge-corrected (i.e., unbiased) empirical values (Besag, 1974; Guyon, 1982). This generalizes to arbitrary bilateral autoregressions and, if X is Gaussian, is equivalent to asymptotic maximum-likelihood estimation. However, here we are concerned with observation on Y in (6.5) rather than on X. The incorporation of treatment effects τ is straightforward but that of random error Z is more problematical, primarily because of edge effects. When Z is ignored, it is generally found that the estimate of $\beta_1 + \beta_2$ is very close to $\frac{1}{2}$. One means of including Z is to make

toroidal assumptions, identifying opposite edges of the field (Besag, 1977). Although this device is of minimal direct practical relevance, one might reasonably expect that it should depress the estimates of β_1 and β_2, and yet the above behavior seems to be reproduced. Thus, we might well abandon the stationarity assumption and rather adopt (6.12) with $\beta_1 + \beta_2 = \frac{1}{2}$; that is, on the infinite lattice, an intrinsic bilateral autoregression of class zero (Künsch, 1987). Such a formulation is of course entirely consistent with the one-dimensional development in §6.3.2; again, we may expect that Z will often be negligible, which if assumed from the start would lead to entirely straightforward estimation, although we do not wish to exclude the possibility of a non-zero α.

Certainly the problems of estimation are not insuperable but they require further research, especially as regards standard errors for treatment contrasts; these must retain approximate validity somewhat outside the narrow confines of the model itself. We briefly consider the assessment of different methods in §6.3.4, but here we conclude with some remarks concerning the structure of intrinsic autoregressions. Thus, suppose $\beta_1 = \beta$ and $\beta_2 = \frac{1}{2} - \beta$ in (6.12). In the absence of stationarity, we need a new measure of covariation and the obvious choice is the *semi-variogram* (of chapter 5), defined by

$$v_k = \frac{1}{2}\text{var}(X_i - X_{i+k}), \quad i, k \in Z^2 .$$

It follows that $v_0 = 0$ and, from (6.12),

$$v_k = -\kappa\epsilon_k + \beta(v_{k_1-1,k_2} + v_{k_1+1,k_2}) + (\frac{1}{2} - \beta)(v_{k_1,k_2-1} + v_{k_1,k_2+1}) ,$$

where $\epsilon_k = 1$ if $k = (0,0)$ and is otherwise zero. It can be shown (Künsch,1987) that the asymptotic growth of the semi-variogram is logarithmic. Although explicit results are not generally available, it may be noted that, when $\beta = \frac{1}{4}$, $v_{1,0} = v_{0,1} = \kappa$,

$$v_{k,k} = \frac{4k}{\pi}\left(1 + \frac{1}{3} + \frac{1}{5} + \ldots + \frac{1}{2k-1}\right) , \quad k = 1, 2, \ldots ,$$

(*cf.* Besag, 1981), and the semi-variogram can be easily evaluated for all lags. On the other hand, $\beta = \frac{1}{2}$ of course reverts to the model of §6.3.2 for independent columns of plots. At first sight, it might appear that the value of β should simply be determined by plot shape, assuming isotropy of the underlying fertility process, but in practice, fertility patterns are also induced by the management of the experiment.

6.3.4 Other Approaches and Further Research

Several other methods of analysis for agricultural field experiments, adopting explicit spatial assumptions, have been proposed in the literature; see, for example, Wilkinson *at al.* (1983), Green *et al.* (1985), Williams (1986), Gleason and Cullis (1987), and Martin (1989). Here we briefly consider two, the first based on a time-series formulation, the second on a data-analytic approach.

We have already noted, in §6.3.1, the equivalence between first-order unilateral and bilateral autoregressions in one dimension. This extends to models of arbitrary order. Thus, for trials that only require one-dimensional adjustment, Martin (1989) proposes that classical time-series methodology should be used to select and fit an appropriate model in advance. He then extends this approach to two-dimensional adjustment by considering only the class of processes that, after row or column differencing, are *separable*; that is, are stationary and have interplot autocorrelations ρ_k satisfying

$$\rho_{k_1,k_2} = \rho_{k_1,0}\,\rho_{0,k_2}\,. \tag{6.15}$$

Separability leads to a considerable simplification in the computation of parameter estimates, though the advantage is diminished with the inclusion of superimposed random error; see Martin (1989) for details. At first sight, the Manhattan metric of (6.15) is unappealing but could be appropriate when fertility patterns are largely the product of cultivation practice and may in any case provide an adequate approximation. Model selection based on very limited data is perhaps the major handicap of the approach, though this aspect could be abandoned. As with other methods, there is a need for further research, including practical investigation.

For a data-analytic viewpoint, we consider the approach proposed by Green *et al.* (1985); this also supplies further insight into Papadakis' and most other methods. Equation (6.1) again provides the starting point. However, equations (6.2) and (6.3) are generalized to

$$\begin{aligned} \tau^\bullet(x^\bullet) &= \mathcal{T}(y - x^\bullet) \\ x^\bullet(\tau^\bullet) &= \mathcal{S}(y - T\tau^\bullet)\,, \end{aligned} \tag{6.16}$$

where $\mathcal{S}$ and $\mathcal{T}$ may be linear or nonlinear operators; $\mathcal{S}$ is a *smoother* of fertility and other extraneous variation, whilst $\mathcal{T}$ allows (e.g., robust/resistant) alternatives to be substituted for the ordinary least-squares estimate (6.2). Here we concentrate on adjustment along a single column of n plots, in which case the basic choices of $\mathcal{S}$ and $\mathcal{T}$ are made as follows. First it is assumed that fertility variation is approximately locally linear, so that second

differences $\Delta_2 x$ are small, where Δ_2 is the $n-2$ by n matrix taking second differences. Then x and τ are estimated by *least-squares smoothing*; that is, x^* and τ^* minimize

$$\alpha x^T \Delta_2^T \Delta_2 x + z^T z\,,$$

with the effect that smoothness of the fitted fertility pattern is offset against the residual variation, according to the value of the "tuning constant" α. Hence x^* and τ^* satisfy (6.16) with T as in (6.2) and

$$\mathcal{S} = (I + \alpha \Delta_2^T \Delta_2)^{-1}\,. \tag{6.17}$$

As might be anticipated, (6.16) determines estimates of treatment contrasts rather than τ^* itself; see below. Green *et al.* (1985) suggest several data-analytic prescriptions for the choice of α, including cross validation, and illustrate their methodology on data from three different trials. The analyses also include approximate standard errors for treatment contrasts and graphs of estimated treatment, fertility, and residual effects across each of the experimental areas. As with other methods of fertility adjustment that involve a tuning parameter, the exact value of α does not seem to be critical.

It is of interest that an alternative derivation of (6.17) is available through the random field formulation (6.5), with the assumption that second differences in X are uncorrelated and have equal variance 2κ. The generalized least-squares estimate of δ in (6.9), based on second differences of the y_i's, is then given by (6.11) but with Δ replaced by Δ_2 in the definitions of F, Q, and U. The equivalence follows since (6.17) implies that

$$\Delta_2 (I - \mathcal{S})^+ \Delta_2^T = \alpha^{-1} Q\,, \tag{6.18}$$

where A^+ denotes any generalized inverse of A. Note that, since (6.18) also holds if Δ_2 is replaced by Δ throughout, the above argument can be inverted to provide a least-squares smoothing interpretation of the first-differences analysis in §6.3.2. In fact, the generalized equations (6.16) are of very wide applicability. For example, they include, on the one hand, the estimates obtained from a classical analysis of an incomplete block design and, on the other, those obtained from the "NN" methodology of Wilkinson *et al.* (1983); for a comprehensive discussion, see Green (1985). Incidentally, NN analysis over a finite region provides an example where a random field formulation is not strictly available: $\mathcal{S}$, though linear, is not completely symmetric because of border plots. However, there are close asymptotic links between NN analysis and that based on first differences.

A very important aspect of any model-based statistical procedure is its robustness to departures from the underlying assumptions. Our initial optimism concerning the adequacy of a fairly crude fertility model is supported

in practice by the general similarity between treatment estimates obtained from different spatial formulations and by close agreement with classical results when sophisticated design and analysis, such as balanced lattice squares, has been used. Furthermore, there is frequent disparity between spatial and conventional estimates when a poor design, such as randomized complete blocks, has been employed, so that fertility adjustment seems to be worthwhile. We briefly discuss quantitative assessment below but it may also be desirable to modify a model-based procedure to accommodate gross anomalies, particularly those caused by measurement errors or by abrupt jumps in fertility, which may be the product of a change in underlying geological structure, for example. Papadakis (1984) reviews his previous work on this topic and Besag and Seheult (1989) summarize a closely related approach geared to first-differences analysis. In the context of (6.16), the two types of anomaly might be catered for by nonlinear resistant versions of $\mathcal{T}$ and $\mathcal{S}$ respectively.

How can quantitative assessments be made? Perhaps the only rigorous method is to use data from uniformity trials in which all plots receive common treatment. If a mock design is superimposed on such a trial, any particular procedure can be used to estimate the relative "treatment" effects and, since these are known to be zero, an assessment of accuracy can be made. Furthermore, the results are relevant to a real experiment under the usual assumption of treatment additivity, provided the method of analysis also acts additively. Predicted standard errors can also be compared with actual variability of estimates. Unfortunately, in a random field framework, each trial provides but a single assessment and many sets of uniformity data are required for a proper evaluation. Moreover, uniformity trials are rarely carried out these days, though, for an early catalog, see Cochran (1937). An alternative is to carry out an assessment within a randomization framework (e.g., Besag and Kempton, 1986, Appendix 2), although of course, this addresses a population for which the procedure was not designed.

Finally, what of Bayesian formulations? They are indeed conspicuous by their virtual absence from the literature. The main difficulty is that of representing in probabilistic terms one's prior beliefs about fertility variation. Thus, considerations are likely to be very similar to those arising in a random field formulation (6.5) but inferential aspects may be more akin to recent developments in Bayesian image analysis.

Bibliography

[1] Bartlett, M. S., Nearest neighbour models in the analysis of field experiments (with discussion), *J. R. Stat. Soc., B* **40** (1978), 147–174.

[2] Bartlett, M. S., Stochastic models and field trials, *J. Appl. Prob.* **25A** (1988), 79–89.

[3] Beaven, E. S., Pedigree seed corn, *J. R. Agric. Soc.* **70** (1909), 119–139.

[4] Besag, J. E., Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Stat. Soc., B* **36** (1974), 192–236.

[5] Besag, J. E., Errors-in-variables estimation for Gaussian lattice schemes, *J. R. Stat. Soc., B* **39** (1977), 73–78.

[6] Besag, J. E., On a system of two-dimensional recurrence equations, *J. R. Stat. Soc., B* **43** (1981), 302–309.

[7] Besag, J. E., and R. A. Kempton, Statistical analysis of field experiments using neighbouring plots, *Biometrics* **42** (1986), 231–251.

[8] Besag, J. E., and A. H. Seheult, Contribution to discussion of Bruce and Martin, *J. R. Stat. Soc., B* **51** (1989), 405–406.

[9] Cochran, W. G., A catalogue of uniformity trial data, *J. R. Stat. Soc. (Suppl.)* **4** (1937), 233–253.

[10] Dahlaus, R., and H. Künsch, Edge effects and efficient parameter estimation for stationary random fields, *Biometrika* **74** (1987), 877–882.

[11] Gleason, A. C., and B. R. Cullis, Residual maximum likelihood (REML) estimation of a neighbour model for field experiments, *Biometrics* **43** (1987), 277–287.

[12] Green, P. J., Linear models for field trials, smoothing and cross-validation, *Biometrika* **72** (1985), 527–537.

[13] Green, P. J., C. Jennison, and A. H. Seheult, Analysis of field experiments by least-squares smoothing, *J. R. Stat. Soc., B* **47** (1985), 299–315.

[14] Guyon, X., Parameter estimation for a stationary process on a d-dimensional lattice, *Biometrika* **69** (1982), 95–105.

[15] Kempthorne, O., *The Design and Analysis of Experiments*, John Wiley and Sons, New York, 1952.

[16] Künsch, H., Intrinisic autoregressions and related models on the two-dimensional lattice, *Biometrika* **74** (1987), 517–524.

[17] Lévy, P., Chaînes doubles de Markoff et fonctions aléatoires de deux variables, *C. R. Acad. Sci., Paris* **226** (1948), 53–55.

[18] Martin, R. J., The use of time-series models and methods in the analysis of agriculture field trials, *Comm. Stat. Theor. Meth.* **19** (1989), 55–81.

[19] Papadakis, J. S., Advances in the analysis of field experiments, *Proceeds. Acad. Athens* **59** (1984), 326–342.

[20] Patterson, H. D., and E. A. Hunter, The efficiency of incomplete block designs in National List and Recommended List cereal variety trials, *J. Agric. Sci.* **101** (1983), 427–433.

[21] Patterson, H. D., and R. Thompson, Recovery of inter-block information when block sizes are unequal, *Biometrika* **58** (1971), 545–554.

[22] Rosanov, Yu. A., On the Gaussian homogeneous fields with given conditional distributions, *Theor. Probability Appl.* **12** (1967), 381–391.

[23] Tiao, G. C., and M. M. Ali, Analysis of correlated random effects: Linear model with two random components, *Biometrika* **58** (1971), 37–51.

[24] Wiancko, A. T., Use and management of check plots in soil fertility investigations, *J. Am. Soc. Agron.* **13** (1914), 368–374.

[25] Wilkinson, G. N., S. R. Eckert, T. W. Hancock, and O. Mayo, Nearest-neighbour (NN) analysis of field experiments (with discussion), *J. R. Stat. Soc., B* **45** (1983), 151–211.

[26] Williams, E. R., A neighbour model for field experiments, *Biometrika* **73** (1986), 279–287.

7

Spatial Statistics in Ecology

Peter Guttorp
University of Washington

7.1 Introduction

Ecological theory is essentially spatial in character. Many methods for analyzing spatial data have been developed in an ecological context (Hertz, 1909; Greig-Smith, 1952; and Kershaw, 1957, are some important early references). Methods from spatial statistics have recently seen an increasing use in this field. Perhaps the most important data for quantitatively oriented plant ecologists are complete maps of the vegetation in an area at different times. While the construction of such maps used to be an incredibly time-consuming fieldwork task, modern digitization techniques enable an increased use of aerial photographs and satellite images. Here, as in many other fields, there has recently been a substantial increase in both the quantity and volume of data potentially available to the ecological modeler. Some overviews of the use of spatial methods in ecological analysis are Ripley (1987) and Legendre and Fortin (1989).

Typically, a large number of factors interact in ecological processes, and the precise nature of these interactions is the subject of study. For example, in the study of forest growth, a limiting factor is availability of light (Ford and Diggle, 1981). The death of a large tree yields sudden possibilities for growth of plants that would otherwise remain very small, and can completely change the competitive advantages between species. The introduction of a new species may eliminate many previously successful competitors (Ford, 1975, Linhart, 1976). In order to evaluate forest resource management plans, it may prove important to develop adequate stochastic models for species growth and competition. The interactions take place at different scales: the extent of a tree crown limits the availability of light, decreasing the potential

for other growth beneath the crown, whereas the availability of nutrients in the local region can increase growth potential on a somewhat larger scale.

In this chapter, we concentrate on one approach to stochastic modeling of ecological communities, namely, spatial point processes. Models for animal communities often need to include movement explicitly. The theory of branching diffusions (Dawson and Ivanoff, 1978, Kulperger, 1979) can sometimes be applied to such situations. There is a plethora of predator-prey models in the applied probability literature, although so far most of them are not specifically spatial in nature. There is a need for more work on spatially nonhomogeneous competition models.

Section 7.2 introduces the general concepts of point processes, discusses nonparametric estimation of second order parameters, and presents some particular models that have found use in the literature. Section 7.3 contains an outline of a point process approach to modeling single species forest growth. It must be emphasized here that the efforts to date of using stochastic models (in particular point process models) and their attendant statistical analysis to aid ecological understanding has had only very limited success. This is due partly to oversimplifications (such as using only homogeneous models or studying only one species rather than the interactions of several), partly to lack of high-quality data, and partly to the difficulty in interpreting interactions at vastly different scales. More work is also needed on how to combine inference from the individual pieces that together make up a model of a complex system.

7.2 Point Processes

A point process is a process of locations of events, taking place in some space $\mathcal{X}$. Each event may have associated with it a mark, taking place in some mark space $\mathcal{Y}$. For example, an event may be a tree, and the mark may be the species of the tree, its crown length, crown angle, height, and diameter. An excellent description of point process theory is Daley and Vere-Jones (1988, especially ch. 7). The random variable $N(A)$ counts the number of events in the set $A \subset \mathcal{X}$. A marked point process is a point process on $\mathcal{X} \times \mathcal{Y}$ with the additional property that the marginal process of locations $N(A \times \mathcal{Y})$, $A \subset \mathcal{X}$ is itself a point process.

A case of particular interest is a multivariate point process, where $\mathcal{Y} = \{1, \ldots, m\}$ for some finite integer m. Harkness and Isham (1983) study a bivariate point process (i.e., $m = 2$) of ant nests for the species *Cataglyphis*

bicolor and *Messor wasmanni*. Their main interest is in assessing whether the locations of *Cataglyphis* nests are dependent upon those of the *Messor* ants. This is suggested on biological grounds, since *Cataglyphis* ants eat dead insects, mainly *Messor* ants, whereas the latter collect seeds for food. An example of a trivariate point process is the data collected by Gerrard (1969) and analyzed by Besag (1977), Diggle (1983, sec. 7.1), and others, which contains locations of hickory, oak, and maple trees in Lansing Woods, Michigan. Of interest here is the interactions between the species. We return to these examples below.

An important class of point processes consists of those whose distribution is invariant under translations; these are called stationary or homogeneous. Those in the subclass of isotropic processes have distributions that additionally are invariant under rotation. The assumptions of homogeneity and isotropy are perhaps made more often than the various applications warrant.

Time series analysis has benefitted much from studying second order properties. These can be estimated nonparametrically, and for a Gaussian series completely determine the probabilistic structure of the series. But even in non-Gaussian cases, second-order parameter functions such as the spectrum or the correlogram convey interesting information. In the case of point processes, second-order parameters are perhaps less informative (Baddeley and Silverman, 1984), but are still an important aspect of the analysis of a point pattern. Diggle (1983, ch. 5; see also Ripley, 1988, ch. 3) presents second-order parameter estimation, and Brillinger (1978) discusses the relation between time series and point process analysis.

In what follows, we concentrate on the spatial case where $\mathcal{X} = \mathbf{R}^2$. The second-order product density of a point process is defined by

$$\lambda_2(x, y) = \lim_{|dx|,|dy|\to 0} \frac{\mathbf{E}\, N(dx)\, N(dy)}{|dx|\,|dy|}\,.$$

For a stationary process, λ_2 depends only on the vector $x - y$, and if the process is also isotropic, it further depends only on the length $t = |x - y|$. A common variant of λ_2 for stationary isotropic processes is (Ripley, 1976)

$$K(t) = \int_0^{2\pi} \int_0^t \frac{\lambda_2(x)}{\lambda} x\, dx\, d\theta\,,$$

where λ is the rate of the process. The parameter function $K(t)$ measures the expected relative rate of events within distance t of an arbitrary event. For example, in a Poisson process (§7.2.1 below) we have

$$K(t) = \pi t^2\,,$$

and for a Poisson cluster process of Neyman-Scott type (§7.2.2)

$$K(t) = \pi t^2 + \mathbf{E}\, S(S-1)H_2(t)/(\rho_c \mathbf{E}^2 S)\,,$$

where S is the number of points in a cluster, H_2 is the cdf of the vector difference between two points in the same cluster, and ρ_c is the rate of the cluster process.

Bartlett (1964) stressed the inferential importance of the distribution of nearest-neighbor distances (which is equivalent to the K-function introduced above). Ripley (1976) proposed to estimate $K(t)$ from points $x_1, \ldots, x_n$ in a set A by

$$\hat{K}(t) = \frac{|A|}{n^2} \sum_{i \neq j} \frac{1(u_{ij} < t)}{w_{ij}}\,,$$

where $u_{ij} = |x_i - x_j|$ and w_{ij} is the proportion of the area of the sphere of radius u_{ij} about x_i inside A. This nonparametric estimator can be used to fit a parametric model by minimizing the distance between the estimate and the parametric form of the function. When comparing to a Poisson process, it is a common practice to use a square root transformation to stabilize the variance of the plotted function. Harkness and Isham (1983) found that this plot for *Messor* nests (Figure 7.1) lay below the envelope for simulated values from a Poisson process for distances below 50 feet, indicating an inhibition between nests, presumably due to the foraging practices of these ants (similar findings for other ant species are reported by Lerings and Franks, 1982). On the other hand, the *Cataglyphis* nests were consistent with spatial randomness.

Further analysis indicated a tendency for *Cataglyphis* nests to be located at or near the mean foraging path for a *Messor* nest (Figure 7.2). Harrison and Gentry (1981) discussed biological and statistical aspects of foraging paths for a single species. The study area consisted of about half scrub and half field, and the *Cataglyphis* nests were located mostly in the field region. Stationarity over the entire study region did not seem to be a reasonable assumption for these nests, and Harkness and Isham separated out the field region in their study.

In the case of anisotropic stationary point processes one can estimate λ_2 directly using the obvious empirical counterpart; essentially a histogram estimator (*cf.* Brillinger, 1978; Ohser and Stoyan, 1981). Standard error for the estimators can often be developed under the Poisson hypothesis (Baddeley, 1980), whereas for more complicated processes, one may have to use Monte Carlo methods to assess the variability (see the examples in

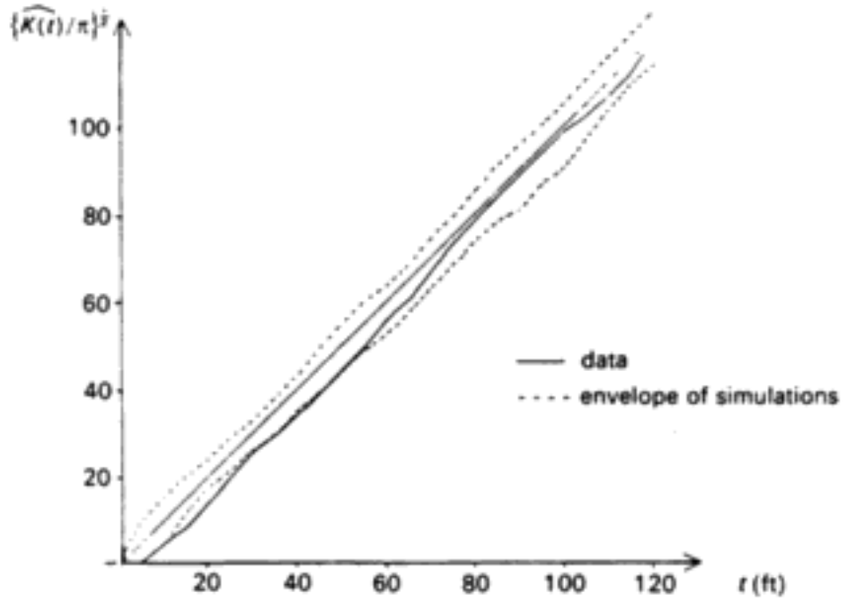


FIGURE 7.1: $\{\widehat{K(t)/\pi}\}^{\frac{1}{2}}$ for the *Messor* nests, calculated at 5-ft intervals, together with the envelope of 19 curves from simulated Poisson data. Reprinted, by permission, from Harkness and Isham (1983). Copyright © 1983 by the Royal Statistical Society.

Besag and Diggle, 1977). Bootstrap and other resampling methods have been proposed in the ecological literature by Solow (1989).

For multivariate processes a cross intensity can be defined. The corresponding K-function is

$$K_{ij}(t) = 2\pi(\lambda_i\lambda_j)^{-1}\int_0^t \lambda_{ij}(u)\,u\,du\,,$$

where λ_i is the rate of points of type i, and λ_{ij} is the cross-intensity function defined by

$$\lambda_{ij}(|x-y|) = \lim_{|dx|,|dy|\to 0} \frac{\mathbf{E}\,N_i(dx)\,N_j(dy)}{|dx|\,|dy|}\,.$$

Corresponding quantities can be defined for more general marked point processes, and their estimation is discussed by Hanisch and Stoyan (1979).

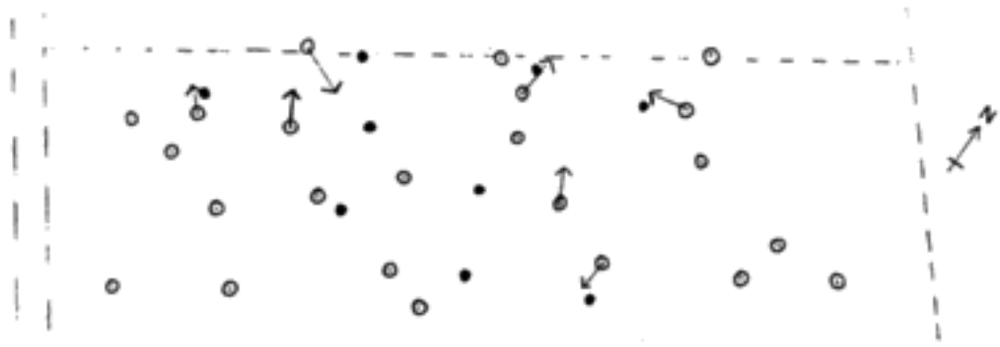


FIGURE 7.2: Mean foraging path for seven *Messor* nests. The solid dots correspond to *Cataglyphis* nests, and the open dots denote *Messor* nests. The field is between 330 and 340 ft wide. Reprinted, by permission, from Harkness and Isham (1983). Copyright © 1983 by the Royal Statistical Society.

7.2.1 The Poisson Process

The simplest model for point processes is the *completely random*, or Poisson process. To define it, assume that there is a finite measure Λ, such that for all finite families of disjoint intervals $A_1, \ldots, A_k$ we have

$$\mathbf{P}(N(A_i) = n_i,\, i = 1, \ldots, k) = \prod_{i=1}^{k} \frac{\Lambda(A_i)^{n_i}}{n_i!} \exp(-\Lambda(A_i))\,.$$

In particular, the counts in disjoint sets are independent, and hence one cannot improve the prediction of the number of points in a set from information about numbers of points in, say, surrounding sets. This is what constitutes the complete randomness of the Poisson process.

There are many equivalent ways of describing the distribution of point processes. For example, one may be able to specify the zero probability function

$$\zeta(A) = \mathbf{P}(N(A) = 0)\,,$$

or the probability generating functional (pgfl)

$$G(h) = \mathbf{E}\left(\exp\left[\int_{\mathcal{X}} \log h(x)\, N(dx)\right]\right)\,,$$

defined for all real measurable functions h with $0 \leq h \leq 1$ such that $1 - h$

vanishes outside a bounded set. The pgfl of a Poisson process is

$$G(h) = \exp\left(-\int_{\mathcal{X}}(1-h(x))\,\Lambda(dx)\right).$$

The Poisson process is often taken as a null hypothesis, to be rejected in favor of some more structured ecologically relevant process. This was common practice in the nineteenth century (Darwin, 1881, Hensen, 1884) and is still a very common hypothesis in ecological models. Besag and Diggle (1977) discuss how to assess such a pattern (as well as more complex ones) using Monte Carlo testing, which enables a researcher to test specific hypotheses by simulating the assumed process, and then to check whether the observed statistic of interest is extreme among the simulations. Among other examples, the authors applied this to the locations of 65 Japanese black pine saplings (Numata, 1961; *cf.* Bartlett, 1964). More specifically, they used Monte Carlo testing on a χ^2-statistic comparing observed intertree distances to what would be expected under spatial randomness. The observed χ^2-statistic, which would have been deemed significant were the intertree distances independent, was in fact found consistent with a Poisson process. Much confusion has arisen in the ecological literature (and elsewhere) from a failure to appreciate the statistical dependence present in inter-event distances of a Poisson process.

A more detailed analysis of spatial patterns of ponderosa pines was performed by Getis and Franklin (1987) who found that, while the overall pattern of locations was consistent with spatial randomness, nearest neighbor distances for individual trees showed evidence of clustering on relatively large scales (about 20 m), and inhibition (presumably due to competition) on smaller scales (about 6 m). Here the mapping was done from aerial photographs, and the smallest resolvable distance was 2.4 m.

7.2.2 Cluster Processes

The concepts of clustering and regularity are important ecological concepts, describing deviations from the completely random process. On an intuitive level, clustering describes the phenomenon of an ecological niche, or local regions with higher than average density, separated by regions of low density, while regularity indicates a tendency towards spacing between individuals.

A cluster point process is a two-tiered process, defined in a conditional fashion. Given a point process N_c of cluster centers, one associates with each of its events a secondary point process $N_s(\cdot|x)$, centered at an event

at x. The cluster process is the superposition of these secondary processes. Formally,

$$N(A) = \int_{\mathcal{X}} N_s(A|x)\, N_c(dx)\,.$$

Usually the secondary processes are assumed independent, in which case the pgfl takes the simple form

$$G(h) = G_c(G_s(h|\cdot))\,,$$

where G_c is the pgfl of the process of cluster centers and $G_s(\cdot|x)$ is the pgfl of a secondary process centered at x. A necessary and sufficient condition for the existence of a cluster process is that

$$\int_{\mathcal{X}} (1 - \zeta_s(A|x))\, N_c(dx) < \infty \quad \text{a.s. } [N_c]\,.$$

A special case which has found many applications is the Poisson cluster process, where N_c is a Poisson process. The pgfl for a Poisson cluster process is

$$G(h) = \exp\left(\int_{\mathcal{X}} (G_s(h|x) - 1)\, \Lambda(dx)\right)\,.$$

It is easily shown that the Poisson cluster process is overdispersed with respect to a Poisson process with the same mean measure, i.e., that the cluster process shows greater variability in the number of events in a set. This overdispersion has often been taken as a definition of clustering in both ecological and engineering literature. However, it is easy to construct clustering processes (where N_c is non-Poisson) which are underdispersed relative to a Poisson process. For example, the findings (described above) of Getis and Franklin (1987), as well as the similar earlier results of Besag (1977), may be described by a cluster process (on larger scales) driven by a primary process that is more regular (on small scales) than a Poisson process.

The most common Poisson cluster process is the Neyman-Scott process, in which a random number of points are laid out in an i.i.d. fashion around the cluster center. This model was introduced by Neyman (1939) to describe the dispersion of larvae in a field. It has since found important applications in astronomy to describe the distribution of clusters of galaxies (Neyman and Scott, 1959; Peebles, 1980), and in hydrology, where it has been used to describe precipitation (Kavvas and Delleur, 1981, Rodriguez-Iturbe *et al.*, 1984; *cf.* ch. 4).

7.2.3 The Cox Process

The doubly stochastic Poisson process (often called a Cox process) arises when the mean measure Λ of a Poisson process is taken as a realization of a nonnegative stochastic process. A detailed discussion can be found in Grandell (1976) and Karr (1986). The pgfl of a Cox process is

$$G(h) = L_\Lambda(1-h)\,,$$

where L_Λ is the Laplace functional of the stochastic process (random measure) Λ. It follows that $\mathbf{Var}\,N(A) = \mathbf{E}\,N(A) + \mathbf{Var}\,(\Lambda(A))$, so that a Cox process is also overdispersed relative to a Poisson process.

As an example, consider the *shot noise* process, used by Vere-Jones and Davies (1966) to model earthquake sequences (including aftershocks). It is a Cox process with Λ given by

$$\Lambda(A) = \sum_i Y_i \int_{A+\tau_i} f(u)\,du\,,$$

where τ_i are the locations of events in a temporal Poisson process of constant rate ν, which triggers stresses of random amplitude Y_i, assumed i.i.d. These can give rise to major earthquakes. The intensity then decays according to some nonnegative integrable function f on $[0,\infty)$, possibly yielding aftershocks. Consequently,

$$G(h) = \exp\left(\int_{\mathcal{X}}\int_{\mathbf{R}_+} \phi[(1-h(t))f(t-x)\,dt - 1]\nu\,dx\right)\,,$$

where $\phi(t) = \mathbf{E}\exp(-tY)$. Comparing this to the Poisson cluster process pgfl given above, we see that it is of the same form. Hence, the shot noise process (so named since the moments agree with the moments derived by Campbell, 1909, for shot noise in vacuum tubes) is a Poisson cluster process (in fact, a Neyman-Scott process), and the two different mechanisms for constructing the process are indistinguishable from data. However, from an ecological point of view the two mechanisms are very different, and need to be distinguished from each other. In order to do so, more complex descriptions (perhaps involving more factors or species) are required.

Although the Cox process is overdispersed (clustered) relative to a Poisson process, a multivariate version can be constructed to model extreme inhibition between patterns. Let $\mathbf{N} = (N_1, N_2)$ be a bivariate point process driven by a bivariate nonnegative stationary stochastic process $\mathbf{\Lambda}(x)$, such

that given Λ, the two components N_1 and N_2 are independent Poisson processes, but $\Lambda_1(x)+\Lambda_2(x) = \nu$, where ν is a positive constant. Then (Diggle, 1983, sec. 6.6.2)

$$\lambda_{12}(u) = -c_{22}(u) + \lambda_1\lambda_2\,,$$

where $c_{22}(u)$ is the covariance density for $\Lambda_2(x)$. Consequently,

$$K_{12}(t) - \pi t^2 = -(2\lambda_1\lambda_2)^{-1}\sum_1^2 (K_{jj}(t) - \pi t^2)\,.$$

A plot of a nonparametric estimate of the left-hand side of this equation against a similar estimate of the right-hand side may indicate the adequacy of this model.

Diggle (1983, sec. 7.7) applied this model to the Lansing Woods data. As demonstrated by Besag (1977), there is a strong negative dependence between maples and hickories. The diagnostic plot mentioned above indicates that the fit of the competing Cox model is reasonable. However, the superposition of maples and hickories, which under this model should exhibit spatial randomness, does not follow a Poisson process. When adding the oaks, the Poisson fit for the superposition is adequate (although there still is some indication of clustering in the superposition process, possibly due to the other kinds of trees that are left out of the analysis). The oaks exhibit much less overdispersion than the other two species. A nonparametric estimate of (local) intensity confirms that a compensatory mechanism may be operating, but does on the other hand cast some doubt over the stationarity assumption. On the whole, this analysis, while providing a nice description of the observed spatial pattern, fails to produce an ecological explanation of it.

7.2.4 Markovian Point Processes

Markovian models, which are defined through a local dependence structure, have found much use in biology. In the spatial context, Markovian point processes were introduced by Strauss (1975) and by Ripley and Kelly (1977). A point process on a finite region A is Markov of range δ if the conditional density of a point at x, given all the points in $A \setminus \{x\}$, only depends on the points in the sphere of radius δ around x (excluding x itself). Call two points *neighbors* if their distance is less than δ, and define a *clique* as a set of mutual neighbors. It is convenient to describe the distribution of a

point process in terms of its likelihood ratio (Radon-Nikodym derivative) with respect to a unit rate Poisson process. In general this can be written

$$l(x_1, \ldots, x_n) \propto \prod_i g_i(x_i) \prod_{j>i} g_{ij}(x_i, x_j) \cdots g_{12\ldots n}(x_1, \ldots, x_n).$$

Ripley and Kelly proved that if a process is Markovian, it must have all of the g-functions identically 1, except when the arguments constitute a clique. This generalizes to other neighborhood systems, not necessarily distance-based.

The simplest nontrivial conditionally specified point process (also called a Gibbs point process) is one in which only pairwise interactions are allowed. Then

$$l(x_1, \ldots, x_n) \propto \exp\left(\sum_{i=1}^{n} \psi_1(x_i) + \sum_{i<j} \psi_2(x_i, x_j)\right).$$

The functions ψ_2 are called point pair potentials. This name comes from statistical mechanics, where models of this sort are used to describe the potential energy of a set of particles. The process is Markovian of range δ if $\psi_2(x, y) = 0$ whenever $|x - y| > \delta$. If the point process is stationary, ψ_2 depends only on the distance between its arguments, and ψ_1 is a constant. Writing $\psi_2(x, y) = V(|x - y|)$, we can specify the type of interaction by specifying V. These models are most commonly used to model repulsive interactions, leading to what is often called a regular point pattern (Strauss, 1975, Ogata and Tanemura, 1984). Examples include $V(r) \equiv 0$, the Poisson process; $V(r) = -\log(1 - \exp(-(r/\sigma)^2))$, a soft core repulsive model; $V(r) = (\sigma/r)^k$, an intermediate case; and $V(r) = \infty$ if $r \leq \sigma$, and 0 otherwise, a hard core rejection model (where no points are closer than σ).

Bartlett (1975, sec. 3.2.2) applied a simple inhibitory model to the spatial distribution of gulls' nests. The idea was to regard the distribution of nests as following a Poisson process, but to allow for the association with each point of a random cutoff, within which radius no other nests can be found. This combined the hard-core rejection model above with features of the Cox process of §7.2.3.

It is difficult to estimate the parameters of Markovian point process models, mainly because the normalizing constant in the likelihood is very hard to evaluate. The two most common approaches are to use approximations developed in statistical physics for the normalizing constant (Ogata and Tanemura, 1984, discuss some of these approximations; see also Ripley, 1988, ch. 4, and recent work using stochastic approximation techniques by

Moyeed and Baddeley, 1989), or to use Besag's method of pseudolikelihood (Besag 1975, 1977; some recent theoretical results are in Jensen and Møller, 1989, and Särkkä, 1989).

7.3 A Spatio-Temporal Point Process Model for Tree Growth

Most situations where spatial point process models can be useful include a temporal aspect. In this section, we discuss a possible approach to modeling tree growth in a pristine forest, with a view toward use for regenerative policies in national parks following major natural disasters. The intent of this section is to indicate how a physically based model may be used to suggest facets of a stochastic model of forest growth. This is different from the statistical (or descriptive) models that have been the main emphasis in the past for such efforts.

In order to construct such a model, it seems reasonable to separate out the occurrence of new growth, the process of growth itself, and the process of tree death, as suggested by Rathbun and Cressie (1989). We will call the three components the birth, growth, and death processes. For simplicity, we will only consider a single species and will use a discrete time scale of, say, a year. We separate trees into adult individuals that are well established and juveniles that are struggling to succeed. Foresters tend to make this classification based on simple measurements such as base diameter. It is assumed that the forest under study is mapped completely (with regard to the species of interest) at regular intervals. Sterner *et al.* (1986) developed models similar to those discussed below for the interaction of four tropical tree species.

The birth process, at any given time t, will be constructed conditional upon the location of mature adults. Potential sites for new juveniles are obtained from a cluster process of Neyman-Scott type with cluster centers given by the mature adult locations. This represents the spread of seeds from the adult trees. The germination of seeds, or more precisely, germination and subsequent establishment of a juvenile plant, is modeled by thinning the potential sites, i.e., by deletion of each cluster point independently, with probability depending on the configuration of adults around the seedling. If the nearest adult is far away, so the seedling is in a relatively open area, it would have a comparatively high probability of germination, whereas if the seedling is located next to an adult, this probability would be low.

The growth process takes into account the amount of sunlight available to a tree by using data on crown angle and height. This process determines the development of marks from year to year, rather than the points themselves.

The death process needs to have several factors. The process of locations is thinned using a probability proportional to size (and thereby, approximately, to age). In addition, competition between juveniles affects their survival probabilities. The effect of large windstorms can be thought of as a constant force (this would usually have a preferred direction) whose mortality effect on a given tree depends on its size and on the configuration and sizes of its neighbors. Large isolated trees have the highest mortality from windstorms, whereas sheltered trees in the middle of a tight cluster have the smallest. Major disasters, such as fires, can be modeled using dependent thinning, where nearby trees have a very high probability of death, conditional upon a given tree to have succumbed to fire. Each year the probability of a major disaster is very small. It can be estimated from tree-ring data. The probability of death from storms is comparatively higher, and may vary from year to year, based on meteorological factors.

The combination of these forces yields an anisotropic process, for which one can determine, at least qualitatively, the behavior of second-order intensities. Since many of the subprocesses are observable, it is possible to assess these aspects of the model using data. The combination of the subprocesses into a complex mechanism and the detailed fitting and inference yields many challenging theoretical problems. The main use for this type of model is to assess effects of changes in the driving forces of the ecological process, and evaluate various possible reseeding policies. It is straightforward to include modest amounts of harvesting in the model, which can then be used to assess various recruitment policies. For assessment purposes, computer simulation is likely to be necessary.

7.4 Conclusion and Further Directions

The use of point process models in ecology to date has perhaps not reaped the expected benefits. While the models sometimes have managed to describe a complex data set in a relatively compact form, there have been very few instances where data-analytic findings have found proper explanation in ecological/biological terms. With the increased quality of aerial maps, the requisite data for certain types of vegetation ecology studies will be made more readily available, and the quality of the analysis should improve.

The interaction between statisticians and subject area scientists is always the key to relating data-analytic findings to scientific explanations. Increased awareness in the ecological community of the methods made available by improved methods of spatial statistical analysis will undoubtedly benefit both statisticians and ecologists. There is a substantial need for more theoretical research into statistical inference based on interacting components of complex systems, and into the comparison of model data (be it the result of simulation, mathematical, or stochastic anaylsis) to "ground truth" measurements.

Bibliography

[1] Baddeley, A. J., A limit theorem for some statistics of spatial data, *Adv. Appl. Prob.* **12** (1980), 447–461.

[2] Baddeley, A. J., and B. W. Silverman, A cautionary example of the use of second-order methods for analyzing point patterns, *Biometrics* **40** (1984), 1089–1093.

[3] Bartlett, M. S.,The spectral analysis of two-dimensional point processes, *Biometrika* **51** (1964), 299–311.

[4] Bartlett, M. S., *The Statistical Analysis of Spatial Pattern*, Chapman and Hall, London, 1975.

[5] Besag, J., Statistical analysis of non-lattice data, *Statistician* **24** (1975), 179–195.

[6] Besag, J., Some methods of statistical analysis for spatial data, *Bull. Int. Stat. Inst.* **47(2)** (1977), 77–92.

[7] Besag, J., and P. J. Diggle, Simple Monte Carlo tests for spatial pattern, *Appl. Stat.* **26** (1977), 327–333.

[8] Brillinger, D. R., Comparative aspects of the study of ordinary time series and of point processes, pp. 33–133 in *Developments in Statistics* **1**, P. R. Krishnaiah, ed., Academic Press, Orlando, 1978.

[9] Campbell, N. R., The study of discontinuous phenomena, *Proc. Camb. Philos. Soc. Math. Phys. Sci.* **15** (1909), 117–136.

[10] Daley, D. J., and D. Vere-Jones, *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York, 1988.

[11] Darwin, C., *The Formation of Vegetable Mould Through the Action of Worms*, John Murray, London, 1881.

[12] Dawson, D., and G. Ivanoff, Branching diffusions and random measures, in *Branching Processes—Advances in Probability and Related Topics*, A. Joffe and P. Ney, eds., Dekker, New York, 1978.

[13] Diggle, P. J., *Statistical Analysis of Spatial Point Patterns*, Academic Press, London, 1983.

[14] Ford, E. D., Competition and stand structure in some even-aged plant monocultures, *J. Ecol.* **63** (1975), 311–333.

[15] Ford, E. D., and P. J. Diggle, Competition for light as a spatial stochastic process, *Ann. Bot.* **48** (1981), 481–500.

[16] Gerrard, D. J., Competition quotient: A new measure of the competition affecting individual forest trees, *Res. Bull.* **20** (1969), Agricultural Experiment Station, Michigan State University.

[17] Getis, A., and J. Franklin, Second-order neighborhood analysis of mapped point patterns, *Ecology* **68** (1987), 473–477.

[18] Grandell, J., *Doubly Stochastic Poisson Processes*, Lecture Notes in Math **529**, Springer, Berlin, 1976.

[19] Greig-Smith, P., The use of random and contiguous quadrants in the study of the structure of plant communities, *Ann. Bot.* **16** (1952), 293–316.

[20] Hanisch, K. H., and D. Stoyan, Formulas for the second-order analysis of marked point processes, *Math. Operationsforsch Stat. Ser. Stat.* **10** (1979), 555–560.

[21] Harkness, R. D., and V. Isham, A bivariate spatial point pattern of ants' nests, *Appl. Stat.* **32** (1983), 293–303.

[22] Harrison, J. S., and J. B. Gentry, Foraging pattern, colony distribution, and foraging range of the Florida harvester ant, *Pogonomyrmex bodins*, *Ecology* **62** (1981), 1467–1473.

[23] Hensen, V., Über die Bestimmung der Planktons oder des im Meer triebenden Materials an Pflanzen und Tieren, *Ber. Comm. Wiss. Untersuch. Deutschen Meere* **5** (1884).

[24] Hertz, P., Über die gegenseitligen durchschnittlichen Abstand von Punkten, die mit bekannter mittlerer Dichte im Raum angeordnet sind, *Math. Annal.* **67** (1909), 387–398.

[25] Jensen, J. L., and J. Møller, *Pseudolikelihood for Exponential Family Models of Spatial Processes*, Research report **203**, Department of Theoretical Statistics, University of Aarhus, 1989.

[26] Karr, A. F., *Point Processes and Their Statistical Inference*, Dekker, New York, 1986.

[27] Kavvas, M. L., and J. Delleur, A stochastic cluster model of daily rainfall occurrences, *Water Resour. Res.* **17** (1981), 1151–1160.

[28] Kershaw, K. A., The use of plant cover and frequency in the detection of pattern in plant communities, *Ecology* **38** (1957), 291–299.

[29] Kulperger, R. J., Parametric estimation for a simple branching diffusion process, *J. Multivar. Anal.* **9** (1979), 101–115.

[30] Legendre, P., and M.-J. Fortin, Spatial pattern and ecological analysis, *Vegetatio* (1989).

[31] Lerings, S. C., and N. R. Franks, Patterns of nest dispersion in a tropical ground ant community, *Ecology* **63** (1982), 338–344.

[32] Linhart, Y. B., Density-dependent seed germination strategies in colonizing versus non-colonizing plant species, *J. Ecol.* **64** (1976), 375–380.

[33] Moyeed, R. A., and A. J. Baddeley, *Stochastic Approximation of the MLE for a Spatial Point Pattern*, Report BS-R8926 (1989), Centrum voor Wiskunde en Informatica, Stichting Math. Centrum, Amsterdam.

[34] Neyman, J., On a new class of 'contagious' distributions, applicable in entomology and bacteriology, *Ann. Math. Stat.* **10** (1939), 35–57.

[35] Neyman, J., and E. L. Scott, Large scale organization of the distribution of galaxies, *Handbuch der Physik* **53** (1959), 416–444.

[36] Numata, M., Forest vegetation in the vicinity of Choshi—Coastal flora and vegetation at Choshi, Chiba Prefecture IV, *Bull. Choshi Marina Lab., Chiba University* **3** (1961), 28–48.

[37] Ogata, Y., and M. Tanemura, Likelihood analysis for spatial point patterns, *J. R. Stat. Soc., B* **46** (1984), 496–518.

[38] Ohser, J., and D. Stoyan, On the second-order and orientation analysis of planar stationary point processes, *Biometr. J.* **23** (1981), 523–533.

[39] Peebles, P. J. E., *The Large-Scale Structure of the Universe*, Princeton University Press, Princeton, 1980.

[40] Rathbun, S. L., and N. A. C. Cressie, A model for a spatial birth process, paper presented at the 1989 Joint Statistical Meetings, Washington, D.C., Department of Statistics, Iowa State University, 1989.

[41] Ripley, B. D., The second-order analysis of stationary point processes, *J. Appl. Prob.* **13** (1976), 255–266.

[42] Ripley, B. D., Spatial point pattern analysis in ecology, pp. 407–423 in *Developments in Numerical Ecology*, P. and L. Legendre (eds), NATO ASI Series **G14**, Springer-Verlag, Berlin, 1987.

[43] Ripley, B. D., *Statistical Inference for Spatial Processes*, Cambridge University Press, Cambridge, 1988.

[44] Ripley, B. D., and F. P. Kelly, Markov point processes, *J. London Math. Soc.* **15** (1977), 188–192.

[45] Rodriguez-Iturbe, I., V. K. Gupta, and E. Waymire, Scale considerations in the modeling of temporal rainfall, *Water Resour. Res.* **20** (1984), 1611–1619.

[46] Särkkä, A., *On Parameter Estimation of Gibbs Point Processes Through the Pseudolikelihood Method*, Technical report 4/1989, Department of Statistics, University of Jyväskylä, 1989.

[47] Solow, A. R., Bootstrapping sparsely sampled spatial point patterns, *Ecology* **70** (1989), 379–382.

[48] Sterner, R. W., C. A. Ribic, and G. E. Schatz, Testing for life historical changes in spatial patterns of four tropical tree species, *J. Ecology* **74** (1986), 621–633.

[49] Strauss, D. J., A model for clustering, *Biometrika* **62** (1975), 467–475.

[50] Vere-Jones, D., and R. B. Davies, A statistical survey of earthquakes in the main seismic region of New Zealand, Part 2—Time series analyses, *N.Z. J. Geol. Geophys.* **9** (1966), 251–281.

8

Spatial Signal-Processing in Radars and Sonars

T. T. Kadota
AT&T Bell Laboratories

8.1 Introduction

Radars and sonars are used for detecting and tracking targets. The surveillance radars and sonars typically employ arrays of sensors (or radiators) placed on the ground or in the ocean. A planar array is used for sonar to detect seismic explosions and for radar to track distant flying objects, and a linear array is used for sonar to detect distant underwater objects. The data thus obtained are in the form of a set of time-series that are related by the spatial configuration of the array. The task of the radar and sonar systems is to process these data to detect the signal transmitted from a target and estimate the signal parameters related to the target location and velocity. Typically, the data contain noise and interfering signals besides the target signal, and the "signal-processing" (processing of these data) requires suppression of the noise and interference and enhancement of the signal.

The literature on spatial signal-processing is enormous: the *IEEE Transactions on Acoustic, Speech and Signal Processing*, the *IEEE Transactions on Aerospace Electronics*, and the *Journal of the American Statistical Association*, just to name a few. Covering the entire area and providing an adequate survey is beyond the scope of this report. Instead, by using a simple example, we give a list of how statistical decision and estimation theory is used on this form of multidimensional data to derive a signal-processing algorithm and indicate how to extend the basic approach to more complex problems in reality.

8.2 Detection and Estimation Problem

The example we have chosen is the signal-processing of underwater acoustic data for detecting a narrow-band signal transmitted from a distant source and determining the direction of its arrival. The signal detection problem is traditionally cast as a problem of testing a null hypothesis H_0 (signal absent) against an alternative hypothesis H_1 (signal present) as follows (Helstrom, 1986):

$$dr_j(t) = \begin{cases} z_j(t)\,dt + dw_j(t)\,, & (H_0) \\ \qquad\qquad\qquad\qquad j = 1, \ldots, J;\ 0 < t \le T\,, & \\ s_j(t)\,dt + z_j(t)\,dt + dw_j(t)\,, & (H_1) \end{cases} \tag{8.1}$$

where T is an observation time interval, $dr_j(t)$ is the (incremental) acoustic data recorded at the j^{th} sensor at time t, $s_j(t)$ is the acoustic signal received at the j^{th} sensor at time t, $dw_j(t)$ is the (incremental) background noise at the j^{th} sensor at time t (assumed to be white Gaussian, independent from sensor to sensor), and $z_j(t)$ is the interference (or additional noise) at the j^{th} sensor at time t. We assume that the background noise w and the interference z are mutually independent.

Adopting the Neyman-Pearson formulation, we obtain the likelihood ratio between the two hypotheses H_0 and H_1 (i.e., the Radon-Nikodym derivative between the two probability measures induced by the random fields of (8.1)). It can be expressed as

$$\frac{dP_1}{dP_0}(r) = \frac{E_z\,\exp[(s+z,r) - \frac{1}{2}\|s+z\|^2]}{E_z\,\exp[(z,r) - \frac{1}{2}\|r\|^2]}\,, \tag{8.2}$$

where

$$\begin{aligned} r &= \{dr_j(t),\, 0 < t \le T,\, j = 1, \ldots, J\}\,, \\ s &= \{s_j(t),\, 0 < t \le T,\, j = 1, \ldots, J\}\,, \\ z &= \{z_j(t),\, 0 < t \le T,\, j = 1, \ldots, J\}\,, \end{aligned}$$

and E_z denotes the expectation with respect to the probability distribution of the z process and the inner product and the norm are defined by

$$(s,r) = \sum_{j=1}^{J} \int_0^T s_j(t)\,dr_j(t)\,, \qquad \|s\|^2 = \sum_{j=1}^{J} \int_0^T s_j^2(t)\,dt\,.$$

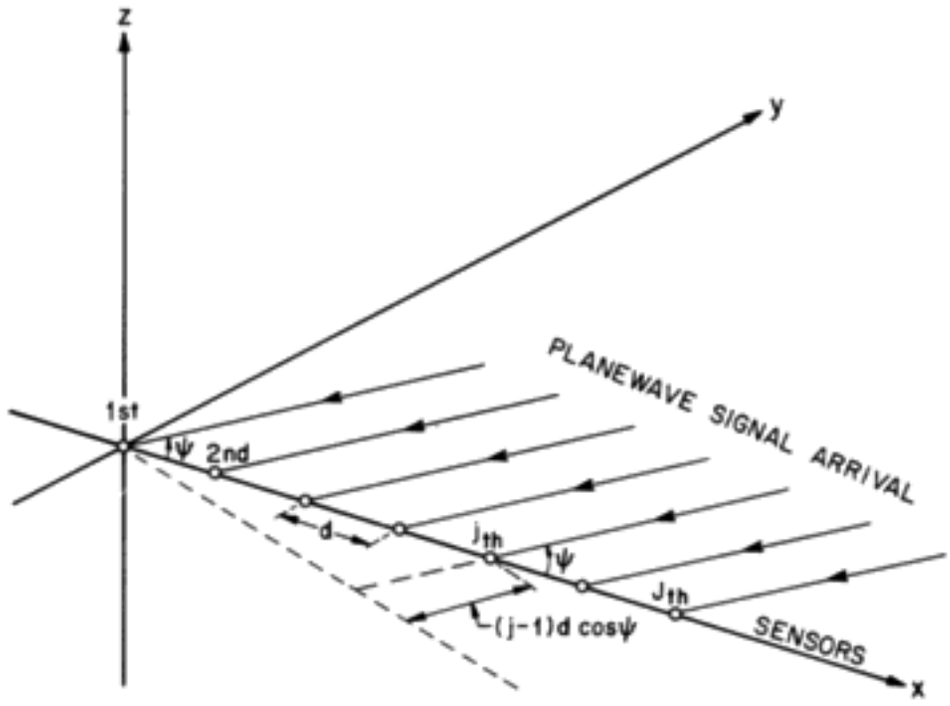


FIGURE 8.1: Linear array of uniformly spaced sensors.

8.3 Using Linear Arrays of Uniformly Spaced Sensors

The most widely studied case is the one with a linear array of equally spaced sensors (see Figure 8.1) and a narrow-band planewave signal with a known carrier frequency, which may be represented by

$$s_j(t) = \mathrm{Re}\, s_e(t) \exp[-i\omega(t-\tau_j)] = \mathrm{Re}\, a_j(\psi) s_e(t) \exp(-i\omega t)\,, \qquad (8.3)$$

where Re denotes the real part of what follows, $\{s_e(t),\, 0 < t \leq T\}$ is the complex signal envelope function, and

ω = the carrier (angular) frequency such that $\omega T \gg 2\pi$,

τ_j = $(j-1)d\cos\psi/c$

= the time delay of the signal planewave arrival at the j^{th} sensor relative to the first,

d = the distance between two adjacent sensors,

c = the velocity of sound propagation,

ψ = the assumed direction of the signal planewave arrival, and

$$a_j(\psi) = \exp[i(j-1)\frac{\omega}{c}d\cos\psi], \quad j = 1,\ldots,J. \tag{8.4}$$

Then, in the absence of the interference z, the logarithm of the likelihood ratio becomes

$$\log\frac{dP_1}{dP_0}(r) = (s,r) - \frac{1}{2}\|s\|^2 = \mathrm{Re}\sum_{j=1}^{J} a_j^*(\psi)\int_0^T s_e^*(t)\,d\bar{r}_j(t) - \frac{1}{2}\|s\|^2, \tag{8.5}$$

where $d\bar{r}_j$ is the envelope function of the narrow-band representation of the data, namely,

$$dr_j(t) = \mathrm{Re}\,d\bar{r}_j(t)\exp(-i\omega t).$$

Suppose a planewave arrives in the direction of ψ_0. Then the signal part of the data-dependent term, the first term of (8.5), is proportional to

$$\mathrm{Re}\,a(\psi)^\dagger a(\psi_0) = \mathrm{Re}\,\frac{1-\exp[iJ\omega d(\cos\psi_0-\cos\psi)/c]}{1-\exp[i\omega d(\cos\psi_0-\cos\psi)/c]}, \tag{8.6}$$

where $a(\psi) = (a_1(\psi),\ldots,a_J(\psi))$ is the direction (or steering) column vector in the ψ-direction and $\dagger$ denotes the complex conjugate transpose. Equation (8.6) is in the form of a "main beam" centered at ψ and "side lobes" on each side of the main beam as ψ_0 varies from $-\pi/2$ to $\pi/2$. Hence, the processing (of the data r) described by (8.6) is called "beam forming" (Steinberg, 1976). On the other hand, if the actual direction of the planewave arrival ψ_0 is fixed and the assumed direction ψ is varied, (8.6) attains the maximum at $\psi = \psi_0$, namely, when the beam is steered at the signal source. Thus, detection of the planewave signal and estimation of its direction of arrival are done by varying ψ (steering the beam) from $-\pi/2$ to $\pi/2$ to find the maximum of (8.6) and comparing it to a preassigned threshold determined by the false-alarm probability (according to the Neyman-Pearson criterion) (Helstrom, 1986). Instead of steering a single main beam, one can place many beams to fill the angular sector $(-\pi/2,\pi/2)$ by providing many direction vectors $a(\psi_m)$, $m = 1,\ldots,M$. Then by comparing the magnitude of (8.6) for each ψ_m, instead of varying ψ, we effectively accomplish the same task of detection and estimation.

In the presence of the interference z, some modification to the beam-forming is necessary. For any interference to be effective against the signal, it must have energy at or near the carrier frequency (otherwise, it can be

simply filtered out). For the sake of simplicity, suppose we have one sinusoidal interference arriving in the ψ' direction with a Gaussian distributed amplitude, namely,

$$z_j(t) = \mathrm{Re}\, u \exp\{-i\omega[t-(j-1)d\cos\psi'/c]\} = \mathrm{Re}\, u\bar{z}_j \exp(-i\omega t)\,, \quad (8.7)$$

where the second equality defines $\bar{z}_j$ and u is a complex Gaussian variable with mean 0 and variance β^2. This represents a planewave arriving through a multipath medium causing the Rayleigh fading. By carrying out the expectation with respect to z in (8.2) (i.e., with respect to u), the data-dependent term of the log likelihood ratio becomes

$$\begin{aligned} \mathrm{Re}\,(\bar{s},\, \bar{r} - E_0\{\bar{z}|\bar{r}\}) &= \mathrm{Re}\left(s_e,\, \left[a(\psi)^\dagger - \frac{\beta^2}{1+\beta^2} a(\psi)^\dagger a(\psi')a(\psi')^\dagger\right] r\right) \\ &= \mathrm{Re}\,((I+R)^{-1}\bar{s},\, \bar{r})\,, \end{aligned} \quad (8.8)$$

where

$$\bar{s}_j(t) = a_j(\psi)s_e(t)\,,$$

and $E_0\{\bar{z}|\bar{r}\}$ is the conditional expectation of $\bar{z}$ given $\bar{r}$ under H_0, and $R = \beta^2 a(\psi')a(\psi')^\dagger$. The first member of (8.8) has an obvious interpretation: the optimum processing is to make the least-mean-square-error estimate of the interference and subtract the estimate from the data before the beamforming. The second member, on the other hand, shows how the conventional beamformer is to be modified due to the presence of the interference. By recalling that $a(\psi')$ is the steering vector in the direction of the interference, the modified beamformer has a considerably reduced output in the direction of the interference, thus acquiring the term, *null-steering* (Steinberg, 1976; Gabriel, 1976; Friedlander and Porat, 1989).

In practice, the interfering source is not known a priori and its covariance matrix R must be estimated. The estimation may be done beforehand or simultaneously with the detection operation, assuming that the direction of the signal arrival is known (which is the case with the fixed multibeam scheme). This simultaneous method is referred to as the adaptive beamforming and is implemented by attaching a variable gain (or weight) and a variable time-delay (which are adjusted as data are obtained) to the output of each sensor. Of course, such an adjustment must be done rapidly so that accurate signal detection and direction-of-arrival estimation can be accomplished. The iterative methods of adjusting and their convergence characteristics have been extensively studied. Monzingo and Miller (1980)

is one of the comprehensive textbooks on the subject. A benchmark paper series edited by Haykin (1980) has many important papers, including those of Gabriel, Applebaum, Widrow, Griffiths, and Owsley.

8.4 Using General Arrays of Sensors

The results presented above can be generalized as follows (Kadota and Romain, 1977): first, instead of a linear array of equally spaced sensors, we can consider a general three-dimensional array (or configuration) with the coordinates (ξ_j, η_j, ζ_j), $j = 1, \ldots, J$. Then the planewave arriving in the (θ, ψ)-direction, where θ and ψ are the elevation and the azimuthal angles, incurs at the j^{th} sensor the phase shift expressed by

$$a_j(\omega, \theta, \psi) = \exp\left[i\frac{\omega}{c}(\xi_j \cos\theta \cos\psi + \eta_j \cos\theta \sin\psi + \zeta_j \sin\theta)\right] . \qquad (8.9)$$

Next, instead of a single planewave, we consider a signal consisting of K planewaves, each having a different frequency ω_k, $k = 1, \ldots, K$, and each arriving in M different directions (θ_m, ψ_m), $m = 1, \ldots, M$. For convenience, we assume that $\omega_k T$ is an integral multiple of 2π for every k. Also, rather than a "slowly varying" envelope function $s_e(t)$, we consider a complex Gaussian variable (independent of time) as the amplitude of each planewave. Thus, the signal at the j^{th} sensor is now given by

$$s_j(t) = \sum_{k=1}^{K} \sum_{m=1}^{M} \operatorname{Re} u_{km} a_j(\omega_k, \theta_m, \psi_m) \exp(-i\omega_k t) , \qquad (8.10)$$

where $\{u_{km}\}$, $u_{km} = \tilde{u}_{km} + i\hat{u}_{km}$, are complex, zero-mean, Gaussian variables with

$$E\,\tilde{u}_{km}\tilde{u}_{k'm'} = E\,\hat{u}_{km}\hat{u}_{k'm'} = \rho_{kk'mm'} , \quad k, k' = 1, \ldots, K;\ m, m' = 1, \ldots, M .$$

We allow some ρ's to be zero since not all frequency components have all M arrival directions. The interference is now generalized to

$$v_j(t) = \sum_{\ell=0}^{\infty} \operatorname{Re} \bar{v}_{j\ell} \exp\left(-i\frac{2\pi\ell t}{T}\right) , \qquad (8.11)$$

where $\{\bar{v}_{j\ell}\}$, $\bar{v}_{j\ell} = \tilde{v}_{j\ell} + i\hat{v}_{j\ell}$, are complex, zero-mean, Gaussian variables. That is, for each j, $v_j(t)$ is a discretized version of the spectral representation

of a general stationary noise. Then the data-dependent part of the log-likelihood ratio takes the following quadratic form in the data:

$$(x,(I+V)^{-1}S(I+V+S)^{-1}x)\,, \tag{8.12}$$

where

$$x=(\tilde{x}_1,\ldots,\tilde{x}_J,\ldots,\tilde{x}_{(K-1)J+1},\ldots,\tilde{x}_{KJ},\hat{x}_1,\ldots,\hat{x}_J,\ldots,\hat{x}_{(K-1)J+1},\ldots,\hat{x}_{KJ})\,,$$

$$\tilde{x}_{(k-1)J+j}=\left(\frac{2}{T}\right)^{\frac{1}{2}}\int_0^T\cos\omega_k t\,dr_j(t)\,,\quad \hat{x}_{(k-1)J+j}=\left(\frac{2}{T}\right)^{\frac{1}{2}}\int_0^T\sin\omega_k t\,dr_j(t)\,,$$

$$S=\begin{bmatrix}\tilde{S} & \hat{S}\\ \hat{S}^{\dagger} & \tilde{S}\end{bmatrix}\,,$$

with the $((k-1)J+j,(k'-1)J+j')^{\text{th}}$ elements of $\tilde{S}$ and $\hat{S}$ given respectively by the real and the imaginary parts of

$$\sum_{m,m'=1}^{M}\rho_{kk'mm'}a_j(\omega_k,\theta_m,\psi_m)a_j^*(\omega_{k'},\theta_{m'},\psi_{m'})\,,\ j,j'=1,\ldots,J;\ k,k'=1,\ldots,K\,,$$

and

$$V=\begin{bmatrix}\tilde{V} & 0\\ 0 & \tilde{V}\end{bmatrix}\,,\quad (\tilde{V})_{(k-1)J+j,(k'-1)J+j'}=E\,\tilde{v}_{j\ell}\tilde{v}_{j'\ell}\delta_{kk'}=E\,\hat{v}_{j\ell}\hat{v}_{j'\ell}\delta_{kk'}\,,$$

$$\ell=\frac{\omega_k T}{2\pi}\,,\qquad k=1,\ldots,K\,.$$

The J sensors constitute spatial samplers of the available (acoustic) data and their configuration specifies the pattern of spatial sampling. This sampling pattern is incorporated into the covariance matrices S and V to influence the detection statistic (8.12) which specifies the data-processing algorithm. Although the linear array (with or without the equal spacing) is the most common configuration, due primarily to the ease of implementation, the sampling pattern can be considered as a factor with respect to which the detection and estimation performance can be optimized. In fact, we show next an interesting example of this application.

So far, we have assumed that the sensor positions are rigidly fixed and their coordinates are known a priori. Although this is the case with the phased-array radars and seismographic sensors, for underwater-acoustic sensors the exact positions in the ocean are difficult to determine and calibration of the array becomes necessary. One way to deal with this problem

is to model the deviation (or fluctuation) of the sensor position (from the presumed value) as an additional noise and incorporate into the optimum processor (for detection and estimation) the sensitivity of the performance to this noise. For example, the array gain (the output signal-to-noise ratio of an array processor for a given direction of signal arrival) may be maximized under the constraint that the array-gain sensitivity to the sensor-position noise be kept below a given level (Cox *et al.*, 1987). An alternative is to devise the sensor configuration so as to make these test statistics immune to the sensor-position fluctuation. The ESPRIT (Estimation of Signal Parameters by Rotational Invariance Techniques) method (Roy and Kailath, 1989) forms pairs of sensors to create an array of doublets such that it consists of two identical subarrays where one is a translate of the other. Suppose the signal consists of M planewaves with complex, zero-mean, Gaussian amplitudes, having the same frequency ω arriving from M directions $(0, \psi_m)$, $m = 1, \ldots, M$. We further assume for simplicity that the interference is absent. Suppose we have already detected the signal and our goal is to estimate the M arrival directions ψ_m, $m = 1, \ldots, M$, which are specified relative to the axis of the doublet (the displacement vector). Denote the data from the two subarrays of sensors by two $(J/2)$-vectors x and y, assuming J to be even,

$$x = (x_1, \ldots, x_{J/2}), \qquad x_j = \left(\frac{2}{T}\right)^{\frac{1}{2}} \int_0^T \exp(i\omega t)\, dr_j(t), \tag{8.13}$$

$$y = (y_1, \ldots, y_{J/2}), \qquad y_j = \left(\frac{2}{T}\right)^{\frac{1}{2}} \int_0^T \exp(i\omega t)\, dr_{\frac{J}{2}+j}(t), \quad j = 1, \ldots, J/2.$$

Then

$$\begin{aligned} R_{xx} &= E\, x x^* &&= AUA^* + I \\ R_{xy} &= E\, x y^* &&= AU\Phi^* A^*, \end{aligned}$$

where A, U, and Φ are $J \times M$, $M \times M$, and $M \times M$ matrices respectively and specified by

$$(A)_{jm} = a_j(\omega, 0, \psi_m), \; (U)_{mm'} = 2\rho_{mm'}, \; (\Phi)_{mm'} = \exp\left(i\frac{\omega\Delta}{c}\sin\zeta_m\right)\delta_{mm'},$$

$$j = 1, \ldots, J, \qquad m = 1, \ldots, M,$$

where Δ is the distance between the two paired sensors. Assuming U to be nonsingular, we observe that the determinant of

$$R_{xx} - I - \gamma R_{xy} = AU(I - \gamma\Phi^*)A^* \tag{8.14}$$

vanishes if and only if

$$\gamma = \exp(-i\frac{\omega\Delta}{c}\sin\psi_m), \quad m = 1,\ldots,M.$$

This fact can be used to estimate ψ_m, $m = 1,\ldots,M$, as follows: regard R_{xx} and R_{xy} as measured covariance matrices. For example, we might subdivide the observation interval T into N equal subintervals $((n-1)T/N, nT/N)$, $n = 1,\ldots,N$, where $N = \omega T/(2\pi)$, replace the integration limits in (8.13) by $(n-1)T/N$ and nT/N, $n = 1,\ldots,N$, and denote the integrals by $x_j(n)$ and $y_j(n)$, $j = 1,\ldots,J$; $n = 1,\ldots,N$. Then, put

$$\hat{R}_{xx} = \frac{1}{N}\sum_{n=1}^{N} x(n)x^*(n), \qquad \hat{R}_{xy} = \frac{1}{N}\sum_{n=1}^{N} x(n)y^*(n).$$

Now substitute these empirical matrices into the left-hand side of (8.14) and find M minima of the absolute value of the determinant as γ moves on the unit circle centered at the origin of the complex plane. Substitute the M γ-values corresponding to these minima and solve for ψ_m, $m = 1,\ldots,M$.

Observe that the knowledge of the sensor positions incorporated into A and of the signal powers $\rho_{mm'}$ is not required. Thus, this method of estimating the signal arrival directions is free of the costly array calibration. The price to be paid for this is that the two subarrays must be identical, with one being a translate of the other.

8.5 Future Research Considerations

The assumption that both the signal and the interference plus noise be Gaussian fields is primarily for mathematical convenience since the problem then is completely treatable by linear operators in Hilbert spaces, and Gaussian fields are the simplest class of the second-order random fields. However, there are evidences, especially in the case of the ocean acoustics, that the probability distributions of the interference fields considerably deviate from the Gaussian distribution (Middleton, 1987). Some simple analytical examples, such as the "contaminated Gaussian" distribution (Martin and Schwartz, 1971), have been proposed for the one-dimensional i.i.d. time series. Although the non-Gaussian interference makes the analytical solution to the optimum processing problem infeasible, some suboptimum processing methods are explored in special cases (Monzingo and Miller, 1980). Since it

is unrealistic to completely specify the probability distribution of the interference, a robust method, such as the min-max solution (Huber, 1981), has been sought. The results so far are restricted to the one-dimensional time series having independent identical distributions (Kassam and Poor, 1985), and generalization to the higher dimensional case with dependent distributions should be sought. Another area of investigation is the case where the interference is a nonstationary and inhomogeneous random field, such as a transient disturbance. In this case, one might use a semideterministic criterion rather than the totally probabilistic Neyman-Pearson criterion, and estimate (maximum likelihood) the interference z in (8.2) rather than average with respect to its probability distribution. One practical problem in dealing with multidimensional data is computational complexity. Even if there is an explicit algorithm for the optimum signal-processing, the complexity may be too prohibitive to justify its use. Thus, a trade-off between the detection-estimation performance and the computational complexity, or the cost of processing the data, must be considered. Study of this trade-off is another area of useful research in the future.

Bibliography

[1] Cox, H., R. M. Zeskind, and M. M. Owen, Robust adaptive beamforming, *IEEE Trans. Acoust., Speech, Signal Process.* **35** (1987), 1365–1375.

[2] Friedlander, B., and B. Porat, Performance analysis of a null steering algorithm based on direction-of-arrival estimation, *IEEE Trans. Acoust., Speech, Signal Process.* **37** (1989), 461–466.

[3] Gabriel, W. F., Adaptive arrays—An introduction, *IEEE Proc. Natl. Aerosp. Electron. Conf.* **64** (1976), 239–272.

[4] Haykin, S., ed., *Array Processing—Applications to Radar*, Benchmark papers in electrical engineering and computer science **22**, Dowden, Hutchinson and Ross, Stroudsburg, PA, 1980.

[5] Helstrom, C. W., *Statistical Theory of Signal Detection*, 2nd ed., Pergamon Press, Oxford, England, 1986, 87–95.

[6] Huber, P. J., *Robust Statistics*, John Wiley and Sons, New York, 1981.

[7] Kadota, T. T., and D. M. Romain, Optimum detection of Gaussian signal fields in the multipath-anisotropic noise environment and numerical evaluation of detection probabilities, *IEEE Trans. Inf. Theory* **23** (1977), 167–178.

[8] Kassam, S. A., and H. V. Poor, Robust techniques for signal processing: A survey, *IEEE Proc. Natl. Aerosp. Electron. Conf.* **73** (1985), 433–481.

[9] Martin, R. D., and S. C. Schwartz, Robust detection of a known signal in nearly Gaussian noise, *IEEE Trans. Inf. Theory* **17** (1971), 50–56.

[10] Middleton, D., Channel modeling and threshold signal processing in underwater acoustics: An analytical overview, *IEEE J. Oceanic Eng.* **12** (1987), 4–28.

[11] Monzingo, R. A., and T. W. Miller, *Introduction to Adaptive Arrays*, John Wiley and Sons, New York, 1980.

[12] Roy, R., and T. Kailath, ESPRIT—Estimation of Signal Parameters via Rotational Invariance Techniques, *IEEE Trans. Acoust., Speech, Signal Process.* **37** (1989), 984–995.

[13] Steinberg, B. D., *Principles of Aperture and Array System Design*, John Wiley and Sons, New York, 1976.

[14] Wegman, E. J., S. C. Schwartz, and J. B. Thomas, *Topics in Non-Gaussian Signal Processing*, Springer-Verlag, New York, 1989.

9

Stochastic Modeling in Physical Chemistry

Peter Clifford and N. J. B. Green
Oxford University

9.1 Introduction

How can corrosion be controlled in the cooling system of a nuclear reactor? What is the most efficient design for a solar cell? How do you build an artificial enzyme? These are just some of the important practical questions that lie behind the prolific research activity taking place in physical chemistry departments around the world.

As a branch of science, physical chemistry is defined not so much by the circumscription of its subject matter as by its method of approach, applicable to a wide diversity of problems arising from physics and chemistry on the one hand to biology and materials science on the other. From a statistician's perspective, a familiar thread within the densely woven fabric of physical theory, mathematical development, and experimental technique is the constant concern with finding simple and expedient models, frequently of a stochastic nature (van Kampen, 1981; Wax, 1954). Thus, although physical theory may in principle provide a complete microscopic description of the problem at hand, in practice the intractability of the mathematical development prevents useful predictions from being made. A classic illustration from physics is that of modeling the motion of a dust particle on the surface of a raindrop. The dust particle moves as a result of collisions with the water molecules. A typical raindrop will contain 10^{20} molecules whose deterministic equations of motion can be formulated as a Hamiltonian system. The solution of the equations is clearly impracticable. The motion of

the dust particle is therefore unresolved. However, a stochastic approximation can be derived, namely, Einstein's theory of Brownian motion, which provides good agreement with experimental observation. It should be noted that although the model fits the data on an observational scale, the trajectories of theoretical Brownian motion contradict physical laws, since infinite acceleration is required.

When chemistry is introduced, things become more complicated. Consider, for example, the effect of a pulse of radiation on the water droplet. Radiation creates chemically reactive species distributed throughout the droplet. Chemical reactions occur when reactive species approach each other as a result of molecular motion. As in the case of Brownian motion it is natural to look for a stochastic approximation to the reaction process, but here we must track the motion of a large number of atomically small reactive species. One approach is to use the heuristics of statistical mechanics, pioneered by Gibbs, to provide joint distributions for molecular positions and velocities. The progress of chemical events following radiation can then be treated as a stochastic process, but on an enormous state space. The stochastic behavior can be viewed as a manifestation of the chaotic character of the solutions of the nonlinear equations of motion. The skill of the physical chemist is to derive and validate parsimonious approximations of the reaction process while attempting to fit experimental data. There are therefore close analogies between the activities of physical chemists and the role of statisticians in applied science, in that the physical chemist must construct models that on the one hand are reasonably faithful to the laws of physics (the client) and on the other are amenable to mathematical manipulation and eventual experimental verification.

9.2 Diffusion Controlled Reaction

A classical theoretical problem in the analysis of the reaction rates in solutions is the modeling of diffusion controlled reactions. In these reactions, molecules of non-zero size diffuse and react instantaneously if they encounter one another. In a typical experiment, very reactive particles are produced randomly in space and essentially instantaneously by, for example, using a pulse of light. These particles diffuse and react with other species already in solution. The number or concentration of the reactive particles is observed in real time, by, for example, optical absorption methods. Computer simulations of liquids can provide insight into the reaction process, but the

results are necessarily subject to statistical error. A great deal of theoretical work has been devoted to deriving and validating good analytical approximations; see for example Balding and Green (1989) in the one-dimensional case. The original theory, owing to Smoluchowski (1917) fixed the coordinate system on a single particle and made the implicit approximation that all other particles diffuse independently in this frame of reference (Noyes, 1961). We will refer to this as the independent pairs (IP) approximation. While this is probably a good approximation for a central, slowly moving molecule surrounded by faster moving molecules (e.g., colloid coagulation), it is certainly not true for the converse problem (fast central molecule in a sea of static traps). In three dimensions the Smoluchowski theory gives the same result for both cases, namely, the survival probability of the central particle is

$$\Omega(t) = \Omega(0) \exp\left[-\int_0^t k(t')\,dt'\,c_s\right],$$

where c_s is the density of traps,

$$k(t) = 4\pi a D\left[1 + \frac{a}{\sqrt{\pi D t}}\right], \tag{9.1}$$

a is the encounter radius, and D is the diffusion coefficient of the mobile molecule(s).

For the latter case, where the Smoluchowski theory might be expected to break down because the intermolecular distances are highly correlated, the survival probability is related to the volume of the Wiener sausage, swept out by the diffusing molecule in the course of its trajectory. This is because the molecule will survive to time t if and only if there is no trap in the volume swept out, $V_a(t)$, and if the traps are distributed according to a Poisson point process with intensity c_s, the survival probability of a given trajectory is $\exp[-V_a(t)\,c_s]$.

The observed survival probability will be the expectation of this random variable

$$\Omega = E[\exp[-V_a(t)\,c_s]].$$

Donsker and Varadhan (1975) have obtained precise asymptotic results for expectations of this kind. They show

$$\lim_{t\to\infty}\left[(Dt)^{d/(d+2)}\,c_s^{2/(d+2)}\,\ln\Omega\right] = -\kappa_d,$$

where d is the dimension and κ_d is a positive constant. This is very different from the simple exponential decay at long times predicted by the

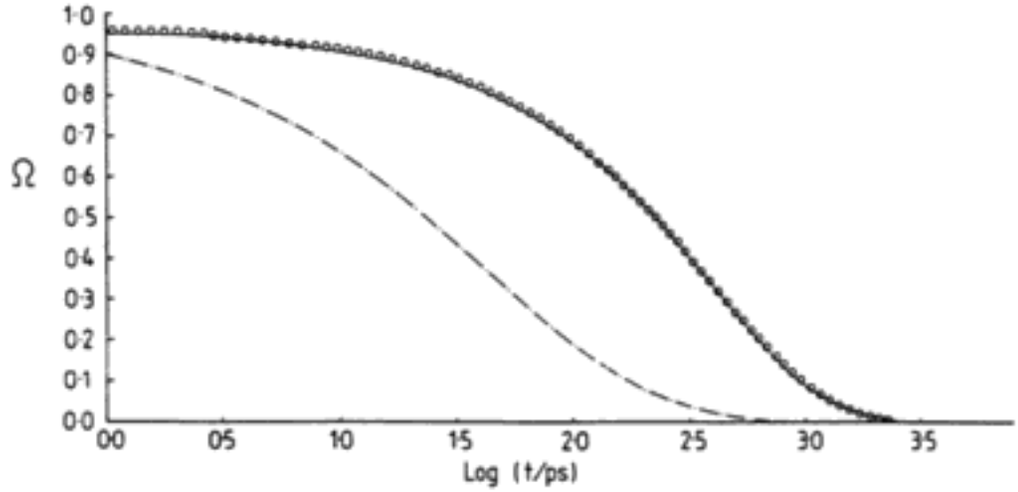


FIGURE 9.1: Survival probability Ω for a particle diffusing in a sea of static traps. Comparison of Monte Carlo simulation (o o o), Donsker-Varadhan asymptotics (– – · – –), and the IP approximation (—).

Smoluchowski theory. Since there is no obvious analytic way to assess the time scale on which the asymptotic behavior will be found, we have developed a simulation technique for this purpose. Early simulation results indicate much better agreement with the Smoluchowski theory than with the Donsker-Varadhan result. The reasons for this observation are not clear at present. See Figure 9.1.

9.2.1 Radiation Spurs

Diffusion controlled reactions are the fastest reactions that occur in solution. Experimental observations of the rate of reaction contain information about the initial spatial distribution of the reactive species. A substantial amount of research has been devoted to the analysis of radiation tracks. If radiation interacts weakly with the liquid (e.g., fast β-particles), the track consists of small isolated spurs, which are clusters of highly reactive particles; the spurs subsequently relax by diffusion and within each spur the particles react with each other on encounter.

The problems of describing reactions in clusters can be illustrated by reference to a number of model systems with simplified chemistry.

The Two-Species Spur

The simplest system contains two types of particles, A and B, which react on encounter to form products AA, AB, or BB. The particles are identical in all but name and have identical spatial distributions. The classical method of dealing with such a system is to make a continuum approximation. The two spatially dependent concentrations (which are identical) obey macroscopic continuum equations of the form

$$\begin{aligned} \frac{\partial}{\partial t} C_A &= D_A \nabla^2 C_A - k^0_{AA} C_A^2 - k^0_{AB} C_A C_B \\ \frac{\partial}{\partial t} C_B &= D_B \nabla^2 C_B - k^0_{BB} C_B^2 - k^0_{AB} C_A C_B \,, \end{aligned}$$

where the first terms on the right-hand sides represent diffusive spreading of the concentration profiles, and the remaining terms represent local depletion by reaction; the rate coefficients k are given by Smoluchowski's theory (*cf.* equation (9.1)). Although these equations are perfectly satisfactory when applied to macroscopic problems, they are not appropriate when dealing with the small number of particles in a spur of finite extent. There are two reasons for this: (1) the small number of particles in the spur ought to be treated as a discrete variable, and (2) the Smoluchowski rate constant is appropriate for a particle initially surrounded by a homogeneous Poisson field of reactants as opposed to the highly clustered distribution in a spur. The necessity for a correct stochastic theory is easily demonstrated in this simple system. If there are initially N_A particles of type A and N_B particles of type B, then simple probabilistic arguments permit us to show that the

TABLE 9.1: Typical Product Yield Ratios

N_A	N_B	AA	AB	BB
1	1	0	1	0
2	2	1	4	1
3	3	1	3	1
4	4	3	8	3
10	10	9	20	9
∞	∞	1	2	1

expected product yields, for all times, are in the ratio

$$N_{AA} : N_{AB} : N_{BB} = \frac{1}{2}N_A(N_A-1) : N_A N_B : \frac{1}{2}N_B(N_B-1).$$

Typical ratios are given in Table 9.1. The continuum approximation always predicts a ratio of 1:2:1 since it corresponds to the case of infinitely many particles. The independent pairs approximation can be used to provide a stochastic theory of spur kinetics. If the state of the spur at time t is labelled with M, N where M is the number of A particles and N is the number of B particles, then $P_{MN}(t)$, the probability of being in state M, N, satisfies the following forward equations:

$$\begin{aligned}\frac{d}{dt}P_{MN}(t) &= \frac{1}{2}[(M+2)(M+1)P_{M+2,N} - M(M-1)P_{MN}]\lambda_{AA}(t) \\ &+ \frac{1}{2}[(N+2)(N+1)P_{M,N+2} - N(N-1)P_{MN}]\lambda_{BB}(t) \\ &+ [(M+1)(N+1)P_{M+1,N+1} - MNP_{MN}]\lambda_{AB}(t),\end{aligned}$$

where the λ's are the reaction rates for isolated pairs of particles whose initial spatial separation is equivalent to that in a cluster. These equations can be solved analytically in special cases, for example, when the particles are identical. In general, though, they must be solved numerically. Comparisons between this approximation, the continuum approximation, and the full Monte Carlo simulation of sample trajectories are given in Figure 9.2, taken from Clifford *et al.*, (1987a). It is evident that the stochastic independent pair (IP) model is in very good agreement with the simulations.

The Ionic System

The two species system can be generalized by including long-range forces between the particles, such as the Coulombic force between ions. Such forces act attractively between A and B particles, but like particles repel each other. The form of the λ's is modified because the survival probability of a pair depends on the force between the particles. The λ's must now be calculated approximately if the forces are weak (Clifford *et al.*, 1987b) or numerically if they are strong (Green *et al.*, 1989). The introduction of forces would be expected to make the IP approximation worse, because of complicated interactions in the many-body system. Comparisons of the IP approximation and the results of full simulations of sample trajectories are shown in Figure 9.3. It is seen that even when the forces are so strong that

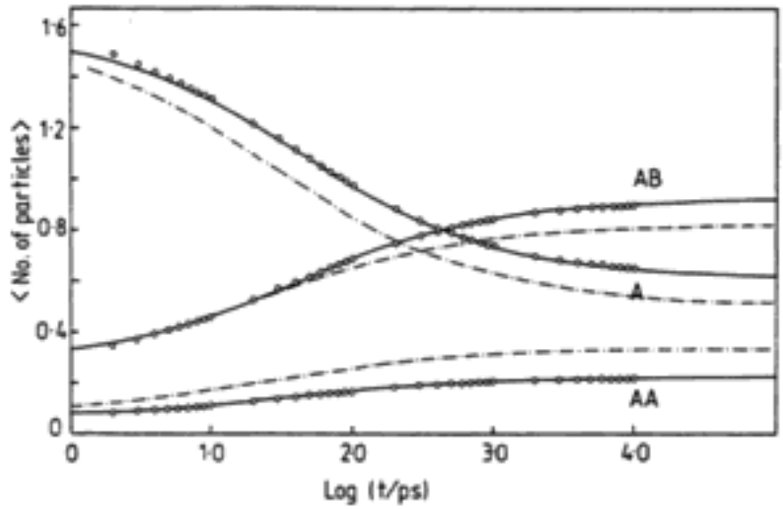


FIGURE 9.2: Kinetics of two species spur. Initial configuration spherical Gaussian. Monte Carlo simulation (o o o); IP approximation (—); continuum approximation (– – · – –). Reprinted, by permission, from Clifford *et al.* (1987a). Copyright © 1987 by the Royal Statistical Society.

the AA and BB encounters are effectively impossible, the IP approximation is still remarkably accurate.

The Scavenging System

The simplest such system is

$$\begin{aligned} A + A &\rightarrow AA \\ A + S &\rightarrow AS, \end{aligned} \tag{9.2}$$

where the species A is clustered in a spur, whereas the species S exists in large numbers uniformly distributed over an extended volume. There is competition between intraspur recombination, the AA interaction, and scavenging, the AS interaction. The relative abundance of the ultimate yields of AA and AS provides information about the scavenging process. In the IP model the forward equation becomes

$$\frac{d}{dt}P_N = \frac{\lambda(t)}{2}[(N+2)(N+1)P_{N+2} - N(N-1)P_N] + k_{AS}^0 C_S[(N+1)P_{N+1} - NP_N],$$

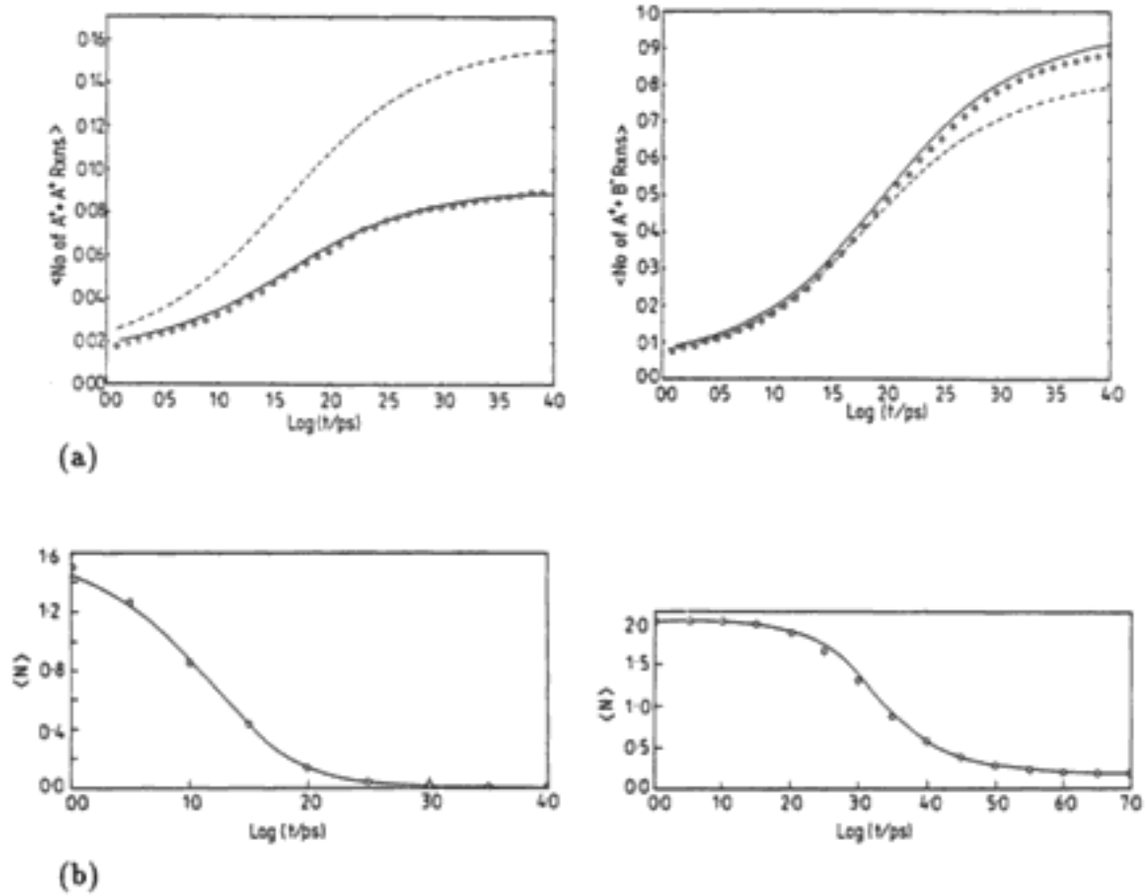


FIGURE 9.3: (a) Ionic reactions in high permittivity solvents. Average number of reactions in a spur containing two ion-pairs: A^+, B^-. Monte Carlo simulation (o o o), simulation using the IP approximation (—), and continuum theory (- - -). Reprinted, by permission, from Clifford *et al.* (1987b). Copyright © 1987 by the American Chemical Society. (b) Ionic reactions in low permittivity solvents. Average number of surviving pairs in a spur containing two ion-pairs: A^+, B^-. Left panel: homogeneous dispersion; right panel: heterogeneous dispersion. Monte Carlo simulation (o o o) and simulation using the IP approximation (—). Reprinted, by permission, from Green *et al.* (1989). Copyright © 1989 by the American Chemical Society.

where $k^0_{AS}(t)$ is given by the Smoluchowski theory (*cf.* equation (9.1)) and N is the number of A particles remaining. The continuum approximation gives the equation

$$\frac{\partial}{\partial t}C_A = D\nabla^2 C_A - k^0_{AA}(t)C_A^2 - k^0_{AS}(t)C_A C_S\,.$$

Typical results are shown in Figure 9.4. Again, the continuum model fails to reproduce the results of a full Monte Carlo simulation, and the IP approximation is superior.

9.3 Computer Simulation of Liquids

Although a full description of a liquid system must be quantum mechanical, almost all liquids (except those containing very small molecules such as helium) can be described adequately using a completely classical deterministic model (McQuarrie, 1976). If we tag and follow one molecule in a computer simulation of such a deterministic system, its motion appears random. If several molecules are followed, their spatial configuration evolves as a spatial point process, marked by the individual velocities. What we would like to do is to find stochastic rules that indicate where the molecules will be, and how fast they will be traveling, as time goes on. For example, in radiation chemistry, where the effect of radiation is to create reactive species distributed throughout the liquid, we are interested in the time taken for such species to encounter and react with each other.

9.3.1 Some History

In physics and chemistry, classical liquids are simulated by one of two techniques. The first relies on the laws of mechanics to provide the equations of motion of a finite system of interacting molecules (Goldstein, 1980). This is known as the molecular dynamics approach. The second technique makes use of statistical arguments that were originally given by Gibbs. This is generally called the Monte Carlo approach. A rigorous and detailed account of the statistical treatment of mechanics can be found in Ruelle (1969). The book *Computer Simulation of Liquids* by Allen and Tildesley (1987) contains a comprehensive history of simulation methodology from the first computer experiments through to the latest ideas. Numerical simulation is the technique most widely used in recent years to study the properties

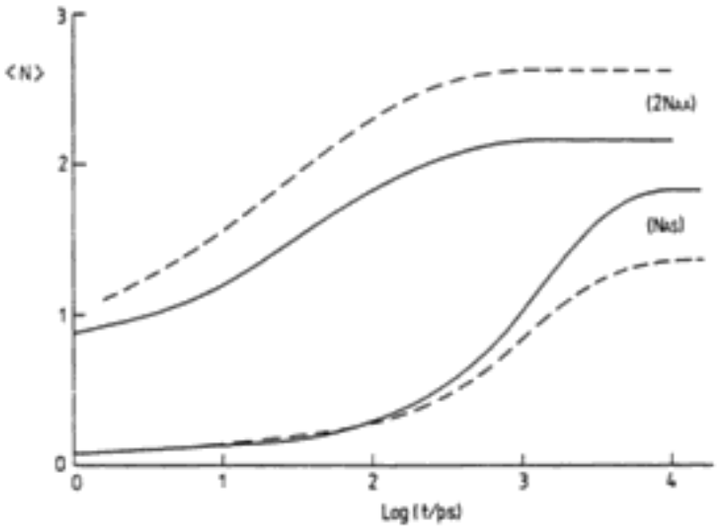


FIGURE 9.4: Average number of reactions in a spur containing four radicals A, A, A, A with scavenging by S molecules. Initial distribution spherical Gaussian. IP approximation (—); continuum theory (- - -). Reprinted, by permission, from Clifford *et al.* (1987a). Copyright © 1987 by the Royal Statistical Society.

of liquids. It is now an extremely large research area in both physics and chemistry, with many hundreds of research groups involved.

The classical mechanics of a system of N structureless molecules is specified by a Hamiltonian $\mathcal{H}$, which is the sum of the kinetic and potential energy of the system: $\mathcal{H}(\mathbf{r},\mathbf{p}) = \mathcal{K}(\mathbf{p}) + \mathcal{V}(\mathbf{r})$. Hamilton's equations of motion are (Goldstein, 1980):

$$\begin{aligned} \dot{\mathbf{p}} &= -\nabla_r \mathcal{V} \\ \dot{\mathbf{r}}_\mathbf{i} &= \mathbf{p}_\mathbf{i}/m_i\,, \end{aligned} \tag{9.3}$$

where $\mathbf{p}$ and $\mathbf{r}$ are the vectors of momenta and positions of the molecules. The kinetic energy is given by

$$\mathcal{K}(\mathbf{p}) = \sum_{i=1}^{N} p_i^2/2m_i\,, \tag{9.4}$$

where m_i is the mass of molecule i and p_i is the magnitude of its momentum. The potential energy $\mathcal{V}(\mathbf{r})$ depends on the positions and orientations of the

particles. It is usually sufficient to assume that $\mathcal{V}$ only depends on the interactions of particles in pairs, and to use a spherical average of the pair potential, although the pair potential may have to be modified to correct for higher order effects. In the absence of an external field, the potential energy then becomes

$$\mathcal{V} = \sum_i \sum_{j>i} \nu_2(r_{ij}),$$

where r_{ij} is the distance between particles i and j.

The computer simulation of liquids and gases was initiated by Metropolis *et al.* (1953), who used Monte Carlo methods to simulate the Gibbs equilibrium distribution of molecular configurations. Their aim was to derive values for stationary (i.e., equilibrium) physical properties such as expected energy and expected pressure. Early work was concerned with the case of a hard sphere potential, $\nu_2(r) = \infty$ for $r < \sigma$ and $\nu_2(r) = 0$ otherwise.

In order to obtain dynamic properties, Alder and Wainwright (1959) developed a method by which the simultaneous equations of motion for many molecules are solved numerically. They illustrated their method by simulations using both hard sphere and square well potentials. Their paper is the first example of molecular dynamics simulation. Simulations using a realistic potential were made by Rahman (1964).

In particular, Rahman estimated the pair-correlation function

$$g(r) = \frac{V}{N} \frac{n(r, \Delta r)}{4\pi r^2 \Delta r}, \tag{9.5}$$

where $n(r, \Delta r)$ is the time averaged number of molecules at a distance between r and $r + \Delta r$ from a given molecule, and N/V is the average density of molecules. He showed that the system has spatial structure that decays slowly over time.

9.3.2 Sampling From Configuration Space

Let us consider a system of N molecules in three-dimensional space, subject to a potential as previously described. We can think of the simultaneous positions and momenta, or equivalently positions and velocities, as coordinates in $6N$-dimensional space, $\mathcal{X}$. We denote a point in this space by x. We call $\mathcal{X}$ the state space and refer to x as a state; in the physics literature, $\mathcal{X}$ is called the phase space. Let $\mathcal{F}$ be a function of x. We refer to $\mathcal{F}(x)$ as an instantaneous evaluation of the property $\mathcal{F}$. For example, $\mathcal{K}(\mathbf{p})$ in equation (9.4) is an instantaneous evaluation of kinetic energy.

A large system can be thought of as the union of many smaller systems. At any instant of time, each small system will have a particular state. A macroscopic property of the large system is an average of the property evaluated over the subsystems. There are two basic ways in which the state space can be explored. The first is to build a dynamic description of molecular motion that will move through the state space according to acceptable physical principles. This is the approach of molecular dynamics. As noted earlier (§9.3.1), the equations of motion are those considered by Hamilton in classical mechanics. The required average is then taken over a succession of times for a single small system; arguing that if the time period is sufficiently large a representative sample of configurations will be obtained.

The second method of sampling states relies on the validity of Gibbs's probabilistic analysis of large mechanical systems. The method involves Monte Carlo simulations. The state of each small system is treated as a random variable drawn from the Gibbs distribution, which is constructed to have maximum entropy subject to certain constraints, giving an explicit form for the density of the distribution. The task of sampling the state space for this method is reduced to that of choosing a random sample, $x_1, x_2, \ldots, x_n$, from a specified density $p(x) : x \in \mathcal{X}$. In typical applications the form of the density lends itself to sampling by the Metropolis method. The required estimate of the macroscopic property is then given by $\overline{\mathcal{F}} = (\mathcal{F}(x_1) + \cdots + \mathcal{F}(x_n))/n$. Since this can be interpreted as an average taken over a number of subsystems, it is usually referred to as a space-average.

Both Monte Carlo and molecular dynamics simulations can be used to sample from the equilibrium distribution of a many-particle system. In principle it is therefore possible to test empirically whether Gibbs's theory agrees with the results of molecular dynamics simulation. In practice, simulations must be run for a large number of time steps until equilibrium is attained. The development of tests for spatial point patterns is an active research area in statistics (Diggle, 1983; Ripley, 1987; Besag and Clifford, 1989). One of our aims is to link methods used by probabilists and statisticians in the study of spatial processes with methods used by physicists and chemists in their study of liquids.

9.3.3 A Typical Computer Experiment

Lynden-Bell *et al.* (1986) investigated the behavior of carbon tetrafluoride CF_4 near its triple point by carrying out a Molecular Dynamics simulation of 256 molecules in a cubic box with periodic boundary conditions, using a variant of the Lennard-Jones potential. They were interested in the struc-

ture of the velocity autocorrelation function, i.e., the empirical correlation between the velocity of a molecule in a particular direction and its velocity in the same direction at some time in the future. For typical liquids, the velocity autocorrelation is strongly positive at short lags, since molecules tend to continue with the same velocity, negative at moderate lags, since molecules eventually bounce off their neighbors, and then slowly approach zero as the lag tends to infinity.

In order to explain certain anomalies in the behavior of the velocity autocorrelation function, Lynden-Bell *et al.* conjectured the existence of "local cages" of molecular configurations. A molecule is said to be in a local cage if its motion is restricted by the proximity of neighboring molecules. They first estimated the density of $\cos\phi(\tau)$, where $\phi(\tau)$ is the angle between the velocity of a molecule at time t, and the velocity of the same molecule at the later time $t+\tau$. They observe that at moderate values of τ the estimated density is approximately uniform. Plotting the height of the estimated density as a function of time τ, they notice that the shape of the curve closely follows that of the velocity autocorrelation function. Stratifying the molecular trajectories by initial kinetic energy, Lynden-Bell *et al.* then repeat their analysis, but for initially fast and slow molecules separately. The results are different for the two groups. The structure is the same, but the magnitude of the effect is much higher for fast molecules. They suggest that high-energy molecules rattle back and forth in cages, while slower molecules diffuse. Lynden-Bell *et al.* finish by looking at the velocity autocorrelation function for the two stratified groups of molecules. They show that the velocity autocorrelation function of slower molecules has, surprisingly, a more pronounced negative portion than that of the faster molecules.

In the simplest statistical mechanical model of a liquid, molecules have independent velocities chosen from the Maxwell-Boltzmann distribution, i.e., multivariate normal. In Atkinson *et al.* (1990), it is shown that, with the possible exception of the last result, all the results of Lynden-Bell *et al.* are consistent with a simple description in which each molecule moves along a random trajectory in such a way that the velocity components in three fixed orthogonal directions are independent Gaussian processes. It is not necessary to propose the existence of local cages.

To see this, notice that $\cos(\phi)$ is essentially a correlation coefficient for three pairs of velocity components. With the Gaussian assumptions above, the density of $R = \cos(\phi)$ is then given by

$$f_R(r,\rho) = \frac{2}{\pi}(1-\rho^2)^{3/2}\sum_{s=0}^{\infty}\Gamma^2\left(\frac{s+3}{2}\right)\frac{(2\rho r)^s}{s!}. \tag{9.6}$$

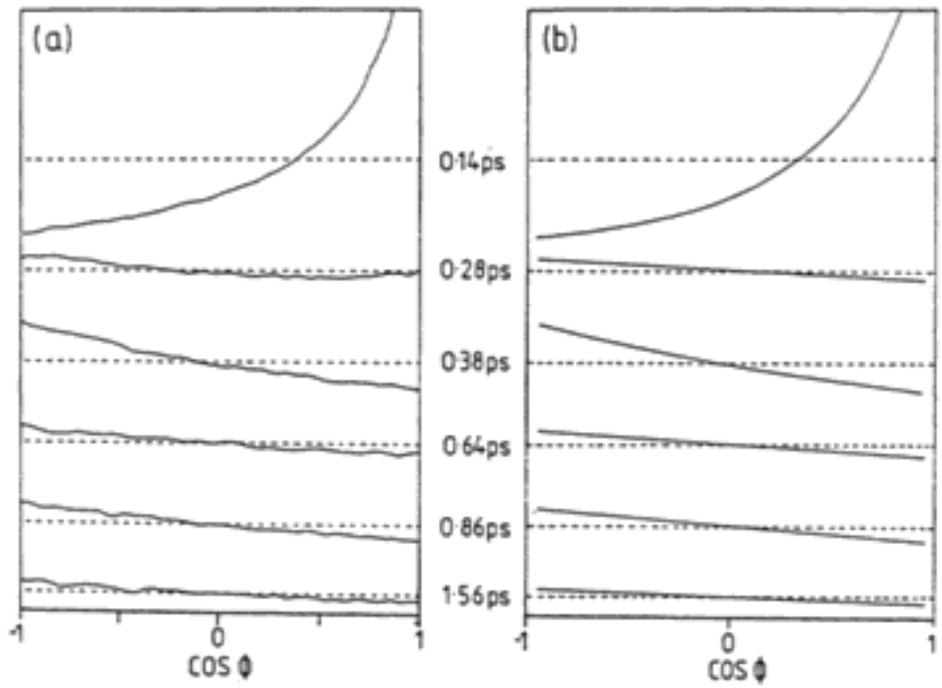


FIGURE 9.5: Histogram of $cos(\phi)$ after various time lags: (a) simulated (Lynden-Bell *et al.*, 1986) and (b) calculated from equation (9.6). The curves are displaced, as in Lynden-Bell *et al.*; in each case the horizontal dashed line represents a uniform density. Reprinted, by permission, from Lynden-Bell *et al.* (1986). Copyright © 1986 by Taylor and Francis, Ltd.

where $\rho = \rho(\tau)$ is the theoretical counterpart of the empirical velocity autocorrelation. In the discussion of their results, Lynden-Bell *et al.* observe that at moderate lags the distribution of R is nearly uniform. If R has a uniform distribution, then the molecule is equally likely to be moving in any direction at this time lag, regardless of its initial velocity. The authors also observe that, at longer time lags, the distribution of R becomes skewed, indicating that the particle's velocity is opposite to the original direction. Lynden-Bell *et al.* consider these results to be paradoxical, however, this is precisely what is predicted by the form of the theoretical density. More importantly, the theoretical predictions give good qualitative fits to the computer-simulated results. See Figure 9.5.

To throw further light on the causes of the nearly uniform distribution of R at the time lag for which the velocity autocorrelation function is zero,

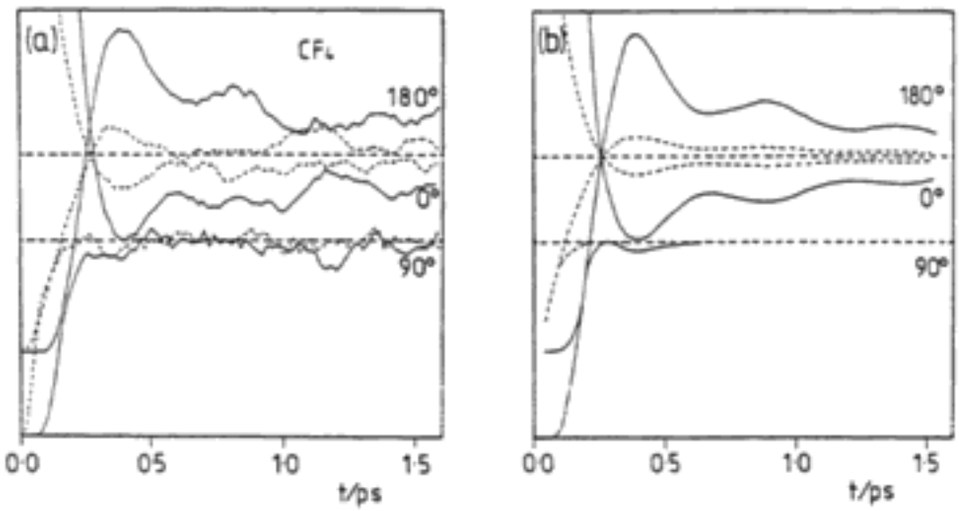


FIGURE 9.6: Time-dependence of the probability density of $\cos(\phi)$ at $\phi = 0°$, $90°$, and $180°$, conditioned on initial kinetic energy. (a) Molecular dynamics simulation (Lynden-Bell *et al.*, 1986) and (b) calculated from equation (9.7). The continuous lines refer to the 7.2% of particles with the highest kinetic energy, and the dashed lines to the 12% with the lowest energy. Reprinted, by permission, from Lynden-Bell *et al.* (1986). Copyright © 1986 by Taylor and Francis, Ltd.

Lynden-Bell *et al.* stratified the trajectories of molecules by their kinetic energy at time zero. Their idea, as noted earlier, was that the observed effect was due to a balance between high-energy molecules rattling back and forth in local cages, while low-energy molecules were diffusing through the whole space.

Taking $\mathbf{v}$ to be the velocity of a molecule at time $t = 0$, and writing $|\mathbf{v}| = \alpha$, the conditional density of R given α obtained under the Gaussian assumptions is

$$f_R(r|\alpha,\rho) = \frac{1}{\sqrt{\pi}} e^{-\alpha^2\rho^2/2(1-\rho^2)} \sum_{s=0}^{\infty} \Gamma\left(\frac{s+3}{2}\right) \frac{1}{s!} \left(\frac{\sqrt{2}\alpha\rho r}{\sqrt{1-\rho^2}}\right)^s . \qquad (9.7)$$

As is shown in Figure 9.6, the fit with the data of Lynden-Bell *et al.* is again excellent.

The final observations of Lynden-Bell *et al.* are difficult to reproduce. We have tried unsuccessfully to confirm these results using a molecular dynamics program. The program was run many times using different numbers

of molecules (32, 108, 256) and a variety of different computers (Vax mainframe, MassComp workstation, Sun workstation). Comparisons of double and single precision calculations were made, and the effect of optimizing compilers was also examined. Different computers, different precision, and the use of an optimizing compiler all substantially changed the configurations and velocities of the molecules. At lags for which the velocity autocorrelation is positive, there was however, a consistent effect with the slow molecules having a slightly greater positive autocorrelation at a fixed time lag than the fast molecules. This is not consistent with the simple hypothesis of Gaussian velocity components.

9.4 Discussion and Future Directions

The main areas of investigation in physical chemistry can be classified as follows:

1. physical properties of matter in equilibrium,
2. dynamical and transport properties of matter,
3. properties of atoms and molecules,
4. statistical mechanics linking the above,
5. energetics and dynamics of chemical reactions, and
6. complex systems.

9.4.1 Physical Properties of Matter in Equilibrium

The properties of bulk matter are reasonably well understood on a qualitative level, and if the substance is made up of simple molecules or atoms, such as the liquid inert gases, numerical simulations of small systems, based on the known intermolecular forces and involving of the order of a thousand molecules, are quite successful in reproducing the observed properties. Molecular dynamics simulations become more complicated when the molecules are neither spherical nor rigid. A great deal of work is still in progress in this area. A typical simulation is a realization of a chaotic spatial temporal process, involving the position and velocity of several hundred molecules for perhaps 10,000 time steps. There are a number of outstanding statistical problems in the design and analysis of these computer experiments. In particular, it is of interest to determine when a simulation has

reached equilibrium. This problem is complicated, in small systems, by the effects of phase transition.

An additional complication, which is receiving attention, is the inclusion of quantum mechanical effects in liquids such as helium. The properties of polymers and biological compounds are also the subject of research activity. Attention has turned in recent years to the study of interfaces between different phases of matter. Monte Carlo and molecular dynamics simulations have concentrated on bubbles and droplets, and a number of experimental techniques have been devised for studying the gas-solid, gas-liquid, and solid-liquid interfaces. This work has relevance to the understanding of catalytic and electrochemical processes.

Recent advances in experimental metallurgy have enabled detailed analyses of the atomic structure of metallic alloys to be carried out. Data are becoming available that record the position, subject to quantifiable error, of up to 60% of the atoms in small three-dimensional regions of a given sample. The analysis of this enormous data base, in particular the task of reconstructing the atomic lattice from partial observations, is a challenging statistical problem, which can be approached by combining simulated annealing as an optimization technique and realistic annealing as a description of the aging process in the atomic lattice.

9.4.2 Dynamical and Transport Properties of Matter

The transport properties of matter, such as viscosity, thermal conductivity, and diffusion, involve transfer of energy or momentum from one molecule to another during a collision. The theoretical relationship between the transport properties of gases and the intermolecular forces has been known for a long time. Recently, physical chemists have attempted to tackle the inverse problem of estimating the intermolecular forces from detailed experimental observations of the viscosity-temperature curve. There is increasing interest in this type of statistical exercise, in particular there are unresolved questions of identifiability.

In solids and liquids, it has become clear that there is a wealth of information about the dynamics of the molecules from light-scattering and neutron-scattering experiments, but the information is in a form that is difficult to extract and interpret. One promising line in this respect is the use of computer simulations as idealized experiments, both to develop tools for the analysis of data and to construct Monte Carlo estimates of dynamical quantities. Applications to more complex systems include the study of the motion of polymers in solution, with particular reference to enzyme activity.

9.4.3 Properties of Atoms and Molecules

A great deal of physical chemistry is involved with investigating the properties of isolated atoms and small molecules. Spectroscopy of these species, using radiation ranging from radio frequency through infrared and the visible spectrum to X-rays, provides basic information about the energies of the accessible quantum states and the symmetries of the corresponding wavefunctions, molecular size and geometry, nuclear spin, dipole moments, magnetic moments, polarizabilities, and so on. Spectroscopy can therefore be used to test the predictions of the great variety of quantum mechanical approximations employed to calculate molecular properties.

9.4.4 Statistical Mechanics

The fundamental molecular properties and their interactions as measured by spectroscopists are the data required by statistical mechanics for the description of bulk matter. Statistical mechanics is the central unifying theory of physical chemistry as it relates the properties of isolated molecules with the bulk. The reconciliation of the statistical mechanical approach with modern theories of chaos in dynamical systems is a problem of outstanding interest to mathematicians. Large deviation theory was used in early attempts to provide a probabilistic interpretation. Recent work on infinite particle systems, has given insight into the phenomenon of phase transition in the classical Gibbs distributions of statistical mechanics. Some of the important applications have been covered in previous sections.

9.4.5 Dynamics of Chemical Reactions

Chemical reactions occur when molecules are transformed by the rearrangement of electrons and nuclei. Physical chemistry concerns itself not only with the energetics of chemical reaction, but also with their rates and with the distribution of energy in the products. Gas phase chemical reactions generally occur as a result of simple collisions between isolated molecules. The classical theory of these processes has recently been revolutionized by experiments in which molecules are produced in collimated mono-energetic beams, which allow many of the parameters of the colliding particles (e.g., speed, quantum state, and orientation) to be fixed, and the energy and angular distributions of the products to be analyzed, thus giving very detailed information about the collision dynamics and the flow of energy between and within molecules.

Classical descriptions of these dynamics have been proposed, which show regions of regular behavior and regions of chaos. It is still not clear how such phenomena will transfer to quantum mechanical descriptions. Reactions in solution are not understood in such a detailed way, although there are quantum mechanical theories of electron transfer reactions and the like. An additional problem of reactions in liquids is that particles diffuse slowly through the liquid and can only react on encounter. There are therefore two limiting cases of reaction: diffusion control, where the rate depends only on the transport through the solution and the rate of encounter; and activation control, where reactive encounters are rare, so that many encounters will take place before reaction can occur. Theories of the former type of reaction have been developed for a long time, but are only now being tested, by numerical solution of the associated stochastic differential equations. For activation-controlled reactions more detailed modeling of the encounter complex is required.

9.4.6 Complex Systems

As well as the fundamental research described above, physical chemistry is involved with description of more complex systems, particularly the evolution of these systems. Frequently, the problems of interest have important spatial aspects that have to be taken into account.

Atmospheric Chemistry—Depletion of Ozone Layer

A realistic model must incorporate chemistry and transport in the atmosphere. It also requires an understanding of interfaces such as those between the air and cloud droplets and ice crystals, which act as sinks for active chemicals.

Combustion

Since the 1950s there has been a series of revolutionary changes in explosives technology, which has resulted in safer but slower-reacting explosive products. In order to maintain product performance, much attention must now be given to understanding the detonation process. The initiation and establishment of the critical conditions for detonation have been subject to little detailed realistic investigation; although of course there are obvious analogies with the stochastic theories of spatial epidemics. The necessary cooperative interaction between small numbers of initiating sites suggests

that a treatment based on macroscopic deterministic approximations may be inappropriate: the number of reacting species being just too small for the averaging implicit in standard treatments. Here again, there is the possibility of nonlinear kinetics producing oscillations and chaotic behavior.

Radiation Chemistry

When a liquid is exposed to ionizing radiation, reactive species are generated in an initially localized spatial distribution. For low linear energy transfer (LET) radiation, isolated clusters, called spurs, are formed. A significant proportion of the chemical reaction following radiation occurs on a short time scale, when the localized distribution has not yet relaxed by diffusion. The chemical process can be treated successively using stochastic methods. Currently, there is interest in extending these results to the products of higher LET radiation, which are formed along linear tracks.

Surface Kinetics and Electrochemistry

The theory of surface kinetics seeks to explain effects such as etching, dissolution of crystals and the formation of corrosion pits. Stochastic models of growth and dissolution have been studied. An interesting class of problems concerns the description of flocculation processes, in which the growth of an aggregrate is limited by diffusion from the surrounding medium.

Electrochemistry is concerned with the understanding of chemical effects produced at electrodes. Recent debates about the feasibility of cold fusion, hinged on the estimation of the probability of favorable molecular encounters at an electrode. Electrochemistry is clearly a branch of physical chemistry in which probabilistic calculations play an important role.

Bibliography

[1] Alder, B. J., and T. E. Wainwright, Studies in molecular dynamics, I. General method, *J. Chem. Phys.* **31** (1959), 459–466.

[2] Allen, M. P., and D. J. Tildesley, *Computer Simulations of Liquids*, Clarendon Press, 1987.

[3] Atkinson, R. A., P. Clifford, and N. J. B. Green, Correlation effects in simple liquids, to appear in *Mol. Phys.*, (1990).

[4] Balding, D. J., and N. J. B. Green, Diffusion-controlled reactions in one dimension: Exact solutions and deterministic approximations, *Phys. Rev. A* **40** (1989), 4585–4591.

[5] Besag, J. E., and P. Clifford, Generalised Monte Carlo tests, *Biometrika* **76** (1989), 633–642.

[6] Clifford, P., N. J. B. Green, and M. J. Pilling, Statistical models of chemical kinetics in liquids, *J. R. Stat. Soc., B* **49** (1987a), 266–300.

[7] Clifford, P., N. J. B. Green, M. J. Pilling, and S. M. Pimblott, Stochastic models of diffusion controlled ionic reactions in radiation induced spurs: (i) High permittivity solvents, *J. Phys. Chem.* **91** (1987b), 4417–4422.

[8] Diggle, P. J. *Statistical Analysis of Spatial Point Patterns*, Academic Press, 1983.

[9] Donsker, M. D., and S. R. S. Varadhan, Asymptotics for the Wiener sausage, *Commun. Pure Appl. Math.* **28** (1975), 525–565.

[10] Goldstein, H., *Classical Mechanics*, 2nd Ed., Addison-Wesley, 1980.

[11] Green, N. J. B., M. J. Pilling, S. M. Pimblott, and P. Clifford, Stochastic models of diffusion controlled ionic reactions in radiation induced spurs: (ii) Low permittivity solvents, *J. Phys. Chem.* **93** (1989), 8025–8031.

[12] Hammersley, J. M., and D. C. Handscomb, *Monte Carlo Methods*, Methuen and Co., Ltd., 1964.

[13] Lynden-Bell, R. M., D. J. C. Hutchinson, and M. J. Doyle, Translational molecular motion and cages in computer molecular liquids, *Mol. Phys.* **58** (1986), 307–315.

[14] McQuarrie, D. A., *Statistical Mechanics*, Harper and Row, New York, 1976.

[15] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, *J. Chem. Phys.* **21** (1953), 1087–1092.

[16] Noyes, R. M., Effects of diffusion rates on chemical reactions, *Prog. React. Kinet.* **1** (1961), 129–160.

[17] Rahman, A., Correlations in the motion of atoms in liquid argon, *Phys. Rev.* **136** (1964), 405–411.

[18] Ruelle, D., *Statistical Mechanics: Rigorous Results*, W. A. Benjamin, Inc., New York, 1969.

[19] Ripley, B. D., *Stochastic Simulations*, John Wiley & Sons, New York, 1987.

[20] Smoluchovski, M. von, Mathematical theory of the kinetics of the coagulation of colloidal particles, *Z. Phys. Chem.* **92** (1917), 129–168.

[21] van Kampen, N. G., *Stochastic Processes in Physics and Chemistry*, North Holland, New York, 1981.

[22] Wax, N., *Selected Papers on Noise and Stochastic Processes*, Dover, New York, 1954.

10

Stereology

Adrian Baddeley
Centre for Mathematics and Computer Science

10.1 Introduction

Stereology is a spatial version of sampling theory. It was initially developed in biology and materials science as a quick way of analyzing three-dimensional solid materials (such as rock, living tissue, and metals) from information visible on a two-dimensional plane section through the material. It now embraces all geometrical sampling operations, such as clipping a two-dimensional image inside a window, taking one-dimensional linear probes, or sampling a spatial pattern at the points of a rectangular grid. Applications include anatomy, cell biology and pathology; materials science, mineralogy and metallurgy; botany, ecology and forestry; geology and petrology; and image processing and computer graphics.

It is not the aim of stereology to reconstruct an entire three-dimensional object. Typically, only a few sections or samples are taken, and their spatial position is not recorded. Further it is usually impossible to model the three-dimensional structure explicitly. Instead, stereology uses simple non-parametric techniques to estimate "geometrical parameters" such as volume and surface area. Simplicity is the key word; the estimation relies only on fundamental geometric facts and classical sampling theory. As a result, stereological methods are almost "assumption free," and are applicable in many different sciences.

Applications and general concepts are described in §10.1. Section 10.2 is a more detailed statistical treatment. Section 10.3 describes newer discoveries and research problems.

10.2 Concepts and Applications

10.2.1 Information from Lower-Dimensional Samples

In 1847 the French mineralogist Delesse published a revolutionary method for measuring the mineral content in a sample of rock [22]. Instead of crushing the rock to separate the different minerals, one simply cuts a plane section through it. Delesse had realized that the *proportion by volume* of a particular mineral can be estimated from its *proportion by area* visible in the section.

Model the rock as a set $X \subset \mathbf{R}^3$ containing a subset $Y \subset X$, the mineral phase of interest. The objective is to estimate the volume fraction

$$V_V = \frac{V(Y)}{V(X)},$$

where $V(\cdot)$ denotes volume. Let T denote a plane in three dimensions, so that $X \cap T$ is the plane section of the rock, and $Y \cap T$ is that part of the section occupied by the mineral phase. Delesse's method estimates V_V from the area fraction

$$A_A = \frac{A(Y \cap T)}{A(X \cap T)},$$

where $A(\cdot)$ denotes area in the two-dimensional section.

This is like a survey sampling problem: X represents the "population" and $X \cap T$ the "sample" from which we want to estimate a population parameter V_V. Astoundingly, A_A is an *unbiased* estimator,

$$V_V = \mathbf{E}A_A \tag{10.1}$$

(under the right sampling conditions), *without any assumptions* about the shape of X and Y. This follows from the basic geometrical fact that the volume of a three-dimensional object is the integral of the areas of its two-dimensional plane slices. Here $\mathbf{E}$ denotes expectation with respect to a suitable random sampling design (not the most obvious one); we give details in §10.2.

Delesse's method was later simplified [74] by placing a grid of parallel lines over the plane section, with the aid of a transparent sheet. Then area fractions A_A can be estimated from *length fractions* L_L, i.e., the relative lengths of the mineral phases on the line grid. This was simplified even further by Glagoleff [23] who showed that if we superimpose a rectangular grid of points over the section plane, the area fraction A_A can be estimated

from the proportion P_P of grid points that "hit" (lie over) the mineral phase. In both cases the estimators are unbiased.

Demonstrate this with a "party trick." Take a sheet of graph paper ruled with (say) thin lines every 1 mm and thick lines every 5 mm. Cut out an arbitrary shape. Ask someone to determine the area of the cutout by counting all the 1 mm squares. Meanwhile estimate the area stereologically by counting the 5 mm crossing points that are visible on the paper, and multiplying by 25. The result will be unbiased, typically accurate to about 5%, and is 25 times as fast to compute.

Similar tricks exist for estimating other geometrical quantities. The length of a plane curve can be estimated from the number of crossing points between the curve and a grid of parallel lines. The surface area of a curved surface in three-dimensional space can be estimated from the length of its trace on a plane section [82]. The length of a curve in space can be estimated from the number of points where the curve hits a section plane. Certain quantities related to curvature can also be estimated [9,21].

TABLE 10.1: Standard Notation for Geometrical Quantities

Space dimension n	set X	Letter	Meaning
3	solid domain	V	volume
	curved surface	S	(surface) area
	space curve	L	curve length
	finite set of objects	N	number of objects
	curved surface	M,K	integral of mean curvature
2	plane domain	A	area
	curve	L,B	curve length
	finite set of points	I,P	number of points
	finite set of objects	N,Q	number of objects
	curve	C	total curvature

These methods are summarized in Table 10.2 with notation listed in Table 10.1. Each quantity in Table 10.2 is an unbiased estimator of the quantity to its left (following the arrow). The table is valid only under very strict assumptions of "uniform sampling" (see §10.2) but with very minimal geometrical assumptions, because it relies only on fundamental relationships between volume, area, and length.

TABLE 10.2: Classical Stereological Formulas

Dimension of Space						
3		2		1		0
V_V	$\leftarrow$	A_A	$\leftarrow$	L_L	$\leftarrow$	P_P
S_V	$\leftarrow$	$\frac{4}{\pi}B_A$	$\leftarrow$	$2I_L$		
L_V	$\leftarrow$	$2Q_A$				

Plate 10.1 (preceding page 71) shows an optical microscope image field from a plane section of the lung of a gazelle (magnification $\times 1500$). A stereological test grid has been superimposed on the image, consisting of 40 test points (circled) and line segments totalling 42 cm in length. Since 7 out of 40 test points hit the tissue (rather than the empty airway), we estimate the volume fraction of tissue as $\hat{V}_V = A_A = 7/40 = 17.5\%$. There are 16 positions where a line segment crosses the tissue-airway boundary, so the surface area of lung/air interface per unit volume of lung is estimated at $\hat{S}_V = 2I_L = 2 \times 16/(42/1500) = 1143$ cm^{-1}. Thus, a cubic centimeter of gazelle lung contains about 1100 cm^2 of lung/air interface.

10.2.2 Stereology is Classical Sampling Theory

Results like (10.1) were known as early as 1733 with the celebrated *needle problem* of Buffon [8] and its successors in *integral geometry and geometrical probability* [84,30,48,75,76,80]. However, the first rigorous statistical foundation was laid out only in 1976 by Miles and Davy [20,61,62].

Unbiased estimation, rather than maximum likelihood or minimum mean squared error estimation, is emphasized for several reasons. The distribution of any statistic is difficult to compute because of geometrical complications, and to do so requires severe assumptions about shape (e.g., assuming that X and Y are spheres). One of the beauties of the estimators above is that they are known to be unbiased without geometrical assumptions: they are effectively nonparametric moment estimators.

A simple test grid requires only a few decisions ("hit" or "not hit") on any image. This is convenient in some applications where it is laborious or difficult to recognize boundaries or identify the objects of interest. Yet it appears to throw away most of the information in the image. This is

in fact desirable, for stereological experiments usually generate hundreds of images; it is not efficient (statistically or economically) to analyze a single image in great detail. There is usually enough replication (sections from different parts of the sampling material, windows from different parts of a section) to dramatically reduce the overall sampling variance. In biological applications, the variance contributions associated with variation between animals, and between parts of the same animal, are usually far greater than the variance due to stereological sampling [17,24].

One of the main stereological discoveries of the 1980s was the pervasive importance of *systematic sampling*. Recall that for a finite population of n individuals, ordered arbitrarily and numbered $1, \ldots, n$, a systematic sample with inverse sampling fraction k is generated by choosing a random number m uniformly distributed in $\{1, \ldots, k\}$ and taking the individuals numbered $m,\ m+k,\ m+2k, \ldots$. The sample has random size, but can be said to consist of a *fixed fraction* of the population. The population total of some variable z_i associated with each individual,

$$Z = \sum_{i=1}^{n} z_i$$

can be estimated unbiasedly by taking k times the sample total,

$$\hat{Z} = k \sum_{j} z_{m+jk},$$

see [11]. The approach is similar for a "continuous population": to estimate an integral $I = \int f(x)\,dx$, the numerical integral

$$\hat{I} = \Delta \sum_{j} f(U + j\Delta) \tag{10.2}$$

is an unbiased estimator of I when U is uniformly distributed over $[0, \Delta]$.

Stereological estimates based on grids of points, lines, and the like, are essentially systematic sampling estimates. A point grid is a two-dimensional systematic sample of the continuous two-dimensional plane.

Estimators based on systematic samples are indeed quite efficient. The estimator of the area of a plane set using a point grid is known to have asymptotic variance $\sim La^3$ as $a \to 0$, where a is the distance between grid points and L is the *perimeter length* of the set. This is of order $n^{-3/2}$ rather than n^{-1}, where n is the expected number of points counted. Negative correlation in systematic samples tends to make them more efficient than independent random samples (depending on the structure of the sampling population).

10.2.3 The Particle Problem

Now the bad news. Suppose that our sampling material contains identifiable individual objects—call them "particles"—such as biological cells, crystal grains in a mineral, or holes in a porous rock. We want to regard these particles as individuals forming a population, and make sampling inferences about them: number of particles, average volume, and so on. Usually we cannot sample from this population directly; we have to take plane sections.

It is impossible to estimate N_V, the number of points or objects per unit volume, from plane sections in the sense of Table 10.2. One indication of this is the mismatch of dimensions or units. For example, $S_V = S(\text{mineral})/V(\text{rock})$ is in units $\text{length}^2/\text{length}^3 = \text{length}^{-1}$; so are the other terms in the same row. Now N_V is in units length^{-3}, and so we would naively expect not to be able to estimate it from lower-dimensional sections.

Notice that V, S, and L are "aggregate" quantities, defined as integrals over the object of interest, whereas N is an "individual" quantity with no such interpretation in general. Miles [60] gives an elegant sketch proof justifying the estimation of aggregate quantities as a straightforward exchange of integration and expectation.

The fundamental problem is that *a plane section through a particle population is a* biased *sample of the population.* To see this, visualize the entire sampling material sliced thinly end-to-end by a series of parallel planes. Randomly choose one of the slices with equal probability. The chance that a given particle is represented on this slice depends on the number of slices through that particle, i.e., is proportional to the projected height of the particle in the direction normal to the section planes. Hence the sampling design has a bias in favor of larger particles.

There are essentially three responses to this problem. We can attempt to numerically "correct" our data for the effect of the sampling bias; we can choose to measure different variables that are more "natural" in this sampling design; or we can change the sampling design so that it becomes unbiased.

In the correction approach to estimating N_V, the two-dimensional quantity, we would naively think of using is Q_A, the number of observed particle profiles per unit area of section. This is indeed related to N_V through the *DeHoff-Rhines equation*

$$Q_A = \bar{H}\, N_V$$

(e.g., [88, p. 142]), where $\bar{H}$ is the mean projected height or mean caliper diameter (i.e., the average over all particles X_i of the mean projected height

$H(X_i)$ defined in (10.12) below). Estimation of particle number is thus confounded by particle shape and size (or involves a nuisance parameter associated with shape and size). Even in the happy case where all particles have the same known shape, the distribution of sizes is usually unknown, and it is hard to estimate $\bar{H}$ from plane sections.

In the second approach, we measure sample quantities only when they are three-dimensionally meaningful. For example, if the objective is to study the proportion of "type X" cells in a given tissue, it is not useful to count cells appearing on the section plane, since there is no direct relation between cell sections and cells. Instead, one should measure the area fraction A_A of type X cells on section, because this can be translated directly into an estimate of the volume fraction V_V of type X cells.

10.2.4 Unbiased Counting and Sampling

A better solution to the problems of sampling bias mentioned above is to avoid them altogether by devising another, unbiased, sampling method.

One example is *disector sampling* [79,28,27]. A disector is a pair of parallel plane sections a fixed distance apart; often these are two consecutive slices through the material. We count a particle only if it appears on one section and not on the other. This gives each particle an equal probability of being sampled. The only assumptions needed are (1) that no particle is small enough to fall between two section planes at this distance and (2) that the experimenter can establish the identity of each particle, i.e., can tell whenever the same particle has been sectioned on two different planes.

Sampling bias is present even in two dimensions. Figure 10.1a shows a sketch of a microscope field-of-view with cell profiles visible. The object is to determine N_A, the number of profiles per unit area. A frame F of known area has been superimposed on the image. Naively one would just count all the objects that lie in or on the frame F and divide by the area $A(F)$. The features so counted are shaded in Figure 10.1a.

This counting rule, dubbed *plus-sampling* by Miles [59], clearly produces a biased sample of profiles. If we imagine the field-of-view to be placed at random on the microscope slide, the larger profiles have a greater probability of being sampled. Hence the plus-sampled estimate of N_A is biased: the expected number of profiles counted is greater than $N_A \times A(F)$.

An alternative is *minus-sampling:* count only those profiles that are completely inside the frame F ([59], illustrated in Figure 10.1b). As the name suggests, this counting rule is negatively biased. Smaller profiles have a

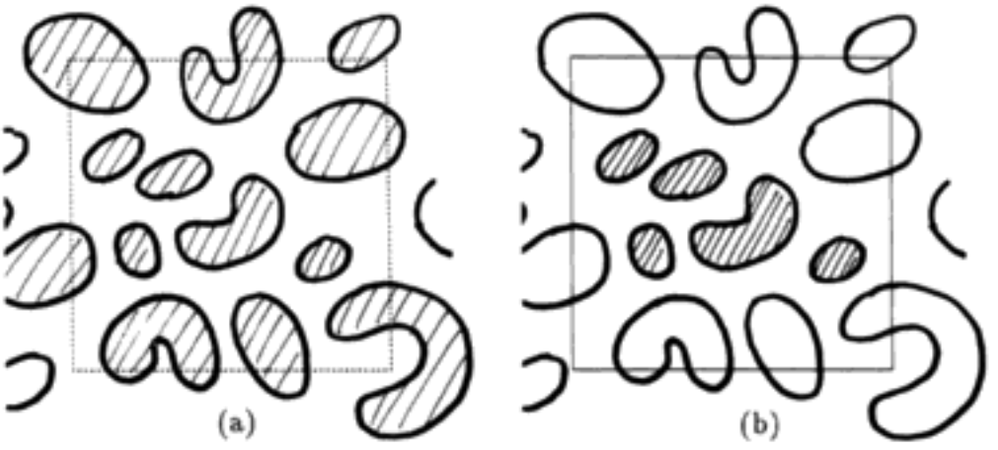


FIGURE 10.1: Two biased counting rules for planar profiles: (a) plus-sampling, (b) minus-sampling.

greater probability of being sampled and counted. Profiles that are actually larger than F will *never* be counted.

A better suggestion is to count only *fractionally* the profiles that hit the boundary of the frame. Count profile X_i with weight $A(X_i \cap F)/A(X_i)$, i.e., the weight is the fraction of area of that profile that is within the window. Using a mean-content formula for windows (§10.3.3), we can verify that the integral of this weight over all translations of F is $A(F)$, so that

$$\hat{N_A} = \frac{1}{A(F)} \sum_i \frac{A(X_i \cap F)}{A(X_i)}$$

is an unbiased estimator of N_A.

An alternative which does not require area calculations is the *associated point* method [59]. Suppose that for any profile X, a unique point $v(X)$ is specified; for example, the centroid of X or the bottom left corner. It is not necessary that $v(X)$ be inside X; we assume only that $v(X)$ is equivariant under translations, $v(X + t) = v(X) + t$ for all vector translations t (if X is shifted then the associated point shifts by the same amount). Then an unbiased estimate of N_A is to count the number of profiles whose *associated points* fall inside F, and divide by $A(F)$. See Figure 10.2a.

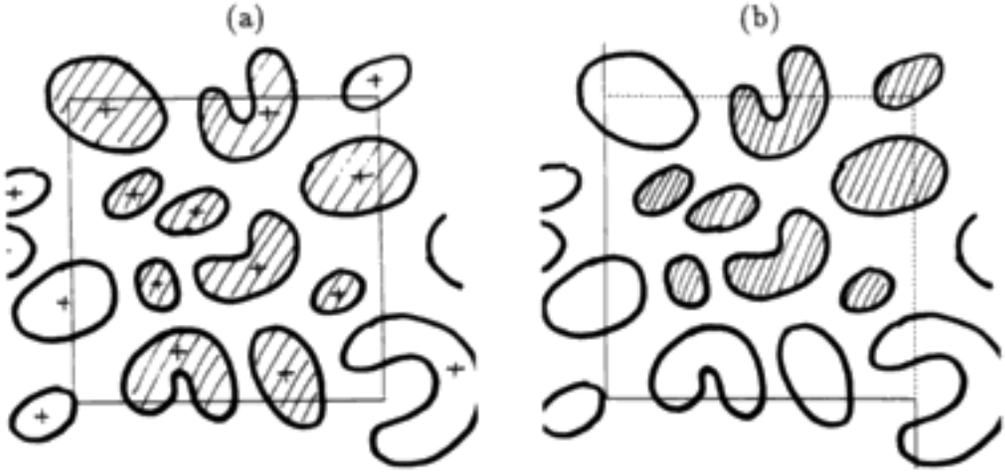


FIGURE 10.2: Two unbiased counting rules for planar profiles: (a) associated point rule, (b) tiling rule.

An even easier alternative suggested by Gundersen [25] employs the special frame illustrated in Figure 10.2b. The solid line, around two sides of the frame and extending to infinity in two directions, is a "forbidden line"; any profile that touches it is not counted. Otherwise any particle that intersects the sampling frame, wholly or partially, possibly crossing the dotted boundary, is counted. The rationale for this rule is, briefly, that if the infinite two-dimensional plane were tiled with copies of this sampling frame (like stacked chairs), then any profile would be counted by exactly one of the frames.

Plate 10.2 (preceding page 71) shows the unbiased estimation of N_V for nuclei in human renal glomerulus using a combination of Gundersen's tiling rule and the disector. Two optical section planes (i.e., different positions of the microscope focal plane) with a separation of 4 μm are shown. To the left is the top (look-up) plane; to the right is the bottom (measuring) plane on which is superimposed a randomly translated tessellation of rectangular counting frames. Nuclei seen clearly on the look-up plane are not counted; on the measuring plane, three new nuclei have come into focus in the counting rectangle just below the center. The counting rectangles have real area $527\mu\text{m}^2$, and so our estimate of N_V is $Q_A^- = 3/(4 \times 527) = 0.001423\mu\text{m}^{-3}$, or roughly 1.4×10^6 nuclei per cubic millimeter of glomerulus.

10.2.5 Spatial Interpretation and Inverse Problems

Its founders envisaged stereology as the spatial interpretation of sections, meaning not only quantitative estimation but also more qualitative reasoning about three-dimensional geometry, including shape and topology. But spatial reasoning is confused by sampling effects. A single three-dimensional object may appear on section as several unconnected objects. A section of a three-dimensional object has smaller diameter than the object itself; while the distance between two objects, or two surfaces (e.g., the thickness of a biological membrane) appears greater on section than in three dimensions. A given three-dimensional object may look very different on different section planes; different three-dimensional objects may fortuitously have identical plane sections.

As we have seen, plane sections and rectangular sampling windows generate biased samples of a particle population, since larger particles have a greater probability of being "caught." Other more subtle biases are caused by selecting a particular orientation for the section plane (for example, always slicing muscle tissue transverse to the muscle fibres) or selecting sections where a particular feature is visible.

"Real" and "ideal" geometry also differ. Since physical slices of biological tissue have nonzero thickness, the microscope image is actually a projection through a translucent slab of material onto the viewing plane. This is the *Holmes effect*: images of sectioned objects are larger than they would be for an ideally thin plane section, and some objects may be obscured by others.

The traditional response was "correction" based on an ideal model, for example, assuming the particles are perfect spheres. Wicksell [94,95] showed that, for a population of spheres, both N_V and the size distribution of the spheres can be determined from sections: if F is the distribution function of sphere radii and G the distribution of circle radii observed on section, then (under suitable sampling conditions [39,70,80]) G has probability density

$$g(s) = \frac{s}{\mu} \int_s^\infty (r^2 - s^2)^{-1/2} \, dF(r) \, .$$

This is an integral equation of Abel type. It is invertible:

$$1 - F(r) = \frac{2}{\pi} \mu \int_r^\infty (t^2 - r^2)^{-1/2} g(t) \, dt \, ,$$

so that F can be uniquely recovered from G. Implicitly this includes the estimation of mean sphere radius μ so that N_V can also be determined.

Similar equations have been encountered in the estimation of the thickness distribution of a biological membrane [42] and the orientation distribution of a curved surface [16].

This is a typical inverse problem, in which an unknown distribution or function is related to an observable function by an integral equation or other operator. The difficulty here is that the inversion of the equation is numerically unstable. For example, the circle distribution G must always have a density. Thus, if we apply a naive inversion procedure to the empirical distribution of circle radii obtained from observations of n circles, the inverted F is not a distribution function [86]. Again, substituting $r = 0$ in the inversion formula shows that μ is proportional to the harmonic mean of G; the estimate of μ will have poor sampling properties.

Part of the trouble is that we are attempting to estimate a whole function F nonparametrically without constraints. An alternative is to model F parametrically and estimate the parameters from observations of G. Nicholson [65,66,67] and Watson [85] also showed that some linear functionals of F, such as its moments, can be estimated reliably from samples of G.

More sophisticated approaches to inverse problems are mentioned in chapter 2 of this report. In the Wicksell context, statisticians have recently proposed kernel smoothing methods [81,14,32,37,83] and iterative methods such as the EM algorithm combined with smoothing [78].

Apart from the considerable numerical hitches, some practical objections to the Wicksell approach are that the geometrical model is unrealistic and untestable (cells are not perfect spheres); extra factors such as the Holmes effect will distort the kernel $f(s|r)$; the amount of data collected in stereological experiments will rarely be sufficient to form a stable estimate of F.

By the 1970s there had been many dubious or even erroneous attempts to avoid section effects, and theoretical stereologists evolved the narrower "party line" that it is only possible to reliably estimate certain aggregate three-dimensional quantities such as volume and surface area. More recently, additions to the list of fundamental formulas (Table 10.2) have made it possible to estimate parameters such as the mean squared particle volume, without any assumptions about particle shape. The list of parameters that can be reliably estimated—without shape assumptions—now includes some quantities related to curvature, orientation, and "shape."

10.2.6 Stochastic Models

Stereological inference and spatial interpretation are difficult because we simultaneously have not enough data (important three-dimensional information is lacking) and too much data (the two-dimensional images are not analyzed closely). Stochastic models can bridge this gap.

Explicit Models

At one extreme, we could build a probability model for the entire spatial structure X using random set models from stochastic geometry [34,48,53, 77,80]. An explicit, parametric model would contain information about the sizes, positions, shapes, relative arrangement, and topological relationships of components in X, which could be estimated by comparatively familiar statistical methods. Explicit models in stochastic geometry are mostly analogues of point processes, the Poisson, Cox, cluster, and Markovian categories described in chapter 7. Some statistical theory is available for them [2,33,69,71,77], and they have proved to be excellent descriptions of some simple structures such as rock fractures and crystalline materials [77]; but realistic models for the highly organized structures of biology and ecology still elude us.

Stationary Models

"Nonparametric spatial modeling" is a less demanding approach where the random spatial process X is not explicitly described, but is assumed to be *stationary* (certain distributions or moments are invariant under translations and/or rotations). Then we can nonparametrically estimate the moments or distributions associated with the process [80, chap. 4]. All the standard stereological results can be rederived in this context (see [57,58,64]) since in fact it is a reformulation of the same sampling problem. The reformulation emphasizes how little need be assumed about the spatial structure X, and suggests new estimators. For example, the locational interaction (such as clumping or dispersion) between parts of a spatial structure can be described by the second-order moment characteristics of the process, which can be estimated nonparametrically from sample data. The K function for point processes [70], described in chapter 7, is one instance.

Semiparametric Models

Commonly, only a part of a spatial structure X is of interest. If only that part is modeled, we have a semiparametric statistical model. For example, the thickness of a curved tube could be modelled by a parametric family of distributions for the radius, without specifying shape or location except to assume that the process is stationary [4]. The distribution of *directions* in a structure (e.g., surface normal vectors, curve tangent vectors) could be modeled by a parametric family of distributions on the unit sphere [16]. In a material consisting of several phases or compartments, one can test whether the arrangement of phases is "random" or whether some phases tend to be associated, by applying standard discrete data models [50].

Data Models

At the other extreme would be a statistical model for the stereological data obtained from a series of samples T_i. For example, Cruz [13] proposed a proportional linear regression model for, say, $A(Y \cap T_i)$ against $A(X \cap T_i)$. This model has been criticized [43], and justifications must remain largely empirical, because it is difficult to derive any distributional theory from probabilistic models of the structure or the sampling design.

10.3 Statistical Theory

Stereological methods can be applied with minimal knowledge of the three-dimensional structure under study. However, the sampling rules must be strictly followed; the experimental protocol must generate a random plane or probe with the correct distribution required by stereological theory. In this section, we describe that theory, and show how simple design mistakes can lead to catastrophic errors.

10.3.1 What To Estimate

It was believed for many years that the normal human brain, alone among all organs, loses cells without replacing them. This was established repeatedly from estimates of N_V (cell number per unit volume) at different ages. However, the quantity of real interest is the total cell number N, not N_V. In 1985, Haug [35,36] pointed out that, since younger tissues shrink more during fixation (chemical treatment prior to embedding and sectioning), the

total brain volume after fixation was effectively an increasing function of age, and this could account for the decrease in N_V estimates. The situation is still unresolved because of other uncontrolled variables; but it may be that the wrong scientific question was pursued for 20 years.

This emphasizes the distinction between a total quantity

$$\beta = \beta(Y)$$

and a relative quantity

$$\beta_V = \frac{\beta(Y)}{V(X)},$$

where X is the containing set and Y is the "feature" of interest ($Y \subset X$).

Estimation of absolute and relative quantities is also different. We can convert estimates of β to β_V and vice versa, given an estimate of $V(X)$; but statistical properties of the estimators are not preserved. For example, the expectation of a ratio of random variables is not generally equal to the ratio of their expectations. Sampling designs and estimators that are unbiased or optimal for estimating β_V may not be appropriate for β and vice versa.

10.3.2 Inference

Statistical inference is called design-based if it relies on the randomness in the sampling design. Expectations are averages over all possible outcomes of the sampling. In design-based stereology it is assumed that the geometrical object X is fixed and the sampling probe T is random. Meanwhile, inference is called model-based if it imagines the sampling population was generated by a stochastic model. Expectations are averages over all hypothetical realizations of this model. In model-based stereology, it is assumed that X is (a bounded sample from) a realization of a random process, and the sampling probe T is arbitrary, say fixed.

This is mainly an issue of correctly specifying the population to which we wish to extrapolate statistical inferences. The design-based approach corresponds to finite population inference for survey samples [11] or randomized design inference, while the model approach corresponds to superpopulation inference. Miles [60] distinguishes three kinds of inference in stereology:

Restricted case: The specimen X is a nonrandom, bounded set that is the sole object of interest. For instance, a whole organ from an experimental animal could be available for study. Typically we want to estimate the total volume, surface area, etc., of the organ.

Extended case: The specimen X available for examination is but a portion sampled from a much larger object W. For instance, a rock sample is typically taken from a large outcrop of rock, and we wish to make inferences about the latter. Either total gold volume or relative gold volume fraction might be of interest.

Random case: A stochastic process really exists that generates the internal structure of the sample. That is, the specimen X is a fixed set, but the feature Y inside X is generated as $Y = X \cap Z$ where Z is a "random set" or "spatial stochastic process." For instance, a metallurgist will regard a small piece of steel cut from a bar, formed at a known temperature, and so on, as a sample from the infinite hypothetical reservoir of steel that could be formed under those same conditions. Quantities like total volume are meaningless here; we are mainly interested in fractions per unit volume of steel.

In the restricted case, we are totally dependent on the randomness of the sampling probe T to guarantee validity of the method; but apart from this we do not need to make any unverifiable assumptions.

In the extended case, it must typically be assumed that X was sampled "randomly" from W. For some purposes, it is not valid to sample a rock outcrop by breaking off a piece with a hammer, since the breakage surfaces will usually depend on the internal rock structure.

In the random case, Y must be independent of X; that is, the internal structure must not depend on the external boundaries of the specimen. This would be inappropriate for objects such as biological organs, which have many levels of internal organization.

10.3.3 Geometrical Identities

Unbiased estimation of properties of a set X from observations of the intersection $X \cap T$ is possible thanks to the *mean-content formulas* or *section formulas* of integral geometry [76,89], which have the general form

$$\int_{\text{positions of } T} \alpha(X \cap T)\, d\mu(T) = c\,\beta(X), \tag{10.3}$$

where α, β are geometrical quantities such as those listed in Table 10.1, and $c = c_{\alpha\beta}$ is a constant. Here μ is a so-called "invariant measure" on the space of all possible probes T; basically, this is an appropriate generalization of

Lebesgue measure, and so the integral represents "uniform integration" over positions of T.

A simple example is the statement that the volume of a three-dimensional object is the integral of the areas of its plane slices:

$$\int_{-\infty}^{\infty} A(X \cap T_h)\, dh = V(X), \tag{10.4}$$

where T_h is the plane $\{(x, y, z)\colon x = h\}$. This is known as *Cavalieri's principle*. In simple terms, the volume of an arbitrarily shaped potato can be determined by slicing the potato into infinitely thin parallel slices and summing the areas of the slices. The slicing direction is fixed and arbitrary; we could also average over all orientations, giving

$$\iint A(X \cap T_{\omega,h})\, dh\, d\omega = 2\pi V(X), \tag{10.5}$$

where $T_{\omega,h}$ denotes the plane with direction given by its unit normal vector ω and displacement h from the origin. This averaged version is no longer practicable for potatoes, since after slicing end-to-end in direction ω_1, we have to reassemble the object and slice it end-to-end from another angle ω_2, and so on.

The basic mean content formulas in three dimensions are summarized in Table 10.3. In general, the formulas involving plane sections or line probes require us to average over all orientations. For example, the surface area $S(Y)$ of a curved surface $Y \subset \mathbf{R}^3$ can be determined from the lengths of plane sections,

$$\iint L(Y \cap T_{\omega,h})\, dh\, d\omega = \frac{\pi^2}{2} S(Y), \tag{10.6}$$

but in this case there is no analogue of (10.6) for planes with fixed orientation. The surface area of a potato cannot be determined from the boundary lengths of parallel slices, unless we are permitted to reassemble and reslice the object many times. A better alternative for surface area is to use the mean content result in which T is a one-dimensional line. Thus we would repeatedly impale the potato on an array of linear spikes, changing the potato's orientation each time, and count the total number of points where the spikes penetrated the surface.

Other sampling probes can also be used. Instead of infinite two-dimensional planes, we can take bounded sampling windows within a plane; the results here are similar except that the right-hand sides also involve the area of the

TABLE 10.3: Typical Mean Content Results (3 dimensions)

Object X	Probe T	Resulting $X \cap T$	Desired β	Required α
Solid domain	plane	plane domain(s)	V	A
	line(s)	linear domain(s)	V	L
	point(s)	point(s)	V	P
Surface	plane	plane curve(s)	S	L
	line	point(s)	S	I
Space curve	plane	point(s)	L	Q
Surface	plane	plane curve(s)	M	C

sampling window. Instead of a single plane T_h in (10.4) we may take a stack of equally-spaced parallel planes

$$G_h = \{\ldots, T_{h-2s}, T_{h-s}, T_h, T_{h+s}, T_{h+2s}, \ldots\};$$

summing the contributions $A(X \cap T_{h+ks})$ in (10.4) gives

$$\int_0^s A(X \cap G_h)\, dh = V(X). \tag{10.7}$$

Note that the range of integration is now the bounded interval $[0, s)$ because the stack of planes is uniquely specified by its "starting position" $h \in [0, s)$.

To take stock of these results we note that

1. They do not depend on the "shape" or position of the objects X, and are true under very minimal regularity conditions;

2. They are valid only when integration is performed "uniformly" over all positions of T (in most cases this requires averaging over orientations); and

3. They are statements about integrals or mean values only.

For some time, stereologists thought that opportunities for finding new mean value formulas were severely restricted by (c). This turned out to be pessimistic, because many properties of a geometrical object can be expressed as integrals. For example, some powers of volume $V(X)^m$ can be stereologically determined. Again, the orientation distribution of a curved surface Y in $\mathbf{R}^3$ is the probability distribution of the unit normal vector at a uniformly distributed random point on Y. This is a distribution Q on the unit sphere

S^2 defined by $Q(U) = A(Y_U)/A(Y)$ for $U \subseteq S^2$, where Y_U is the subset of Y consisting of all points $y \in Y$, where the unit normal $\omega(y)$ lies in U. The observed orientation distribution of a two-dimensional plane section $Y \cap T$ is related to Q by an integral equation reminiscent of Wicksell's equation for particle size [16].

10.3.4 Design-Based Estimation

The design-based approach is analogous to sampling design for finite populations [11] but has interesting geometrical complications. The set X is fixed (Miles's restricted or extended models); the probe T is generated by a random sampling design. For example, T could be a single random plane (the analogue of simple random sampling) or a stack of parallel planes (the analogue of systematic sampling). The choice whether to estimate total or relative quantities ($V(X)$ or V_V) affects the choice of sampling design.

Thus we need to convert mean content formulas (10.3) into results of the form

$$\mathbf{E}\alpha(Y \cap T) = c_T \beta(Y), \tag{10.8}$$

$$\mathbf{E}\frac{\alpha(Y \cap T)}{\alpha'(X \cap T)} = c\frac{\beta(Y)}{\beta'(X)}, \tag{10.9}$$

$$\frac{\mathbf{E}\alpha(Y \cap T)}{\mathbf{E}\alpha'(X \cap T)} = c\frac{\beta(Y)}{\beta'(X)}, \tag{10.10}$$

where $Y \subset X$ are fixed sets, T is a random probe, and $\mathbf{E}$ denotes expectation with respect to the distribution of T. In (10.8) c_T is a constant associated with this distribution; in the other versions $c = c_{\alpha\beta}/c_{\alpha'\beta'}$ is a "geometrical constant." Usually $\alpha' = A$ and $\beta' = V$ so that in (10.9 and 10.10) we are estimating β_V from α_A.

The stereological equivalent of a uniform random sample is a so-called *isotropic uniformly random (IUR)* probe. Suppose we wish to generate a random probe T intersecting a set $X \subset \mathbf{R}^n$. The probe is said to be IUR if it has constant probability density with respect to the invariant measure μ that features in (10.3):

$$\begin{aligned} dP(T) &= \tfrac{1}{H(X)}\, d\mu(T) && \text{if } T \cap X \neq \emptyset \\ &= 0 && \text{if } T \cap X = \emptyset. \end{aligned} \tag{10.11}$$

Here, $H(X)$ is the appropriate normalizing constant,

$$H(X) = \int_{T \cap X \neq \emptyset} d\mu \,, \tag{10.12}$$

i.e., the total μ-measure of all positions of T that intersect X.

For example, if the "probe" T is just a single point, the invariant measure μ is Lebesgue (volume) measure, so that $H(X) = V(X)$, and an IUR point probe hitting X is just a random point uniformly distributed in X.

As a less trivial example, a straight line T in two dimensions is uniquely specified by its direction θ and its distance from the origin:

$$T(\theta, p) = \{(x, y) : x \cos\theta + y \sin\theta = p\} \,,$$

where $\theta \in [0, \pi)$, $p \in \mathbf{R}$. The invariant measure for lines turns out to be [48,76]

$$d\mu = dp\, d\theta \,,$$

i.e., Lebesgue measure on the (θ, p) coordinate space. Thus, an IUR line probe T hitting X is a line with random coordinates θ and p, such that the pair (θ, p) is uniformly distributed over the permissible range

$$\{(\theta, p) : T(\theta, p) \cap X \neq \emptyset\} \,.$$

In the special case where X is a disc of radius r centered at the origin, an IUR line through X is generated by making θ and p independent and uniformly distributed over $[0, \pi)$ and $[-r, r]$, respectively; hence the term IUR. However, note that for a general set X the coordinates of an IUR line are *not* independent and their marginal distributions are *not* uniform. A practical method of generating IUR lines through an arbitrary set X is to enclose X by a larger disc $D \supset X$, generate a sequence of IUR lines intersecting D, and take the first line that happens to intersect X. This is just an application of the *rejection method* of Monte Carlo simulation.

The probability that an IUR line through X will hit $Y \subset X$ is $H(Y)/H(X)$ with H as defined in (10.12). In other words, the probability that an IUR line intersects a particular target is related to the *mean projected height* of the target. This does not depend on the position of Y within X; so in a sense the IUR line is a uniform sample through X.

Returning to the estimation problem, clearly we can derive (10.8) from (10.3) by taking T to be an IUR probe hitting X, so that

$$\mathbf{E}\alpha(Y \cap T) \;=\; \int_{\text{positions of } T} \alpha(Y \cap T)\, dP(T)$$

$$
\begin{aligned}
&= \frac{1}{H(X)} \int \alpha(Y \cap T)\, d\mu(T) \\
&= \frac{c_{\alpha\beta}}{H(X)} \beta(Y);
\end{aligned}
$$

so we have an unbiased estimator of $\beta(Y)$, provided we can determine the normalizing constant $H(X)$.

However, a similar argument will not work for formulas (10.9) such as Delesse's principle. The problem is that the expectation of a ratio of random variables in general does not equal the ratio of their expectations. Historically there were many incorrect derivations of Delesse's principle; but the result is just not true for IUR planes. Miles and Davy [20,61] showed that a solution is to take T with the *weighted distribution* with probability density proportional to $\alpha'(X \cap T)$,

$$
dP(T) = \frac{\alpha'(X \cap T)}{G(X)}\, d\mu(T)\,,
$$

where α' must be nonnegative (e.g., A or L but not C). The normalizing constant is

$$
G(X) = \int \alpha'(X \cap T)\, d\mu(T) = c_{\alpha'\beta'}\, \beta'(X)\,.
$$

Then, using $\mathbf{E}_W$ to denote averages with respect to this weighted distribution, we have

$$
\begin{aligned}
\mathbf{E}_W \frac{\alpha(Y \cap T)}{\alpha'(X \cap T)} &= \int \frac{\alpha(Y \cap T)}{\alpha'(X \cap T)}\, dP(T) \\
&= \int \frac{\alpha(Y \cap T)}{\alpha'(X \cap T)} \frac{\alpha'(X \cap T)}{G(X)}\, d\mu(T) \\
&= G(X)^{-1} \int \alpha(Y \cap T)\, d\mu(T) \\
&= \frac{c_{\alpha\beta}}{c_{\alpha'\beta'}} \frac{\beta(Y)}{\beta'(X)}\,.
\end{aligned}
$$

This holds provided $\alpha'(X \cap T) > 0$ whenever $\alpha(Y \cap T) \neq 0$. Note that the proportionality constant c is now a geometrical constant not dependent on T. Another, closely related, solution is to estimate the numerator and denominator of (10.10) separately on a large number of replicated samples: in other words, when replication is present, take the ratio of means, not the mean of the ratios.

The problem encountered here was that plane sections and other stereological analogues of simple random sampling actually do not yield fixed sample size. The sample mean is biased when the sample size is random. We must instead use samples with probability proportional to size, or take replicated estimates and numerically weight them in proportion to size.

Most stereological sampling designs do not have fixed sample size. Different plane sections of a bounded three-dimensional object have different size and shape. Thus, simple random sampling does not generalize easily to most stereological situations. There are some exceptions: a sampling window is a fixed-size sample, if the object of interest always fills the entire window.

Of course, systematic sampling does generalize well to stereology, as we have seen. Stereological estimates based on grids of points, lines, and the like, are essentially systematic sampling estimates. Cavalieri's principle for a stack of planes (10.7) is just an application of (10.2) to the function $f(h) = A(X \cap T_h)$ appearing in the original Cavalieri formula (10.4).

The parameter space describing all positions of a grid or systematic sample is totally bounded, and the invariant measure μ can be integrated over the entire space. In (10.7) the position of a plane grid was specified by a value $h \in [0, s)$. Thus an IUR grid is defined to have uniform probability density with respect to μ over the entire space,

$$dP(T) = \frac{1}{H}\, d\mu(T)\,, \tag{10.13}$$

where the normalizing constant is now the total μ measure of all positions of T,

$$H = \int d\mu(T)\,.$$

Typically, H depends only on the grid spacing. Thus, estimation of a population total according to (10.8) is relatively easy when T is a systematic grid. An unbiased estimate of the volume of a potato can be obtained by cutting it into thick slices by parallel planes at constant separation d, summing the areas of the slices, and multiplying by d.

10.3.5 Model-Based Estimation

In model-based stereology we convert (10.3) into

$$\mathbf{E}\frac{\alpha(Y \cap T)}{\alpha'(X \cap T)} = c_{\alpha\beta\alpha'\beta'}\mathbf{E}\frac{\beta(Y)}{\beta'(X)}\,, \tag{10.14}$$

where Y is a random set $Y = Z \cap X$ inside a fixed three-dimensional domain X. The probe T is now arbitrary (say, fixed). This time $\mathbf{E}$ is present on both sides and denotes expectation with respect to the random structure Y. Note the denominators are constant.

The nonparametric modeling approach described in §10.1.5 is simply to assume that the random process Z is statistically stationary and derive (10.14) by studying moments associated with Z. For example, here is a sketch proof of the model-based Delesse formula,

$$\mathbf{E}\frac{A(Y \cap T)}{A(X \cap T)} = \mathbf{E}\frac{V(Y)}{V(X)}. \tag{10.15}$$

Suppose the random process Z is such that for any $x \in \mathbf{R}^3$ the indicator variable

$$I(x) = \begin{cases} 1 & \text{if } x \in Z \\ 0 & \text{if not} \end{cases}$$

is a well-defined random variable. Let

$$p(x) = \mathbf{E}I(X) = \mathbf{P}\{x \in Z\}.$$

Then

$$\mathbf{E}V(Y) = \mathbf{E}\int_X I(x)\,dx = \int_X \mathbf{E}I(x)\,dx = \int_X p(x)\,dx$$

by exchanging integration and expectation. Assume Z is *first-order stationary* in the sense that $p(x) = p$ does not depend on x. Then this integral is

$$\mathbf{E}V(Y) = pV(X).$$

By a similar argument

$$\mathbf{E}A(Y \cap T) = pA(X \cap T),$$

so that both sides of (10.15) are equal to p, and the result is proved.

This example needed only a simple exchange of integration and expectation. For the other stereological formulas, we need the integral geometric results (10.3), and first-order stationarity assumes (roughly) that certain first moments associated with Z are invariant under translation and/or rotation. The formal apparatus is the moment theory of stationary random measures [57,58,80].

Other, higher-order expectations can be calculated similarly. For example, the variance of A_A can be expressed in terms of the second-order characteristics of the process. We now need to assume Z is *second-order stationary*,

which in this case means that $\mathbf{E}I(x)I(y) = \mathbf{Pr}\{x \in Z \text{ and } y \in Z\} = r(x,y)$ depends only on $y-x$. The resulting formula gives the variance as an integral in terms of r: this is equivalent to the basic variance result of geostatistics (see chapter 5).

Characteristics of "infinite order" can be considered exactly as in the design-based case, for example, the orientation distribution of a curved surface.

10.4 Recent Research and New Directions

10.4.1 Variance of Systematic Sampling

Systematic sampling usually leads to more efficient estimation than simple random sampling with comparable sample size. However, there are fewer general results about the variance under systematic sampling because this depends on the "structure" of the population [11]. At worst, there could be a periodicity in the population that matches the periodicity of the systematic sample, and the variance would be elevated. The classic example is an army population where every tenth serial number is allocated to a sergeant. In stereology, such cases do arise when a biological structure such as a corrugated sheet is sampled by a test grid with the same spacing.

The estimator of the area of a plane set based on a point grid has recently been studied extensively [15,29,45,51,54] using earlier results about the systematic sampling estimator (10.2) of an integral [52]. The variance of (10.2) is

$$\text{var}(\hat{I}) = \Delta \sum_j g(j\Delta) - \int_{-\infty}^{\infty} g \,,$$

where g is the *covariogram* of f,

$$g(x) = \int_{-\infty}^{\infty} f(y) f(x+y)\, dy \,,$$

see [52] and chapter 5. For a wide class of applications,

$$\text{var}(\hat{I}) \sim -\frac{1}{6} g'(0)\, \Delta^2$$

as $\Delta \to 0$. The point-counting estimator of area of a plane set has been found [46,47,15,29] to have variance

$$\text{var}(\hat{A}) \approx 0.0724\, La^3$$

as $a \to 0$, where L is the perimeter length of the set, this being a good approximation for a wide variety of shapes.

10.4.2 Fractionator Sampling

A simple yet extremely powerful application of systematic sampling is the *fractionator* technique [26]. Suppose we wish to estimate the total volume or number of cells in a large animal. Effectively, we are in the "extended case" where it is not feasible to study more than a tiny sample from the object. Worse, it would seem that we have to generate a uniformly distributed random sample of this complex object in order to get valid estimates.

On the other hand, it is easy to generate a *systematic* sample of an animal. We start by dismembering the animal—in any fashion we choose—and arranging the pieces in arbitrary order (e.g., in ascending order of size; or at whim). Then we take a systematic sample of this finite population (inverse sampling fraction k_1) and throw away the remaining material. The retained sample is then cut into smaller pieces and again arranged in arbitrary order; a systematic subsample of this material (inverse fraction k_2) is taken. The process is repeated until we have a subsample that is small enough to analyze microscopically. Then we apply a design-based method to estimate the total amount of material in this ultimate subsample. Finally the total for the entire animal is estimated by the subsample total estimate times the product of the successive inverse sampling fractions $k_1 \cdots k_n$. Clearly this estimator is unbiased. Sampling fractions as low as 10^{-9} are routinely used in brain tissue, meaning that only ≈ 100 cells are actually counted.

At the lowest level of the experiment we still have the problem of estimating the total number (say) of cells in the sample. But here we can often employ a modification of the disector method. If the last stage of subsampling is carried out by slicing the material into sections and systematically subsampling the sections, then we just apply the disector counting rule to each section, and *sum* the disector counts. *This does not require knowledge of section thickness.* Indeed the section planes can be separated by different distances, and even be nonparallel [26,28,27].

Little is known theoretically about the variance of fractionation sampling, although the estimator clearly has a martingale structure. Current practice is to form a jackknife estimate of variance, by initially dividing the specimen into two comparable halves, forming a fractionator estimate from each half, and estimating variance from the absolute difference of the two estimates.

FIGURE 10.3: A point-sampled intercept through a plane section of a particle.

10.4.3 New Estimation Formulas

Perhaps the most exciting area of stereology is the discovery of new mean content results (10.3). Some of the new quantities β are associated with "shape," "size," orientation, curvature, or spatial arrangement. Other results apply in sampling situations where it was previously thought impossible to estimate anything.

Let x be a point inside a three-dimensional set X, which we assume convex (for convenience only). Let $\ell(x,\omega)$ be the infinite ray (half-line) through x in direction ω, where ω is a unit vector. Then the mean cubed length of the intersection between this line and X is proportional to the volume of X:

$$\frac{1}{4\pi}\int_{S^2} L(X \cap \ell(x,\omega))^3 \, d\omega = \frac{3}{4\pi} V(X); \tag{10.16}$$

this is an application of elementary calculus. Note that x is a fixed, arbitrary point. A similar but more complicated formula holds if X is not convex and/or x is outside X.

This can be used [40] to estimate the mean squared volume of particles in a three-dimensional population. First take an area-weighted plane section of the sample material; superimpose a point grid on the section, and at every grid point which hits a particle profile, place a line in a random direction through the grid point and measure the cubed intercept (i.e., length of the intersection between the line and the particle profile). See Figure 10.3. Un-

der this sampling regime, the particles have been selected with probabilities proportional to their individual volumes,

$$p_i = \mathbf{P}\{X_i \text{ selected}\} = \frac{V(X_i)}{\sum_j V(X_j)}.$$

The cubed intercept lengths estimate the individual volumes; so the mean cubed intercept length is an estimate of the *volume-weighted* mean particle volume,

$$\overline{v_V} = \sum_i p_i V(X_i) = \frac{\sum_i V(X_i)^2}{\sum_i V(X_i)},$$

i.e., this is the ratio of mean square volume to mean volume. The mean volume can be estimated separately from estimates of total volume and total number; thus we have reliable (approximately unbiased) estimates of the first two moments of particle volume. Methods exist for some higher moments. In some applications, particularly in pathology, the mean square volume (or variance of volume, etc.) has proved very useful in detecting differences between particle populations.

Another application of (10.16) is useful in studying materials that do not consist of separate particles. Let Y be any set in three dimensions. For example, Y might be the union of all the cells in a tissue, or the empty space in a porous material. At any point x, define the *star set* $S(x,Y)$ to be the set of all points y such that the line segment from x to y lies wholly inside Y. See Figure 10.4. If $x \notin Y$, then $S(x,Y)$ is empty; otherwise, $S(x,Y)$ is a "star-shaped" set, and if Y is convex, then $S(x,Y) = Y$. Consider the *mean star volume*, i.e., the mean of $V(S(x,Y))$ over all points x. This can be estimated on plane sections by the mean cubed length of an intercept through a point in the section. The star volume gives us an interesting measure of the average "local size" of holes in a porous material.

Variations on the star volume, involving other moments of intercept length, have recently been considered as indicators of "shape" [29].

Covariance, and other second-order parameters, can be estimated without bias. This is easiest to describe when X is a stationary random set in $\mathbf{R}^3$. The (noncentered) spatial covariance of X at lag $h \in \mathbf{R}^3$ is

$$C(h) = \mathbf{E}\, 1_X(0)\, 1_X(h),$$

where 1_X is the indicator function of the set X. In other words, this is the expected volume fraction of points x in space where both x and $x+h$ simultaneously lie inside X. If we are willing to assume that X is isotropic, then $C(h)$ depends only on the length $|h|$ and not on direction, and we

FIGURE 10.4: The star volume.

can estimate $C(h)$ as a function of $|h|$ from the sample covariance of plane sections of X. This has been applied to extract detailed information about a material [33,77]. Second-order statistics have also been used to define indices of mineral liberation [19,18].

Non-uniform sampling designs are a very important development. As remarked in §10.3.3, the general formulas for estimating quantities other than volume require random section planes with (roughly speaking) uniform distribution over all possible orientations and all possible positions. However, many experimenters cannot adhere to this requirement. For example, about a third of all stereological applications require that the section plane be cut in a particular direction, either for physical reasons, or because the structure of interest can only be identified when cut this way.

A common case is "vertical" sectioning, where the section plane must be aligned with a specified axis, in other words, normal to some well-defined plane we can call the "horizontal." Thus, there is only one degree of rotational freedom for plane orientation and one degree of translational freedom. An unbiased estimate of surface area from vertical sections has recently been found [5] that uses a test grid consisting of cycloid arcs.

Sampling designs that are non-uniform in position and orientation have recently been studied [41,90,91,92,93]. Mattfeldt and Mall [56,55] proposed samples involving three mutually orthogonal section planes.

10.4.4 Research Frontiers

Here we speculate about future advances (other than those already in progress and covered above).

Three Dimensions

New imaging modalities (such as confocal optical microscopy, infrared Fourier transform imaging) have been developed that can "see" directly into three-dimensional structures such as biological soft tissue and solid bone. Three-dimensional images can also be reconstructed computationally from serial optical sections or tomographic data. Rather than making stereology redundant, this technology has released a flood of interesting new problems. Stereological sampling techniques are needed, e.g., for counting three-dimensional particles [38], and the methods of two-dimensional spatial statistics (see chapters 4 and 7) need to be adapted and refined for three dimensions [6,49].

Structured Models

One reason for the overwhelmingly "nonparametric" character of stereology is that explicit stochastic process models have not succeeded in reproducing the very high degree of organization seen in real (especially biological) microscopic structures. This may change in the next five years. Much recent activity in stochastic geometry [80] is focusing on models where the realizations have a prescribed, ordered appearance such as random tessellations [63], random dense packings, and random fibre processes.

Markov Models

Particularly promising is the development of several kinds of Markov models for spatial processes [1,7,10,73,71,72]. These are one step more complex than completely random Poisson processes, in that a stochastic interaction is allowed between "neighbouring" elements of the process, for example, pairwise interactions between the points in a point process. Markov point processes and random sets can easily be simulated using Monte Carlo methods, and they are convenient for likelihood-based inference [68].

Bootstrap Methods

Bootstrap resampling methods were introduced to stereology by Hall [31,33] in connection with the point-counting estimator of area fraction A_A. The basic idea was to break the sampling region into strips or pieces that are

sufficiently separated for any dependence to be ignored, and to resample these pieces as if they were i.i.d. observations. It seems likely that such methods will prove a useful alternative to parametric modeling, as a way of getting information about variances and confidence levels. The difficulty is in finding acceptable ways of bootstrapping a spatial process with all its inherent spatial dependence.

Bibliography

[1] Arak, R. J., and D. Surgailis, Markov random fields with polygonal realizations, *Probability Theory and Related Fields* **80** (1989), 543–579.

[2] Ayala, G., *Inferencia en Modelos Booleanos*, Ph.D. dissertation, University of Valencia, Spain, 1988.

[3] Baddeley, A. J., Stochastic geometry: an introduction and reading-list, *Int. Stat. Rev.* **50** (1982), 179–183.

[4] Baddeley, A. J., and P. Averback, Stereology of tubular structures, *J. Micros.* **131** (1983), 323–340.

[5] Baddeley, A. J., H. J. G. Gundersen, and L. M. Cruz-Orive, Estimation of surface area from vertical sections, *J. Micros.* **142** (1986), 259–276.

[6] Baddeley, A. J., C. V. Howard, A. Boyde, and S. Reid, Three-dimensional analysis of the spatial distribution of particles using the tandem-scanning reflected light microscope, *Acta Stereologica* **6** (supplement II, 1987), 87–100.

[7] Baddeley, A. J., and J. Møller, Nearest-neighbour Markov point processes and random sets, *Int. Stat. Rev.* **57** (1989), 89–121.

[8] Leclerc, G. L., Comte de Buffon, Essai d'arithmétique morale, in *Supplément à l'Histoire Naturelle*, vol. 4, 1777.

[9] Cahn, J. W., The significance of average mean curvature and its determination by quantitative metallography, *Trans. Amer. Inst. Min., Mett., Pet. Eng.* **239** (1976), 610.

[10] Clifford, P., Markov random fields in statistics, in *John Hammersley Festschrift*, to appear, 1990.

[11] Cochran, W. G., *Sampling Techniques*, 3rd edition, John Wiley and Sons, New York, 1977.

[12] Coleman, R., *An Introduction to Mathematical Stereology*, Memoirs no. 3, Department of Theoretical Statistics, University of Aarhus, Denmark, 1979.

[13] Cruz-Orive, L. M., Best linear unbiased estimators for stereology, *Biometrics* **36** (1980), 595–605.

[14] Cruz-Orive, L. M., Stereology: recent solutions to old problems and a glimpse into the future, *Acta Stereologica* **6/III** (1987), 3–18.

[15] Cruz-Orive, L. M., On the precision of systematic sampling: a review of Matheron's transitive methods, *J. Micros.* **153** (1989), 315–333.

[16] Cruz-Orive, L. M., H.. Hoppeler, O. Mathieu, and E. R. Weibel, Stereological analysis of anisotropic structures using directional statistics, *Appl. Stat.* **34** (1985), 14–32.

[17] Cruz-Orive, L. M., and E. R. Weibel, Sampling designs for stereology, *J. Micros.* **122** (1981), 235–257.

[18] Davy, P. J., Liberation of points, fibres and sheets, pp. 69–78 in *Proceedings of the Oberwolfach Conference on Stochastic Geometry*, Teubner, Leipzig, 1984.

[19] Davy, P. J., Probability models for liberation, *J. Appl. Prob.* **21** (1984), 260–269.

[20] Davy, P. J., and R. E. Miles, Sampling theory for opaque spatial specimens, *J. R. Stat. Soc., B* **39** (1977), 56–65.

[21] DeHoff, R. T., The quantitative estimation of mean surface curvature, *Trans. Amer. Inst. Min., Mett., Pet. Eng.* **239** (1967), 617.

[22] Delesse, M. A., Procédé mécanique pour déterminer la composition des roches, *C. R. Acad. Sci. Paris* **25** (1847), 544.

[23] Glagolev, A. A., On geometrical methods of quantitative mineralogic analysis of rocks, *Trans. Inst. Econ. Min.* (Moscow) **59** (1933), 1.

[24] Gundersen, H. J., and R. Østerby, Optimizing sampling efficiency of stereological studies in biology: or 'Do more less well!', *J. Micros.* **121** (1981), 65-74.

[25] Gundersen, H. J. G., Estimators of the number of objects per area unbiased by edge effects, *Micros. Acta* **81** (1978), 107–117.

[26] Gundersen, H. J. G., Stereology of arbitrary particles. A review of unbiased number and size estimators and the presentation of some new ones, in memory of William R. Thompson, *J. Micros.* **143** (1986), 3–45.

[27] Gundersen, H. J. G., and others, The new stereological tools: disector, fractionator, nucleator and point sampled intercepts and their use in pathological research and diagnosis, *Acta Pathol. Immun. Scand.* **96** (1988), 857–881.

[28] Gundersen, H. J. G., and others, Some new, simple and efficient stereological methods and their use in pathological research and diagnosis, *Acta Pathol. Immun. Scand.* **96** (1988), 379–394.

[29] Gundersen, H. J. G., and E. B. Jensen, The efficiency of systematic sampling in stereology and its prediction, *J. Micros.* **147** (1987), 229–263.

[30] Hadwiger, H., *Vorlesungen ueber Inhalt, Oberflaeche und Isoperimetrie*, Springer-Verlag, Berlin, 1957.

[31] Hall, P., Resampling a coverage pattern, *Stoch. Processes Appl.* **20** (1985), 231–246.

[32] Hall, P., and R. L. Smith, The kernel method for unfolding sphere size distributions, *J. Comput. Phys.* **74** (1988), 409–421.

[33] Hall, Peter, *An Introduction to the Theory of Coverage Processes*, John Wiley and Sons, New York, 1988.

[34] Harding, E. F., and D. G. Kendall, eds., *Stochastic Geometry: A Tribute to the Memory of Rollo Davidson*, John Wiley and Sons, New York, 1974.

[35] Haug, H., *Nervenheilkunde* **4** (1985), 103–109.

[36] Haug, H., History of neuromorphometry, *J. Neurosci. Methods* **18** (1988), 1–17.

[37] Hoogendoorn, A. W., Estimate the weight undersize distribution for the Wicksell problem, *Statistica Neerlandica*, to appear, 1990.

[38] Howard, C. V., S. Reid, A. J. Baddeley, and A. Boyde, Unbiased estimation of particle density in the tandem-scanning reflected light microscope, *J. Micros.* **138** (1985), 203–212.

[39] Jensen, E. B., A design-based proof of Wicksell's integral equation, *J. Micros.* **136** (1984), 345–348.

[40] Jensen, E. B., and H. J. G. Gundersen, The stereological estimation of moments of particle volume, *J. Appl. Prob.* **22** (1985), 82–98.

[41] Jensen. E. B., and H. J. G. Gundersen, Fundamental stereological formulas based on isotropically orientated probes through fixed points with applications to particle analysis, *J. Micros.* **153** (1989), 249–267.

[42] Jensen, E. B., H. J. G. Gundersen, and R. Østerby, Determination of membrane thickness from orthogonal intercepts, *J. Micros.* **115** (1979), 19–33.

[43] Jensen, E. B., and R. Sundberg, Statistical models for stereological inference about spatial structures; on the applicability of best linear unbiased estimators in stereology, *Biometrics* **42** (1986), 735–751.

[44] Jensen, E. B., A. J. Baddeley, H. J. G. Gundersen, and R. Sundberg, Recent trends in stereology, *Int. Stat. Rev.* **53** (1985), 99–108.

[45] Kellerer, A. M., Exact formulas for the precision of systematic sampling, *J. Micros.* **153** (1989), 285–300.

[46] Kendall, D. G., On the number of lattice points inside a random oval, *Q. J. Math. (Oxford)* **19** (1948), 1–26.

[47] Kendall, D. G., and R. A. Rankin, On the number of points of a given lattice in a random hypersphere, *Q. J. Math. (Oxford)* **4** (series 2, 1953), 178–189.

[48] Kendall, M. G., and P. A. P. Moran, *Geometrical Probability*, Griffin's Statistical Monographs and Courses no. 10, Charles Griffin, London, 1963.

[49] König, D., N. Blackett, C. J. Clem, A. M. Downs, and J. P. Rigaut, Orientation distribution for particle aggregates in 3-D space based on point processes and laser scanning confocal microscopy, *Acta Stereologica* **8/2** (1990), 213–218.

[50] Kroustrup, J. P., H. J. G. Gundersen, and M. Vaeth, Stereological analysis of three-dimensional structure organisation of surfaces in multiphase specimens: statistical models and model-inferences, *J. Micros.* **149** (1988), 135–152.

[51] Matérn, B., Precision of area estimation: A numerical study, *J. Micros.* **153** (1989), 269–284.

[52] Matheron, G., *Les Variables Régionalisées et leur Estimation*, Masson, Paris, 1965.

[53] Matheron, G., *Random Sets and Integral Geometry*, John Wiley and Sons, New York, 1975.

[54] Mattfeldt, T., The accuracy of one-dimensional systematic sampling, *J. Micros.* **153** (1989), 301–313.

[55] Mattfeldt, T., H.-J. Möbius, and G. Mall, Orthogonal triplet probes: an efficient method for unbiased estimation of length and surface of objects with unknown orientation in space, *J. Micros.* **139** (1985), 279–289.

[56] Mattfeldt, T., and G. Mall, Estimation of length and surface of anisotropic capillaries, *J. Micros.* **135** (1984), 181–190.

[57] Mecke, J., and D. Stoyan, Formulas for stationary planar fibre processes I—General theory, *Math. Operationsforsch Stat., Ser. Stat.* **12** (1980), 267–279.

[58] Mecke, J., and D. Stoyan, Stereological problems for spherical particles, *Math. Nach.* **96** (1980), 311–317.

[59] Miles, R. E., On the elimination of edge-effects in planar sampling, pp. 228–247 in *Stochastic Geometry: A Tribute to the Memory of Rollo Davidson*, E. F. Harding and D. G. Kendall, eds., John Wiley and Sons, New York, 1974.

[60] Miles, R. E., The importance of proper model specification in stereology, pp. 115–136 in *Geometrical Probability and Biological Structures: Buffon's 200th Anniversary*, R. E. Miles and J. Serra, eds., Lecture Notes in Biomathematics, No. 23, Springer-Verlag, New York, 1978.

[61] Miles, R. E., and P. J. Davy, Precise and general conditions for the validity of a comprehensive set of stereological fundamental formulas, *J. Micros.* **107** (1976), 211–226.

[62] Miles, R. E., and P. J. Davy, On the choice of quadrats in stereology, *J. Micros.* **110** (1977), 27–44.

[63] Møller, J., Random tessellations in R^d, *Adv. Appl. Prob.* **21** (1989), 37–73.

[64] Nagel, W., Dünne Schnitte von stationäre räumlichen Faserprozessen, *Mathematisch Operationsforsch Stat., Ser. Stat.* **14** (1983), 569–576.

[65] Nicholson, W. L., Estimation of linear properties of size distributions, *Biometrika*, **57** (1970), 273–297.

[66] Nicholson, W. L., Estimation of linear functionals by maximum likelihood, *J. Micros.* **107** (1976), 323–336.

[67] Nicholson, W. L., Application of statistical methods in quantitative microscopy, *J. Micros.* **113** (1978), 223–239.

[68] Ogata, Y., and M. Tanemura, Likelihood analysis of spatial point patterns, *J. R. Stat. Soc., B* **46** (1984), 496–518.

[69] Preteux, F., and M. Schmitt, Boolean texture analysis and synthesis, pp. 379–400 in *Image Analysis and Mathematical Morphology. Volume II: Theoretical Advances*, J. Serra, ed., John Wiley and Sons, New York, 1988.

[70] Ripley, B. D., *Spatial Statistics*, John Wiley and Sons, New York, 1981.

[71] Ripley, B. D., *Statistical Inference for Spatial Processes*, Cambridge University Press, Cambridge, 1988.

[72] Ripley, B. D., Gibbsian interaction models, pp. 1–19 in *Spatial Statistics: Past, Present and Future*, D. A. Griffiths, ed., Image, New York, 1989.

[73] Ripley, B. D., and F. P. Kelly, Markov point processes, *J. London Math. Soc.* **15** (1977), 188-192.

[74] Rosiwal, A., Über geometrische Gesteinsanalysen, *Verh. K. K. Geol. Reichsanst. Wien*, 1898, 143.

[75] Santaló, L. A., *Introduction to Integral Geometry*, Actualités Scientifiques et Industrielles, no. 1198, Hermann, Paris, 1952.

[76] Santaló, L. A., *Integral Geometry and Geometric Probability*, Encyclopedia of Mathematics and Its Applications, vol. 1, Addison-Wesley, 1976.

[77] Serra, J., *Image Analysis and Mathematical Morphology*, Academic Press, New York, 1982.

[78] Silverman, B. W., M. C. Jones, J. D. Wilson, and D. W. Nychka, A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography, *J. R. Stat. Soc., B* **52** (1990).

[79] Sterio, D. C., The unbiased estimation of number and size of arbitrary particles using the disector, *J. Micros.* **134** (1984), 127–136.

[80] Stoyan, D., W. S. Kendall, and J. Mecke, *Stochastic Geometry and its Applications*, John Wiley and Sons, Chichester, 1987.

[81] Taylor, C. C., A new method for unfolding sphere size distributions, *J. Micros.* **132** (1983), 57–66.

[82] Tomkeieff, S. T., Linear intercepts, areas and volumes, *Nature* **155** (1945), 24.

[83] van Es, B., and A. Hoogendoorn, Kernel estimation in Wicksell's corpuscle problem, *Biometrika* **77** (1990), 139–145.

[84] von Blaschke, W., *Vorlesungen über Integralgeometrie*, Chelsea, New York, 1949.

[85] Watson, G. S., Estimating functionals of particle size distributions, *Biometrika* **58** (1971), 483–490.

[86] Watson, G. S., Characteristic statistical problems of stochastic geometry, in *Geometrical Probability and Biological Structures: Buffon's 200th Anniversary*, Lecture Notes in Biomathematics, no. 23, Springer-Verlag, New York, 1978.

[87] Weibel, E. R., *Stereological Methods, 1. Practical Methods for Biological Morphometry*, Academic Press, London, 1979.

[88] Weibel, E. R., *Stereological Methods, 2. Theoretical Foundations*, Academic Press, London, 1980.

[89] Weil, W., Stereology: a survey for geometers, pp. 360–412 in *Convexity and its Applications*, P. M. Gruber and J. M. Wills, eds., Birkhauser, Stuttgart, 1983.

[90] Weil, W., Point processes of cylinders, particles and flats, *Acta Applicandae Mathematicae* **9** (1984), 103–136.

[91] Weil, W., Expectation formulas and isoperimetric properties of Boolean models, *J. Micros.* **151** (1988), 235–245.

[92] Weil, W., Translative integral geometry, pp. 75–86 in *Geobild 89*, A. Hübler *et al.*, eds., Akademie Verlag, Berlin, 1989.

[93] Weil, W., Iterations of translative integral formulas and nonisotropic Poisson processes of particles, *Mathematische Zeitschrift*, to appear 1990.

[94] Wicksell, S. D., The corpuscle problem, I, *Biometrika* **17** (1925), 84–89.

[95] Wicksell, S. D., The corpuscle problem, II, *Biometrika* **18** (1926), 152–172.

Bibliographic Notes

Introductory references to stereology are [3,12,44] for statisticians, [87,88, 28,27,14] on applied stereology especially in biology, [80,48] on probabilistic modeling, and [89] from the viewpoint of convex geometry. Many research papers on stereology appear in the *Journal of Microscopy* and *Acta Stereologica*, which are the official journals of the International Society for Stereology; and in *Biometrika*, *Journal of Applied Probability*, and *Advances in Applied Probability*.

11
Markov Models for Speech Recognition

Alan F. Lippman
University of Washington

11.1 Introduction

The goal of speech recognition research is to enable machines to reduce human speech to a list of printed words. By imposing restrictions such as a limited vocabulary and grammar, automated speech recognition systems have been produced during the last 15 years that, within those constraints, approach human performance. Sustained research efforts have resulted in systems that place substantially fewer restraints on the speaker. Earlier recognition systems typically required that the words spoken belong to a fixed small vocabulary, that the speaker pause between each word, and that the speaker participate in a training period during which the system would automatically adjust to that particular speaker (and no other).

Each of the constraints above was used to reduce the complexity inherent in natural speech. This chapter presents an introduction to concepts underlying much of the work in speech recognition and, in the process, explains how the constraints above simplify the problem. The chapter then presents a detailed description of a simple probabilistic speech recognition system, modeled after the SPHINX system [14], that implements hidden Markov models (HMMs).

Hidden Markov models are the basis for many recently developed speech recognition systems and are related to Markov Random Fields (MRFs), which have been successfully applied to some image-processing tasks (see chapter 3). Both approaches rely on similar probabilistic frameworks to describe and exploit the relationship between the items of interest (e.g., the

microphone recording and its transcript, the degraded and "true" images). Both approaches share some fundamental tools: maximum likelihood estimation to estimate parameters, maximum *a posteriori* (MAP) estimation to perform recognition/restoration, and the use of Markovian models to make these estimation problems tractable. However, recorded speech, unlike images, is not a spatial process, it is a function of pressure (or through a microphone, voltage) versus time. This chapter provides a quite different view of some of the methods of chapter 3.

11.2 Basic Speech Concepts

Some of the basic units of speech are the sentence, the word, and the phoneme. There are approximately 40 phonemes, each corresponding to a distinctive sound. The symbol list that dictionaries provide after each word (as a pronunciation guide), is a list of phonemes. Phonemes, words, and sentences are all fundamental to the way we produce speech; people think in terms of sentences and words, while phonemic descriptions are necessary for people to pronounce words correctly. Modern speech recognition systems function in a similar way: a hierarchical approach is used where sentences are modeled as series of words; words are modeled in terms of phonemes; and phonemes are modeled as series of features of the signal. By nesting these three layers of models, printed sentences (our desired goal) are related to the speech signal.

Neither words nor phonemes are easily identified in the speech signal. Even finding the boundaries between two words can be a difficult task, since in the speech signal, phonemes as well as words may merge together and no simple signal processing can separate them. A spoken phrase like "this is," can easily be interpreted as "the sis" (or even by some graduate students as "thesis"). Most people have had the experience of listening to a foreign language and hearing only a steady stream of sounds, illustrating that one must understand speech to find word boundaries. (This type of difficulty is familiar to those who work in image segmentation; it is often impossible to find the boundaries between objects without some knowledge about what the objects are.)

Isolated-word recognition systems require that the boundaries between words be obvious. This is typically accomplished by requiring a speaker to pause between words. These silent pauses can be easily identified in the signal. Individually spoken words also tend to be enunciated more clearly,

aiding recognition. The main drawback of such systems is that the speaker is constrained to speak in a slow, unnatural manner. *Continuous-speech* recognition systems lack this constraint.

The construction of even an isolated-word speech recognizer is difficult. The speech signal associated with a word or a phoneme is extremely variable, and can vary greatly depending on both the speaker's identity and the manner in which the word was spoken. Variability is caused by anatomical differences between speakers, such as sex or vocal tract length, as well as by differences in style, health (presence or absence of a cold), speaking rate, stress, or accent. A quickly spoken word will frequently be slurred or have parts "missing." Accents can result in the wholesale shifting of vowels [9]. In addition, some words have many allowed (as opposed to recommended) pronunciations. This is especially true for common words, like "the," that are typically articulated poorly ("the" is often pronounced as "dee," "dah," "dih," "thah," or "thih"). *Speaker-dependent* systems simplify the recognition problem by adapting themselves to one particular speaker, removing some of the causes of variability.

The speech signal associated with a phoneme also varies depending on the context in which it is pronounced. This effect is called co-articulation. As people speak, the tongue, lips, and other articulators must move from a position necessary to pronounce the current phoneme to a position that will result in the next phoneme being produced. The articulators tend to move only far enough for speech to be intelligible, so current positioning is affected by the previous positioning. The placing of the articulators can also be influenced by the position that should be occupied next, a form of "thinking ahead" that people perform automatically.

For *small-vocabulary* recognition systems, the concept of phonemes typically is not used. Many examples of each word are used to discover a direct relationship between the word and the speech signal. In this way the effect of co-articulation is modeled implicitly. Since the required number of examples will grow linearly as a function of vocabulary size, this type of approach is almost impossible for vocabularies containing more than a thousand words. Phonemes, or some similar concept, are often employed by *large-vocabulary* systems.

Perhaps the most challenging problem in speech recognition research is that of modeling sentences. Unless words are enunciated very clearly, confusions between phonetically similar words are inescapable. While a person would pick the sentence that makes the most sense, an automated system must rely on a sentence model. Different types of models have been used,

ranging from classification trees [1] to N-step Markov chains [2] (the probability of the current word conditioned on all past words being only dependent on the previous N words). If the speaker obeys a highly constrained grammar specified by the system (e.g., the word "come" can only be followed with "here," "over," or "back"), it becomes much easier to automatically recognize sentences. A measure of the constraining power of a grammar is the perplexity (for a definition see [2]). Automated systems that allow large vocabularies and employ a grammar whose perplexity is close to the perplexity of spoken English, can be said fairly to handle *natural tasks.*

11.3 Some Recent Systems

All speech recognition systems require restricted input to achieve good accuracy (around one word in twenty wrong). Table 11.1 provides a short guide to some recent systems and the constraints under which they operate.

Almost universal is the requirement that the speech be recorded in a low-noise environment; a device that operates in a cocktail lounge or on a construction site may not be seen for decades. Other standard requirements are described by the terms *continuous speech*, *speaker independent*, *large vocabulary*, and *natural task*. Large vocabulary systems in this table recognize more than three hundred words.

TABLE 11.1: Some Speech Recognition Systems and Their Abilities

SYSTEM	DATE	Speaker Independence	Continuous Speech	Large Vocabulary	Natural Task
NTT	1975	No	No	No	No
DRAGON	1975	No	Yes	No	No
HEARSAY	1975	No	Yes	Yes	No
HARPY	1976	No	Yes	Yes	No
BELL '82	1982	Yes	No	No	No
FEATURE	1983	Yes	No	No	No
TANGORA	1985	No	No	Yes	Yes
BYBLOS	1987	No	Yes	Yes	No
BELL '88	1988	Yes	Yes	No	No
SPHINX	1988	Yes	Yes	Yes	No

11.4 Signal Processing

Speech signals are typically sampled at 8–16 kHz (8 to 16 thousand samples per second), where the value of each sample is specified by 10–12 bits. The first step in most recognition systems is to process this signal, both to reduce the amount of data and in an attempt to extract the features of the signal that are relevant for recognition.

Some processing methods begin by taking short-term Fourier transforms; short (on the order of 20 milliseconds), overlapping (by on the order of 10 milliseconds) frames of the signal are transformed into vectors whose components are the energies associated with (approximately 10) differing frequency bands. In this manner the speech signal would be represented by a small number (on the order of 100) of vectors per second. Other processing methods fit autoregressive (AR) models (of order 8 to 16) to these short, overlapping frames of the speech signal. In that approach the speech signal is represented by a sequence of AR parameter vectors. Note that whereas each frame of the signal may contain 320 values (20 ms of a 16 kHz sampled signal), this first processing step reduces it to a vector of dimension 16 or less.

The final step of most processing algorithms is the use of vector quantization [19], which reduces each vector to an *acoustic label* belonging to a small discrete set (containing on the order of 256 elements).

Briefly described, the use of vector quantization first requires that standard techniques be used to find cluster centers in a set of vectors obtained from a representative variety of recorded speech. Each of these cluster centers is given a label. Vector quantization replaces any vector in the high-dimensional space with the label of the closest cluster center. In this way, a 16-dimensional vector of AR parameters could be represented by a single byte.

Although good signal processing is critical to the successful performance of a recognition system, it is beyond the scope of this discussion, and we refer the reader to [2], [14], and [7] for further details. For the remainder of this discussion, it is assumed that the speech signal is already described by a series of acoustic labels, each of which belongs to a small, fixed set.

11.5 Probabilistic Recognition

The most successful approaches to the speech recognition problem use probabilistic modeling. The processed speech signal $\mathbf{y} = (y_1, \ldots, y_n)$ is considered to be an observation of n random variables (R.V.s) $\mathbf{Y} = (Y_1, \ldots, Y_n)$. A

sentence or string of words $\mathbf{w} = (w_1, \ldots, w_m)$ is considered to be an observation of m R.V.s $\mathbf{W} = (W_1, \ldots, W_m)$. For a fixed series of recorded acoustic labels $\mathbf{y}$, the value of the conditional distribution $P(\mathbf{W} = \mathbf{w} | \mathbf{Y} = \mathbf{y})$ specifies the probability that the words $\mathbf{w}$ are the "script" of the recording. Speech recognition can be accomplished by finding the word string that maximizes this conditional distribution. (This is MAP estimation.)

By Bayes' rule,

$$\begin{aligned} P(\mathbf{W} = \mathbf{w} | \mathbf{Y} = \mathbf{y}) &= P(\mathbf{Y} = \mathbf{y}, \mathbf{W} = \mathbf{w}) / P(\mathbf{Y} = \mathbf{y}) \\ &= P(\mathbf{Y} = \mathbf{y} | \mathbf{W} = \mathbf{w}) P(\mathbf{W} = \mathbf{w}) / P(\mathbf{Y} = \mathbf{y}) . \end{aligned}$$

For any fixed recording, the value of $P(\mathbf{Y} = \mathbf{y})$ is a constant and the $\mathbf{w}$ that maximizes $P(\mathbf{W} = \mathbf{w} | \mathbf{Y} = \mathbf{y})$ also maximizes $P(\mathbf{Y} = \mathbf{y}, \mathbf{W} = \mathbf{w})$ and $P(\mathbf{Y} = \mathbf{y} | \mathbf{W} = \mathbf{w}) P(\mathbf{W} = \mathbf{w})$. Instead of constructing the conditional distribution $P(\mathbf{W} = \mathbf{w} | \mathbf{Y} = \mathbf{y})$, we shall consider two, wholly separate problems. The first is the construction of a distribution $P(\mathbf{W} = \mathbf{w})$ on sentences; this is the modeling of grammar. The second is the construction of a distribution $P(\mathbf{Y} = \mathbf{y} | \mathbf{W} = \mathbf{w})$ on acoustic label strings; this is the modeling of speech production.

The remainder of this chapter is designed to provide a brief introduction to the techniques used to implement the above approach. The construction of $P(\mathbf{W} = \mathbf{w})$ is not discussed. (The interested reader should refer back to §11.2 for a few references regarding the choice of a grammar.) We concentate instead on presenting in some detail a simplified parametric model for $P(\mathbf{Y} = \mathbf{y} | \mathbf{W} = \mathbf{w})$ similar to the SPHINX baseline system [14], which formed the basis for the SPHINX system, a successful large-vocabulary, continuous speech, speaker-independent recognition system. (We recommend [14] as a detailed guide to the construction of a complex and functional speech recognition system.) Discussion follows on estimating the parameters of this distribution (§11.11). The final topic is the identification of the word string that has maximal probability conditioned on the spoken signal.

11.6 Image-Processing Parallels

This probabilistic approach to speech recognition has many points in common with Bayesian image processing using MRFs. A typical digital picture is a specification of intensities or colors at the sites of a finite two-dimensional lattice $L = \{(i,j)\}_{i,j=0}^{N}$. Modeling can be accomplished by introducing two

types of random variables, $\mathbf{F} = \{F_{ij}\}_{ij \in L}$, corresponding to an observed picture, and $\mathbf{U} = \{U_{ij}\}_{ij \in L}$, corresponding to an unobserved "true" picture. One method of image restoration is to calculate for any observed image $\mathbf{f}$ the $\mathbf{u}$ that maximizes $P(\mathbf{U} = \mathbf{u} | \mathbf{F} = \mathbf{f})$. Bayes' rule is applied, just as it was in speech, to reduce the construction of $P(\mathbf{U} = \mathbf{u} | \mathbf{F} = \mathbf{f})$ to the construction of two separate distributions: $P(\mathbf{U} = \mathbf{u})$ and $P(\mathbf{F} = \mathbf{f} | \mathbf{U} = \mathbf{u})$. $P(\mathbf{U} = \mathbf{u})$ is called the *prior* distribution and has the same role that $P(\mathbf{W} = \mathbf{w})$, the sentence model, did for speech. $P(\mathbf{F} = \mathbf{f} | \mathbf{U} = \mathbf{u})$ is the degradation model, and is the analogue of the speech production model.

Both image-processing tasks and speech recognition require the placement of distributions on very large ($\gg 2^{1000}$), but finite, spaces. Both rely on Markov assumptions to allow computational feasibility. Major differences are that the speech problem is inherently one-dimensional, whereas pictures are multidimensional. The inherent one-dimensionality of speech signals allows the use of fast estimation and search techniques. Although some image models allow the use of similar techniques [11], the class of such models is highly restricted and may not be particularly useful.

11.7 Modeling of Speech

The recognition system we describe uses phoneme models as an intermediate stage between acoustic labels and words. For every phoneme we will form a distribution on acoustic label strings produced during the enunciation of the phoneme. These distributions take the form of HMMs. In our simplified presentation, the effects of co-articulation will be ignored; the distribution of the acoustic labels associated with a given phoneme will be assumed to be independent of the neighboring phonemes.

A more ambitious speech recognition system would model phonemes in context. In such a system, the goal would still be to put a distribution on acoustic strings produced during the enunciation of the phoneme. However, the distribution would also depend (commonly) on the preceding and following phonemes (e.g., a distribution would be formed for the acoustic label strings associated with the phoneme \R\ when that phoneme is in the phoneme string \TH\ \R\ \IY\). The fundamental difficulty of this approach is that the number of such distributions would be approximately 40^3, and parameter estimation becomes impossible. However, clever implementations of context-dependent phoneme models have been made, and we refer the reader to [14] for details.

The use of phoneme models (either context-dependent or context-independent) usually necessitates the assumption that the distribution of the acoustic labels produced during a given phoneme, or that lies within a given context for context-dependent models, be independent of the acoustic labels produced during other phonemes. This assumption allows the production of acoustic labels for a string of phonemes to be considered as though each phoneme was produced independently and the results concatenated. This type of assumption is necessary in order to build models by "instantiation," a technique that is described in §11.10.

Although many words have multiple pronunciations, this model assumes that each word has a unique pronunciation, and therefore a unique phonemic representation. This assumption is used in some state-of-the art systems. Such an assumption forces the phoneme models to model implicitly some of the variability in pronunciation.

In the remainder of this chapter it is assumed that a grammar, and thus $P(\mathbf{W}=\mathbf{w})$, has been specified. The modeling strategies and assumptions above will be used to produce $P(\mathbf{Y}=\mathbf{y}|\mathbf{W}=\mathbf{w})$. Combining this with the value for $P(\mathbf{W}=\mathbf{w})$ yields $P(\mathbf{W}=\mathbf{w},\mathbf{Y}=\mathbf{y})$.

11.8 Hidden Markov Modeling

First, we introduce hidden Markov models (HMMs), and then describe their use in speech recognition. A HMM describes the behavior of two series of discrete R.V.s, call them $\mathbf{X} = (X_1, X_2, \ldots)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots)$. The $\mathbf{X}$ series is called the *hidden* R.V.s, and the other series is called the *observed* R.V.s. The conditional distribution of X_i given the values of all the previous R.V.s $(X_j, Y_j, \forall j < i)$ will be dependent only on the value of X_{i-1} (and not on i). The conditional distribution of Y_i, given the values of all other R.V.s (both hidden and observed), will be dependent only on the value of X_i and X_{i-1} (and not on i):

$$\begin{aligned} P(X_i=x_i|X_j=x_j, Y_j=y_j, \forall j<i) &= P(X_i=x_i|X_{i-1}=x_{i-1}) \\ P(Y_i=y_i|X_j=x_j, Y_j=y_j, \forall j\neq i) &= P(Y_i=y_i|X_i=x_i, X_{i-1}=x_{i-1}). \end{aligned}$$

The "hidden" part of a HMM arises from its use. Only observations of the R.V.s Y will be used to estimate the transition matrix $P(X_i|X_{i-1})$ and the conditional distribution $P(Y_i|X_i, X_{i-1})$. These conditional probabilities are sometimes called the *parameters* of the model.

For each phoneme, we construct a distribution on acoustic label strings $\mathbf{y}$ and hidden state strings $\mathbf{x}$. Based on these distributions we can construct the distribution $P(\mathbf{Y}=\mathbf{y}, \mathbf{X}=\mathbf{x}|\mathbf{W}=\mathbf{w})$. Note that this distribution specifies the distribution $P(\mathbf{Y}=\mathbf{y}|\mathbf{W}=\mathbf{w})$.

11.9 The SPHINX Phoneme Model

We now present a model for the production of one phoneme. Note that although a phoneme may be pronounced either quickly or slowly, the acoustic labels are produced at a constant rate. A phoneme model must allow the production of variable numbers of acoustic labels. As shown in Figure 11.1, let X_i take on seven allowed values, and call them $S_1, .., S_7$. X_1 equals S_1 with probability 1. In Figure 11.1, arrows connect two states S_i and S_j (possibly the same) if $P(X_k=S_j|X_{k-1}=S_i)$ is allowed to be nonzero. With every allowed transition (from S_i to S_j) there is an associated character, B, M, or E, denoting the beginning, middle, or end of the pronunciation of the phoneme. The distribution of Y_i conditioned on observations of all the other R.V.s will only depend on the character associated with the transition from X_{i-1} to X_i (e.g., from Figure 11.1, $P(Y_i=y_i|X_{i-1}=S_2, X_i=S_3) = P_M(Y_i=y_i)$). When S_7 is reached, the pronunciation of the phoneme is complete.

Note that P_B, P_M, and P_E are distributions on Y, and recall that an acoustic label can typically have any one of a few hundred values. The distributions P_B, P_M, and P_E will not be parametrized, so the specification of each distribution requires the specification of many probabilities. For the non-baseline SPHINX system, approximately 700 probabilities are needed to describe each of the distributions P_B, P_M, and P_E.

Hidden Markov models possess desirable properties. The observed R.V.s (Y) behave in a very simple fashion (they are independent) when conditioned on the hidden R.V.s (X), but the marginal distribution of Y possesses no such independence. The behavior of the observed R.V.s is quite complicated. The model above allows variable length label strings, and allows the probabilities of strings of length 1, 2, and 3 to be anywhere between 0 and 1. The loops at states S_2, S_3, and S_4 allow the phoneme to idle at a particular state, helping to model the various ways in which phonemes can be extended in duration (e.g., some phonemes may occasionally have long "middles," other phonemes may always have short "middles").

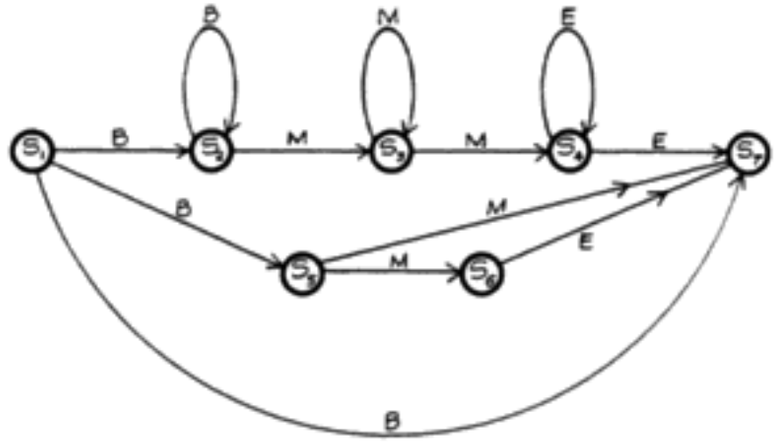


FIGURE 11.1: Allowed transitions.

11.10 Instantiation

We now use the phoneme models and the assumptions listed in the portion on modeling to construct $P(\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x}|\mathbf{W}=\mathbf{w})$. The distribution $P(\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x}|\mathbf{W}=\mathbf{w})$ is formed by using the values of the probabilities $P(X_t|X_{t-1})$ and $P(Y_t|X_t,X_{t-1})$ as specified by appropriate phoneme models. The manner in which this is done is called "instantiation," and is described below. For each sentence $\mathbf{w}=(w_1,w_2,..,w_m)$ there is an associated string of phonemes $\mathbf{p}=(p_1^1,p_1^2,\ldots,p_1^{n_1},p_2^1,\ldots,p_2^{n_2},\ldots,p_m^{n_m})$, where p_i^j is the j^{th} phoneme in the i^{th} word. The total number of phonemes associated with the sentence is $n=\sum_i n_i$. The distribution $P(\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x}|\mathbf{W}=\mathbf{w})$ is a HMM where the hidden variables X_i can take on values in

$$\{S_1(p_1^1),\ldots,S_6(p_1^1),S_1(p_1^2),\ldots,S_6(p_1^2),\ldots,S_1(p_m^{n_m}),\ldots,S_7(p_m^{n_m})\}\,.$$

This set contains $6n+1$ states, six for each phoneme and one state signifying the end of the sentence. Note that the state of the hidden variable X_i specifies the word, phoneme, and part of phoneme that is being produced at time i.

The distribution $P(\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x}|\mathbf{W}=\mathbf{w})$ is formed by defining the transition probabilities

$$P(X_t=S_j(p_k^l)|X_{t-1}=S_i(p_k^l)),\ 1\le i,j\le 6,\ 1\le k<m,\ 1\le l\le n_k$$

to be the same as those specified by the phoneme model for the phoneme

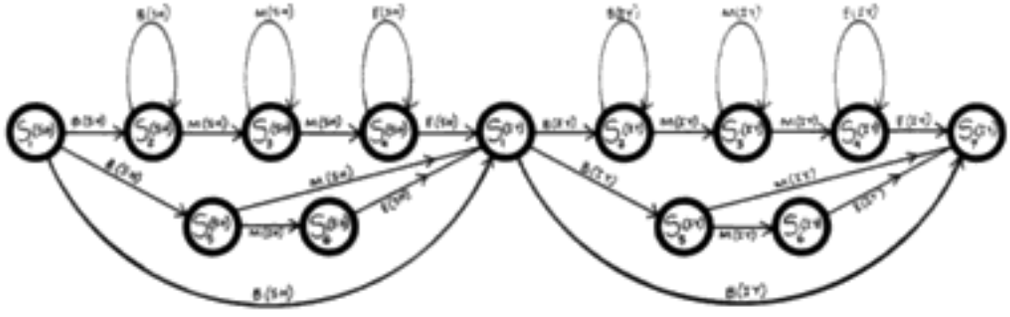


FIGURE 11.2: Graphical representation of the HMM for the word string *she*.

p_k^l. The values of

$$
\begin{aligned}
P(X_t = S_1(p_k^{l+1}) | X_{t-1} = S_i(p_k^l)), &\quad 1 \le i \le 6, \quad 1 \le k \le m, 1 \le l < n_k \\
P(X_t = S_1(p_{k+1}^1) | X_{t-1} = S_i(p_k^l)), &\quad 1 \le i \le 6, \quad 1 \le k < m, l = n_k \\
P(X_t = S_7(p_k^l) | X_{t-1} = S_i(p_k^l)), &\quad 1 \le i \le 6, \quad k = m, l = n_m
\end{aligned}
$$

have the values that transitions to S_7 had in the phoneme model for p_k^l. All the other transition probabilities have value zero. When X_t is observed to have value $S_7(p_m^{n_m})$, the sentence has been completed.

The distributions $P(Y_i|X_i, X_{i-1})$, similarly, are those associated with the related phoneme model. To account for between-word pauses, a nonzero term $P(X_t = S_1(p_k^1)|X_{t-1} = S_1(p_k^1))$ can be added where $P(Y_t = \text{silence}|X_t = S_1(p_k^1), X_{t-1} = S_1(p_k^1)) = 1$; for the sake of clarity, we will not discuss this detail.

Figure 11.2 is a graphical representation of the HMM for the word string *she*. Notice that the distribution $P(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x}|\mathbf{W} = \text{“she”})$ contains no explicit mention of phonemes or words. It is a distribution on strings of acoustic labels and hidden states.

As indicated previously, any string of hidden states $\mathbf{x}$ is associated with a unique word string. Note that this forces the value of $P(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w})$ to be either zero or equal to $P(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x})$.

11.11 Computational Issues: Parameter Estimation and Searches

The model for $P(\mathbf{W} = \mathbf{w}, \mathbf{Y} = \mathbf{y})$ is based on phoneme models. The estimation of parameters for this distribution consists primarily of estimating the parameters of the phoneme models. The simplest way to estimate the parameters of phoneme models would be to excise examples of phonemes from speech, and estimate parameters for each phoneme separately. This is not done, perhaps because the excision of phonemes (accomplished by hand) cannot be done in a manner consistent with the assumptions implicit in the phoneme models. Most current systems estimate parameters using training data consisting of utterances of whole known sentences. Below we present the algorithm used to perform this very difficult computational task. We also briefly present the typical approach used to speed the computations associated with the use of a trained system.

We wish to estimate the parameters of a HMM from one or more observations of $(Y_1, \ldots, Y_T)$. It may seem counter-intuitive that the parameters of a HMM can be estimated. We are, after all, trying to estimate the behavior of variables that are never observed. However, some thought yields examples of HMMs where we should be able to estimate parameters. For example, consider the simplest possible HMM. Let both the hidden and observed values take on only the values 0 and 1. Let $P(X_i = 1|X_{i-1} = 1) = .9$ and $P(X_i = 0|X_{i-1} = 0) = .7$, and let $P(Y_i = 1|X_i = 1) = .85$ and $P(Y_i = 0|X_i = 0) = .8$. These four terms completely describe the behavior of the HMM. An excerpt from a simulation of this HMM is below:

x : 00000000001100111111111111111101111110011111111111111111110111

y : 11000000001100111111111111111111111110011111101100101110101111

Notice that there are long strings of both 1s and 0s in the observation y. Remember that the Y_i are conditionally independent given X. The strings must be caused by the behavior of the hidden variables. Since long strings exist, we can guess that there is not much "noise," and that $P(Y_i = 0|X_i = 0)$ and $P(Y_i = 1|X_i = 1)$ should both be close to 1 or both close to 0. (The distribution of Y given X might turn most 0s to 1s, and most 1s to 0s. It is impossible to tell from the observations whether the underlying process is as above or is governed by $P(X_i = 0|X_{i-1} = 0) = .9$, $P(X_i = 1|X_{i-1} = 1) = .7$, $P(Y_i = 0|X_i = 1) = .85$ and $P(Y_i = 1|X_i = 0) = .8$.) We can then guess that both $P(X_i = 0|X_{i-1} = 0)$ and $P(X_i = 1|X_{i-1} = 1)$ should be $> .5$ in order

to consistently produce strings, and that one of these terms should be $> .8$ because a string of length 22 would otherwise be quite unlikely.

One of the reasons for the success of HMM is the existence of a computationally efficient method for approximate maximum likelihood parameter estimation (as opposed to the completely ad hoc estimation above). Starting with an initial guess for the parameters, and an observation $\mathbf{y} = (y_1, \ldots, y_T)$ of $\mathbf{Y} = (Y_1, \ldots, Y_T)$, the Baum-Welch algorithm [5,4] (also known of as the forward-backward algorithm) is an iterative method for selecting parameters that ensures that at every iteration the likelihood increases. Convergence to a local maxima of the likelihood is guaranteed. The Baum-Welch algorithm is equivalent to, and predates, the expectation maximization (EM) method [10].

We present here the Baum-Welch algorithm. Let

$$\begin{aligned} a_{jk} &= P(X_i = k | X_{i-1} = j) \quad j,k = 1,\ldots,s \\ b_{jk}(l) &= P(Y_i = l | X_i = k, X_{i-1} = j) \quad j,k = 1,\ldots,s, \quad l = 1,\ldots,r\,. \end{aligned}$$

Assume, for simplicity, that $a_j = P(X_0 = j)$ and $b_j(l) = P(Y_T = l | X_T = j)$ are known. Let P_{ab} be the distribution on $(X_0, \ldots, X_T, Y_1, \ldots, Y_T)$ generated by a and b. The likelihood of y is $\sum_x P(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x})$, which can be written as

$$P_{ab}(Y = y) = \sum_{j_0, j_1, \ldots, j_T = 1}^{s} a_{j_0} a_{j_0 j_1} b_{j_0 j_1}(y_1) a_{j_1 j_2} b_{j_1 j_2}(y_2) \cdots a_{j_{T-1} j_T} b_{j_T}(y_T)\,.$$

Note that a naive calculation of the expression above would require an extreme amount of computation, the performance of $2T \times s^{T+1}$ multiplication operations. A more efficient approach is to rewrite it as

$$P_{ab}(Y = y) = \sum_{j_0=1}^{s} a_{j_0} \sum_{j_1=1}^{s} a_{j_0 j_1} b_{j_0 j_1}(y_1) \sum_{j_2=1}^{s} a_{j_1 j_2} b_{j_1 j_2}(y_2) \cdots \sum_{j_T=1}^{s} a_{j_{T-1} j_T} b_{j_T}(y_T)$$

and perform the computation by first calculating $\sum_{j_T} a_{j_{T-1} j_T} b_{j_T}(y_T)$ and storing its value for all values of j_{T-1}. Then the sum over j_{T-1} can be computed and stored for all values of j_{T-2}. Repeating this until the likelihood is evaluated results in $2T \times S^2$ multiplications being conducted.

Each iteration of the Baum-Welch algorithm results in new estimates $\{\bar{a}_{jk}\}$ and $\{\bar{b}_{jk}(l)\}$, based on the data y and on the previous estimates $\{a_{jk}\}$ and $\{b_{jk}(l)\}$:

$$\bar{a}_{jk} = \frac{\sum_i P_{ab}(X_{i-1} = j, X_i = k | \mathbf{Y} = \mathbf{y})}{\sum_i P_{ab}(X_{i-1} = j | \mathbf{Y} = \mathbf{y})}$$

$$\bar{b}_{jk}(l) \quad = \quad \frac{\sum_{i:y_i=l} P_{ab}(X_{i-1}=j, X_i=k|\mathbf{Y}=\mathbf{y})}{\sum_i P_{ab}(X_{i-1}=j, X_i=k|\mathbf{Y}=\mathbf{y})} .$$

These re-estimation equations are the heart of the algorithm. For those more familiar with the EM algorithm, the above formulas can be interpreted as the calculation of $\sum_i E_{ab}(\mathcal{X}_{(j,k,l)}(X_i, X_{i-1}, Y_i)|\mathbf{Y}=\mathbf{y})$, where $\mathcal{X}_i$ denotes the indicator function for the value i; $\mathcal{X}_i(j) = 1$ if $i = j$, and 0 otherwise. These indicator functions are the natural sufficient statistics when a HMM is represented as an exponential family (which can be accomplished via the Gibbs-MRF equivalence). The maximization step of the EM algorithm becomes trivial in this case. We shall not present any proof that the above re-estimation increases the likelihood of $\mathbf{y}$; instead we refer the reader to [4,5].

The Baum-Welch algorithm in addition to the formulas above, specifies a method to calculate the new estimates quickly. This is essential since the distribution $P_{ab}(X_{i-1} = j, X_i = k|\mathbf{Y} = \mathbf{y})$ is dependent on all the values of $\mathbf{Y}$, and a naive calculation would require as many operations as a naive calculation of the likelihood. However, we can implement a computational strategy similar to that introduced above to compute the likelihood. The re-estimation equations above can be written as

$$\begin{aligned} \bar{a}_{jk} \quad &= \quad \frac{\sum_i \alpha_i(j)\beta_{i+1}(k)a_{jk}b_{jk}(y_{i+1})}{\sum_i \alpha_i(j)\beta_i(j)} \\ \bar{b}_{jk}(l) \quad &= \quad \frac{\sum_{i:y_i=l} \alpha_i(j)\beta_{i+1}(k)a_{jk}b_{jk}(y_{i+1})}{\sum_i \alpha_i(j)\beta_{i+1}(k)a_{jk}b_{jk}(y_{i+1})} , \end{aligned}$$

where α and β can be defined inductively in i, forward and backward respectively, by

$$\alpha_{i+1}(k) = \sum_{j=1}^{s} \alpha_i(j)a_{jk}b_{jk}(y_{i+1}) , \quad \beta_i(j) = \sum_{j=1}^{s} \beta_{i+1}(k)a_{jk}b_{jk}(y_{i+1}) .$$

The implementation of the Baum-Welch algorithm for speech recognition depends, obviously, on the form of training data. The standard scenario, as mentioned above, is that there is a list of known sentences and pronunciations of them. We can therefore construct a HMM $P(\mathbf{Y}=\mathbf{y}, \mathbf{X}=\mathbf{x}|\mathbf{W}=\mathbf{w})$ for each known sentence (as in §11.10). The estimation of parameters for the HMMs can proceed by the Baum-Welch algorithm with the simple modification that the iterations be performed synchronously. One iteration will be conducted for each HMM sentence model, then the estimated parameters

for each sentence will be combined into one estimate for all the parameters of all the phoneme models at the end of every iteration.

The performance of the Baum-Welch algorithm depends highly on the quality of the initial estimates. While every iteration of the algorithm guarantees an increase in the likelihood, a bad initial guess may cause convergence to a bad (low) local maximum of the likelihood. Whereas whole sentences are used to train the phoneme models, the initial estimates for the distributions $P(Y_i|X_i, X_{i-1})$ often come from excised phoneme data. In [15], for every phoneme the three distributions $P_B(Y_i)$, $P_M(Y_i)$, and $P_E(Y_i)$ are initialized to a histogram of the acoustic label values associated with that phoneme in hand-labeled data. The initial estimates of the transition probabilities of the hidden states were chosen so that all allowed transitions from a state had equal probability.

Another interesting implementational detail is that the Baum-Welch algorithm is typically used for only 2 or 3 iterations [7, page 35], [14]. On the other hand, EM is well known for slowness after the first few iterations. In [14] the performance of the recognition system is stated to worsen with continued iterations, suggesting to us that an overfit of the training data is occurring.

Once parameter estimation is accomplished, there remains the use of the recognition system. As was stated at the beginning of this section, our goal is the calculation of the string $\mathbf{w}$ that maximized $P(\mathbf{W} = \mathbf{w}|\mathbf{Y} = \mathbf{y})$. The use of Bayes' rule allowed us to modify this to the calculation of the string $\mathbf{w}$ that maximized $P(\mathbf{Y} = \mathbf{y}, \mathbf{W} = \mathbf{w})$. Recall that we have constructed $P(\mathbf{Y}=\mathbf{y}, \mathbf{X}=\mathbf{x}, \mathbf{W}=\mathbf{w})$, which equals $P(\mathbf{Y}=\mathbf{y}, \mathbf{X}=\mathbf{x})$ when $\mathbf{w}$ is the word string associated with $\mathbf{x}$, and zero otherwise.

The string $\mathbf{w}$ that maximizes $P(\mathbf{Y}=\mathbf{y}, \mathbf{W}=\mathbf{w})$ is usually approximated by the $\mathbf{w}$ associated with the $\mathbf{x}$ that maximizes $P(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x})$. The principal justification for this approximation, besides its success and computational simplicity, is that the most likely word string should have at least one corresponding hidden state string that will also be very likely. The most likely string of hidden states, for small vocabulary and simple grammar systems, can be found by a simple dynamic programming [6] scheme called the Viterbi algorithm [2]. For more complicated systems the search is performed by more ad hoc methods that are based on dynamic programming [2,14]

11.12 Future Directions

The construction of a large-vocabulary, speaker-independent, complicated-grammar speech recognition system from scratch is a daunting task. How-

ever, it is one that new researchers interested in speech recognition will not have to face. Databases for speech recognition are becoming commonly available. As a result, various fascinating and extremely challenging subproblems can be approached by single researchers on current generation workstations. One such problem is the speaker-independent recognition of phonemes in continuous speech; another is the recognition of connected digits.

Whereas HMMs have been the most successful approach to date, the fundamental reason for their current superiority is the dedication and creativity of those who have implemented them. Preliminary research indicating that other approaches can be as accurate and computationally feasible is presented in [17]. It is hoped that, as the computational resources to approach the speech recognition problem become available to a larger community, a diversification of approaches will occur and that this chapter encourages research in this direction.

Bibliography

[1] Bahl, L. R., P. F. Brown, P. V. De Souza, and R. L. Mercer, A tree based statistical language model for natural language speech recognition, *IEEE Trans. Acoust. Speech Signal Processing* **37** (1989), 1001–1008.

[2] Bahl, L. R., F. Jelinek, and R. L. Mercer, A maximum likelihood approach to continuous speech recognition, *IEEE Trans. Pattern Anal. Machine Intel.* **5** (1983), 179–190.

[3] Baker, J. K., The DRAGON system—An overview, *IEEE Trans. Acoust. Speech Signal Process.* **23** (1975), 24–29.

[4] Baum, L. E., An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes, in *Inequalities III*, Academic Press, 1972.

[5] Baum, L. E., T. Petrie, G. Soules, and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math. Stat.* **41** (1970), 164–171.

[6] Bellman, R. E., *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[7] Brown, P. F., *The Acoustic-Modeling Problem in Automatic Speech Recognition*, Ph.D. thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, 1987.

[8] Cole, R. A., R. M. Stern, M. S. Phillips, S. M. Brill, P. Specker, and A. P. Pilant, Feature-based speaker independent recognition of English letters, *IEEE Inter. Conf. Acoust. Speech Signal Process.* (Oct. 1983).

[9] Cooke, P., Are accents out?, *New York Times Magazine* (Nov. 19, 1989), 50.

[10] Dempster, A. P., N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc., B* **39** (1977), 1–38.

[11] Devijver, P. A., and M. M. Dekesel, Learning the parameters of a hidden Markov random field image model: A simple example, pp. 141–163 in *Pattern Recognition Theory and Applications*, P. A. Devijver and J. Kittler, eds., NATO ASI Series, vol. F30.

[12] IBM speech recognition group, A real-time, isolated word, speech recognition system for dictation transcription, *IEEE Inter. Conf. Acoust. Speech Signal Process.* (Mar. 1985).

[13] Itakura, F., Minimum prediction residual principle applied to speech recognition, *IEEE Trans. Acoust. Speech Signal Process.* **23** (1975), 67–72.

[14] Lee, Kai-Fu, *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*, Ph.D. thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, 1988.

[15] Lee, Kai-Fu, and H. W. Hon, Speaker-independent phone recognition using hidden Markov models, *IEEE Trans. Acoust. Speech Signal Process.* **37** (1989), 1641–1648.

[16] Lesser, V. R., R. D. Fennell, L. D. Erman, and R. D. Reddy , The Hearsay II speech understanding system, *IEEE Trans. Acoust. Speech Signal Process.* **23** (1975), 11–24.

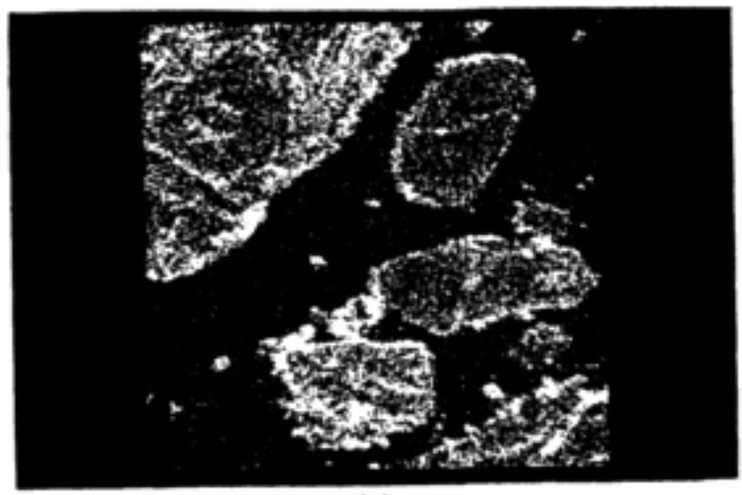
[17] Lippman, A., Data determined Markov models for speech recognition, *Proceedings of the 1990 Asilomar Conference on Signals, Systems, and Computers* (to appear).

[18] Lowerre, B. T., *The HARPY Speech Recognition System*, Ph.D. thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, 1976.

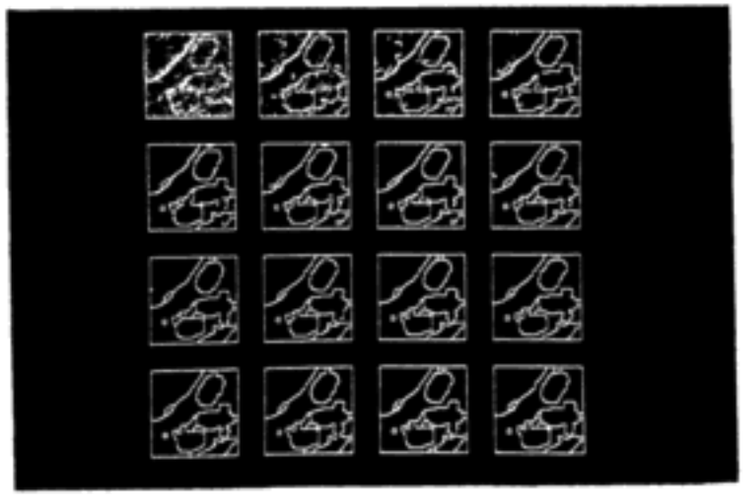
[19] Makhoul, J., S. Roucos, and H. Gish, Vector quantization in speech coding, *Proc. IEEE* **73** (1985), 1551–1588.

[20] Rabiner, L. R., J. G. Wilpon, and F. K. Soong, High performance connected digit recognition using hidden Markov models, *IEEE Int. Conf. Acoust. Speech Signal Process.* (Apr. 1988).

[21] Wilpon, J. G., L. R. Rabiner, and A. Bergh, Speaker-independent isolated word recognition using a 129-word airline vocabulary, *J. Acoust. Soc. Amer.* **72** (1982), 390–396.

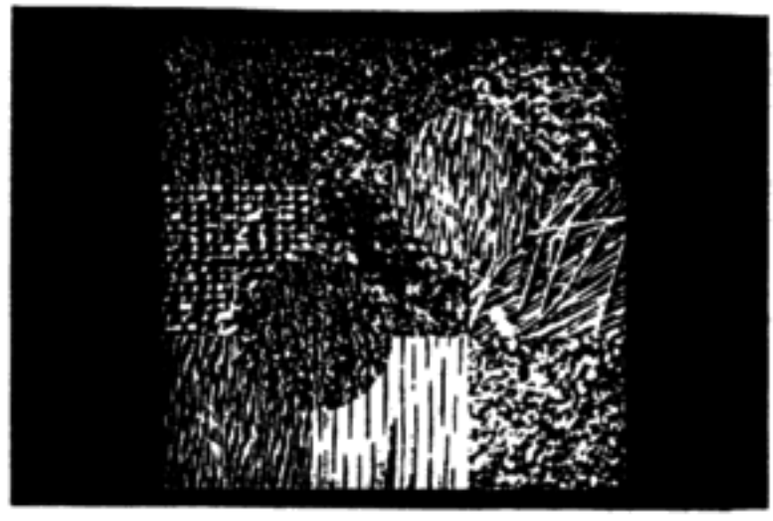


(a)

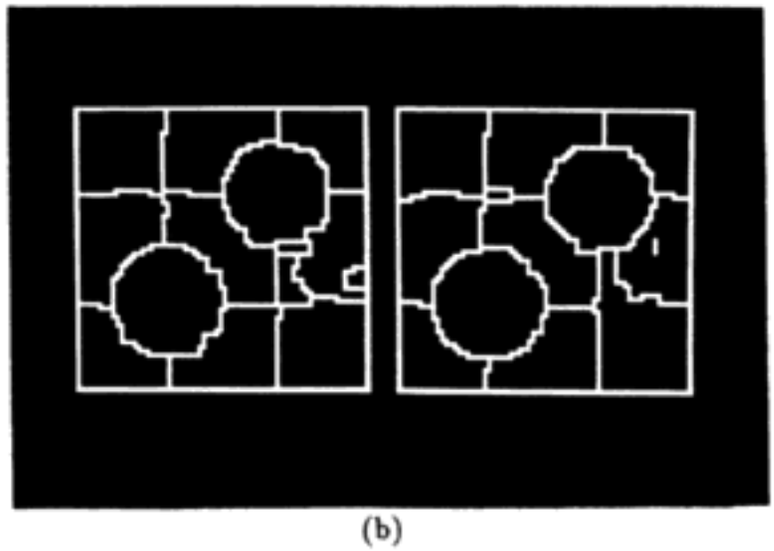


(b)

PLATE 2.1: L-band synthetic aperture radar (SAR) image of ice floes in the ocean: (a) original image, 512×512 (pixel resolution is about 4m $\times$ 4m), (b) evolution of segmentation via stochastic relaxation with constraints; shown are sixteen "snapshots" from sixty sweeps (every third sweep) of stochastic relaxation (upper left panel shows the random starting configuration of edges, and the lower right panel shows the final configuration of the boundaries). Reprinted, by permission, from Geman *et al.* (1990).



(a)



(b)

PLATE 2.2: Collage composed of nine Brodatz textures: leather, grass, and pigskin (top row), raffia, wool, and straw (middle row), and water, wood, and sand (bottom row). Two of the textures, leather and water, are repeated in the two circles; (a) original image 384 x 384, individual textures all 128 x 128; (b) estimated boundaries via deterministic (left panel) and stochastic (right panel) algorithms. Reprinted, by permission, from Geman *et al.* (1990).

PLATE 2.3: Single photon emission computer tomography (SPECT) reconstruction of a slice of a human skull across the eyes, from real (hospital) data: (a) filtered back projection (FBP) reconstruction, (b) reconstruction via the iterative conditional expectations (ICE) algorithms using $\beta = 2.7$, the ML estimator. Note that in (b) one can distinguish details such as nose bone, eyes, and brain, most of which cannot be distinguished in (a).

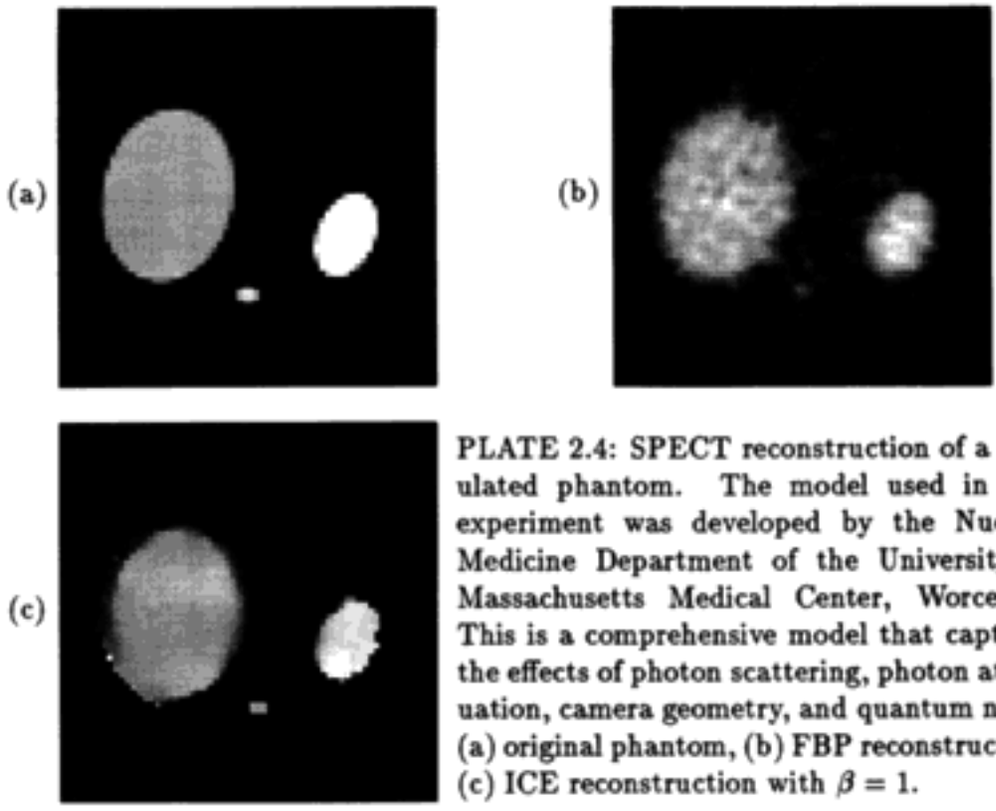


PLATE 2.4: SPECT reconstruction of a simulated phantom. The model used in this experiment was developed by the Nuclear Medicine Department of the University of Massachusetts Medical Center, Worcester. This is a comprehensive model that captures the effects of photon scattering, photon attenuation, camera geometry, and quantum noise: (a) original phantom, (b) FBP reconstruction, (c) ICE reconstruction with $\beta = 1$.

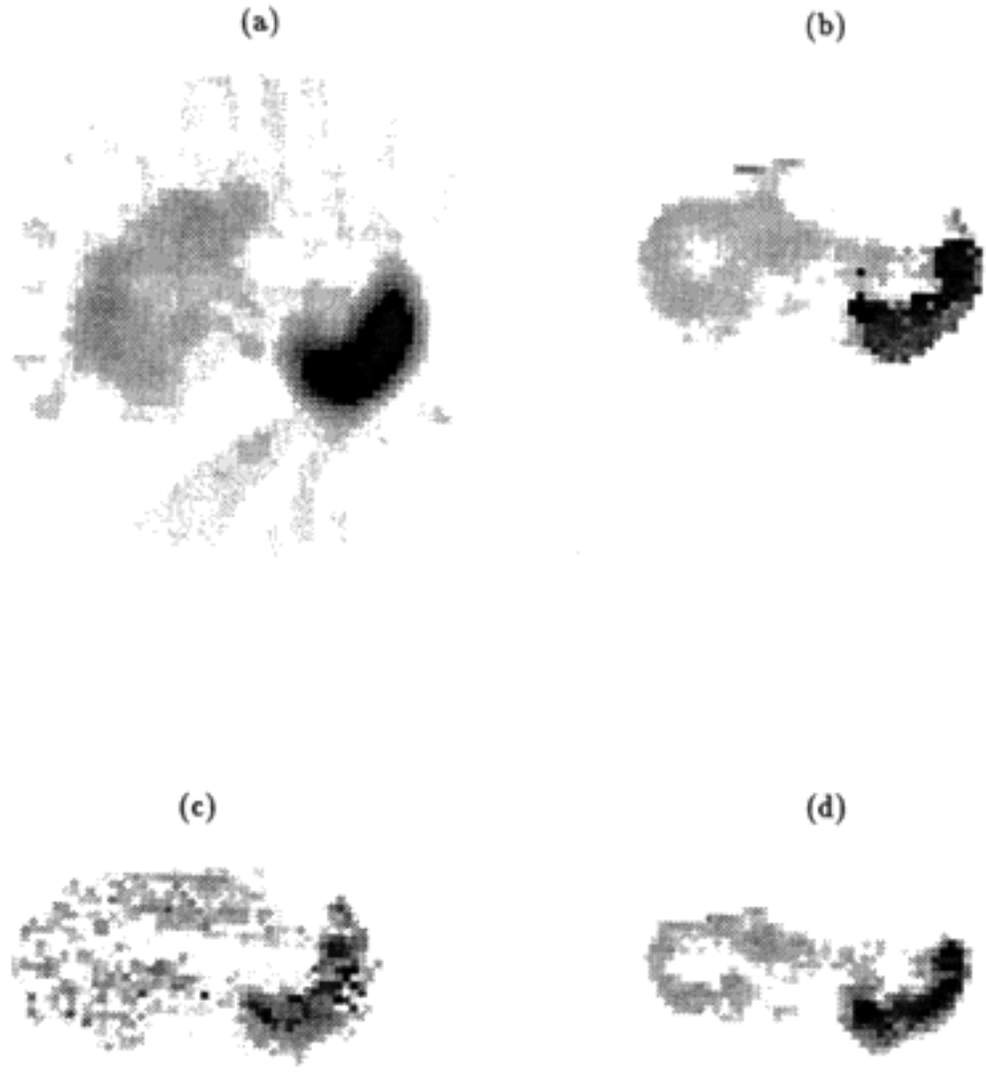


PLATE 2.5: SPECT reconstruction of a human liver/spleen scan, from real (hospital) data: (a) FBP reconstruction, (b) ICE reconstruction with $\beta = 3$, the ML estimator, (c) ICE reconstruction with $\beta = 0$, (d) ICE reconstruction with $\beta = 20$; (c) and (d) demonstrate the significance of the parameter β (see text).

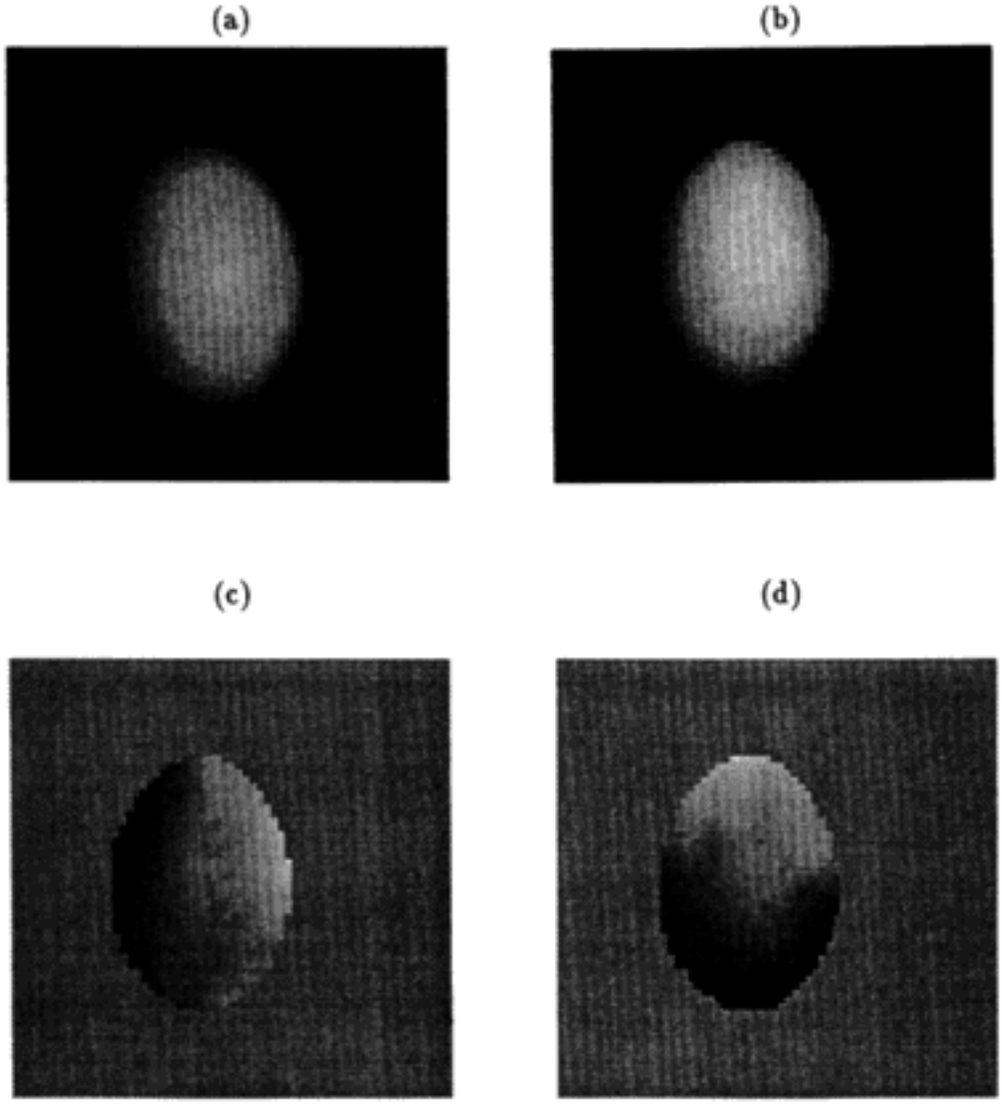


PLATE 2.6: A shape-from-shading experiment with an egg image under uncontrolled illumination. The surface of the egg was assumed to be Lambertian with unknown albedo; the algorithm (a combination of constrained annealing and iterative conditional modes (ICM)) estimated, in addition to the configuration N of unit normals, the albedo ρ of the egg (and of the background) and an effective light source direction $\vec{S}$: (a) original image, 64×64, (b) reconstruction (simultaneous estimation of N, ρ, and $\vec{S}$), (c) reconstructed scene illuminated from the x-direction, (d) reconstructed scene illuminated from the y-direction.

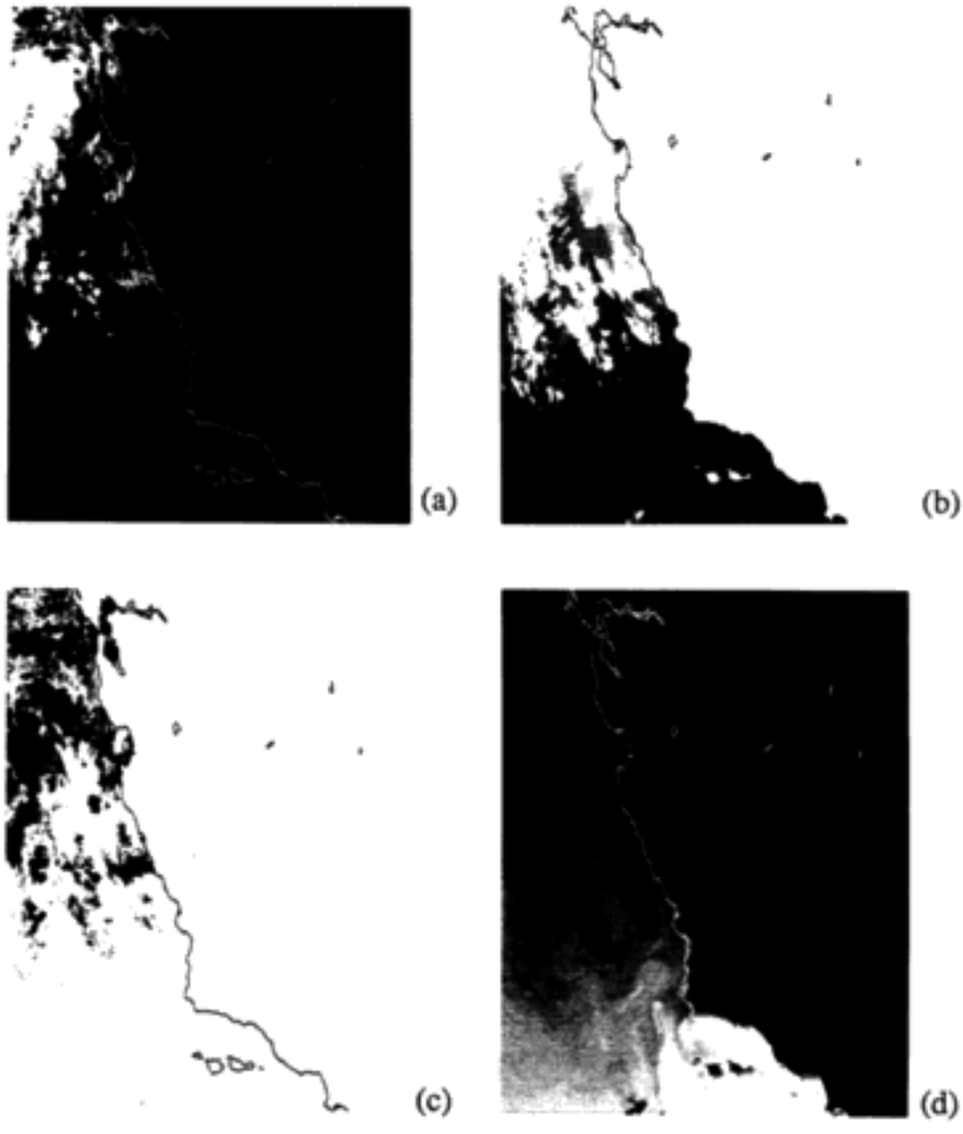
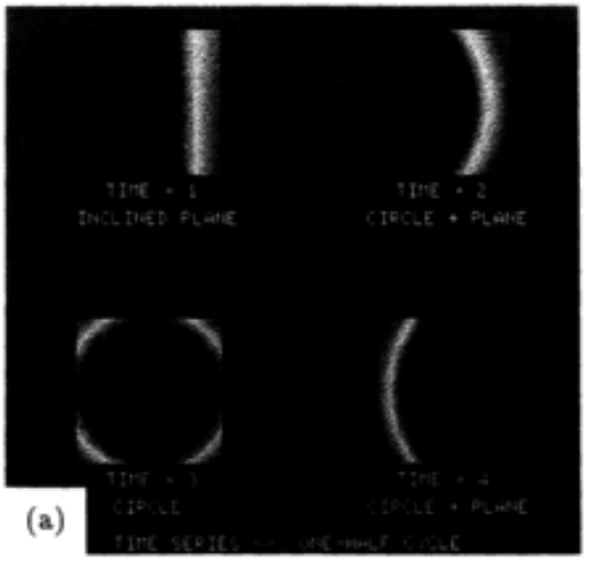


PLATE 3.1: (a) AVHRR band 2 (albedo) with clouds shown as white. (b) AVHRR band 4 (infrared temperature), dark gray scales are warm. (c) Segmented image produced by the PCTSMC algorithm. (d) Final cloud-masked image (clouds and land are black) produced by the PCTSMC algorithm. Details of the different gray scale maps used in the panels of Plate 3.1 are given in the text.



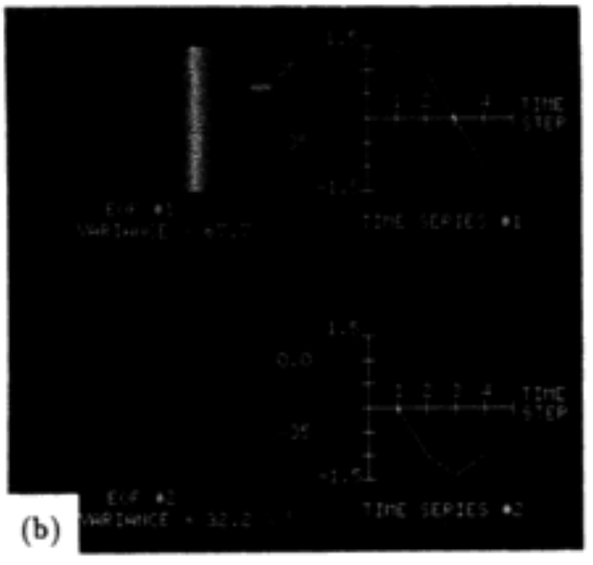


PLATE 3.2: (a) One-half cycle of the image sequence constructed by superposition of an image of an inclined plane with that of a circle. Note, the circle is out of phase with the inclined plane and the range of data in the inclined plane is about twice that of the circle. (b) Dominant patterns of variance determined from the EOF analysis of the image sequence.

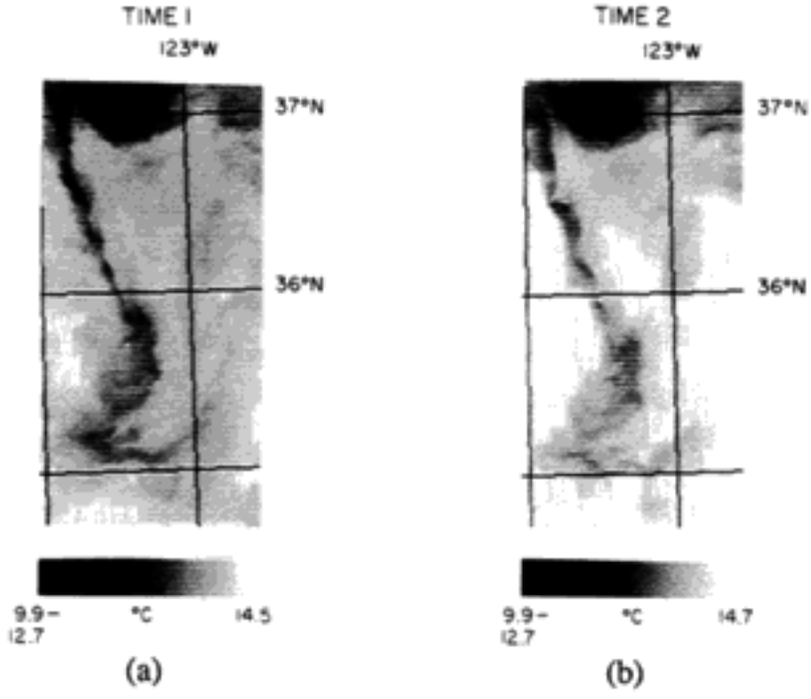


PLATE 3.3: Image sequence and geographical information of a coastal filament observed off central California by the Advanced Very High Resolution Radiometer on the polar-orbiting NOAA-9 and -10 satellites. Individual gray scale mappings were used to optimize feature recognition. The lower temperature ranges (i.e., 9.9°–12.7°C for time step 1) were mapped to a single gray scale. (Wahl and Simpson, 1990b)

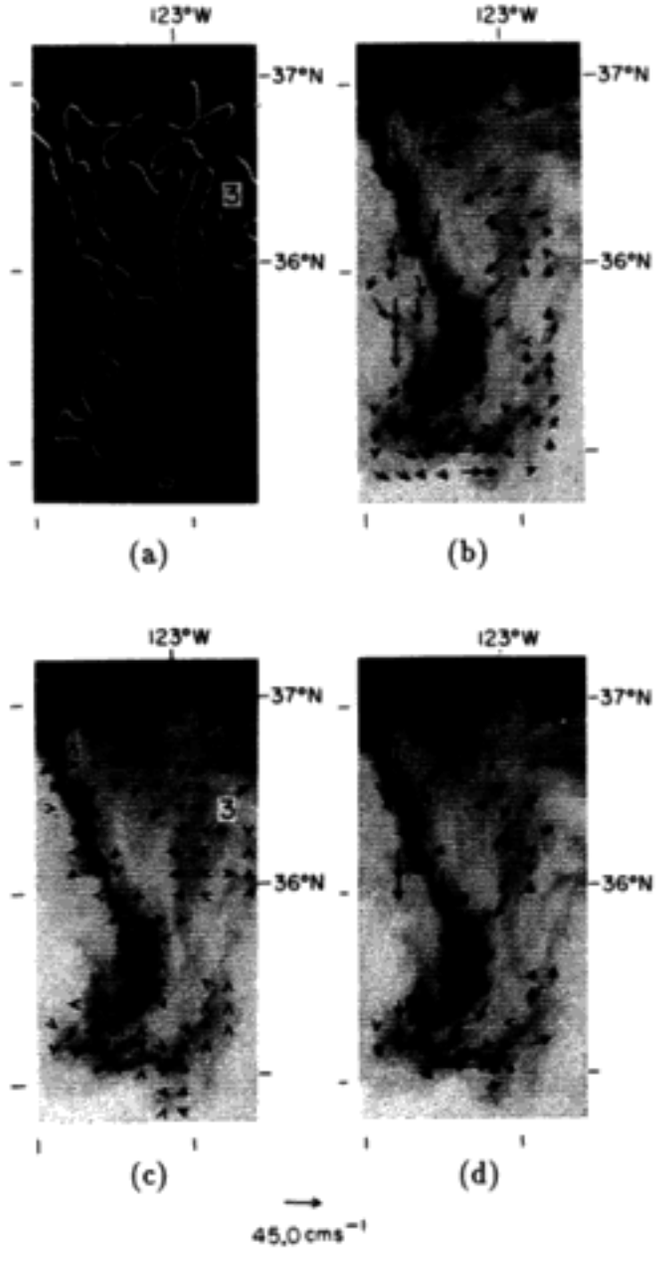
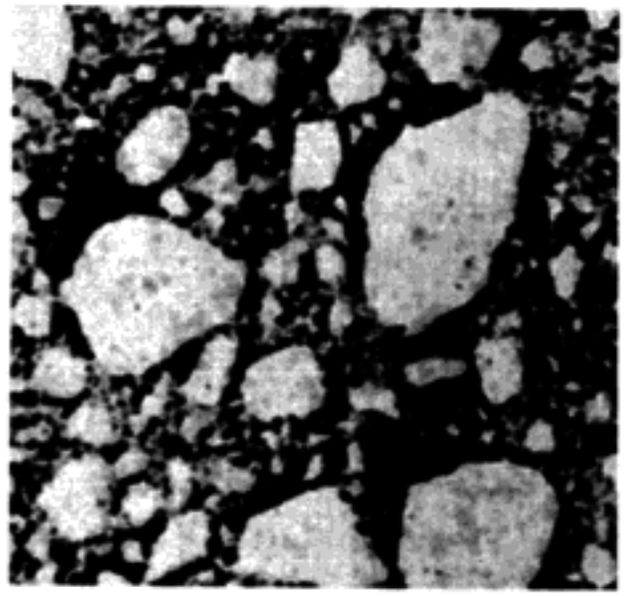
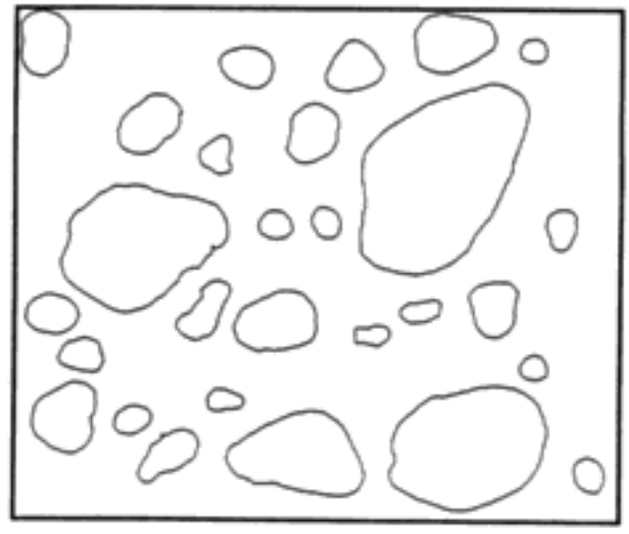


PLATE 3.4: (a) Edge maps. (b) Total velocity field from pattern-matching. (c) The MU normal component of velocity. (d) The tangential component of velocity computed as a difference of (b) and (c) for time step 2 of the image sequence shown in Plate 3.3. (Wahl and Simpson, 1990b)



(a)



(b)

PLATE 3.5: (a) A polar Landsat image showing ice floes as light gray structures against a dark background. (b) The corresponding distribution, size, and shape of the ice floes. Reprinted, by permission, from Banfield and Raftery (1989).

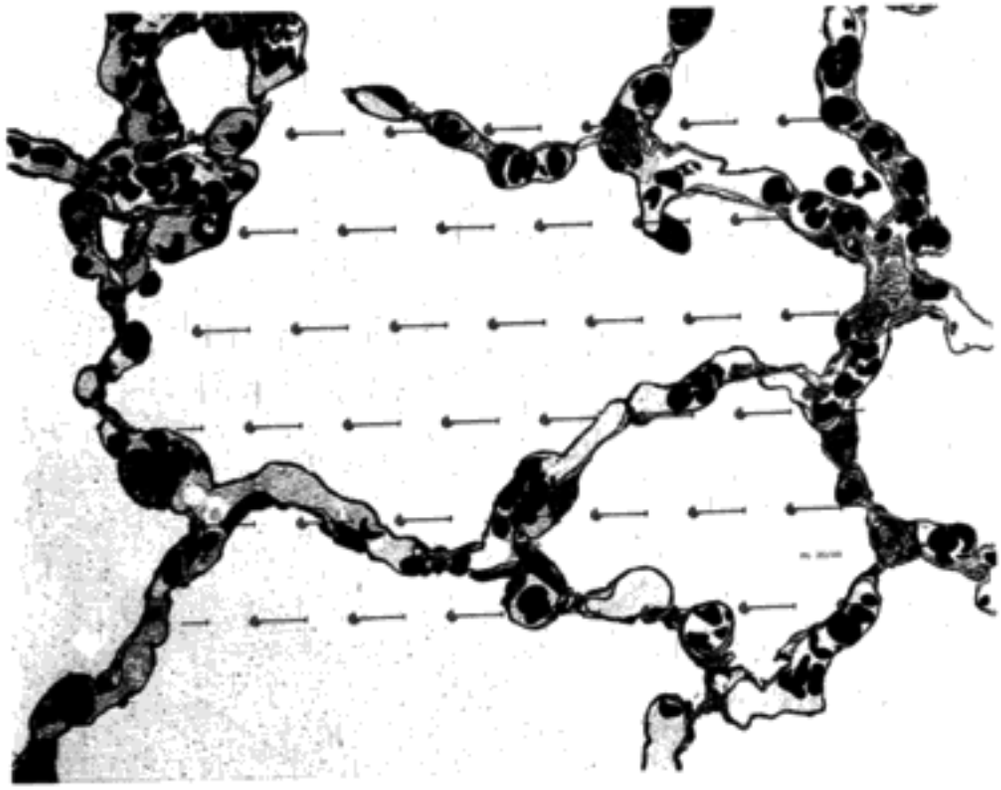


PLATE 10.1: Plane section of a biological structure with stereological test system superimposed. Lung of Grant's gazelle; white space is airway, dark blobs are red blood cells. Microtome thin section, optical microscope image field, magnification ×1500. Standard test system on transparency, randomly translated over photographic print. Reprinted, by permission, from Cruz-Orive and Weibel (1981).

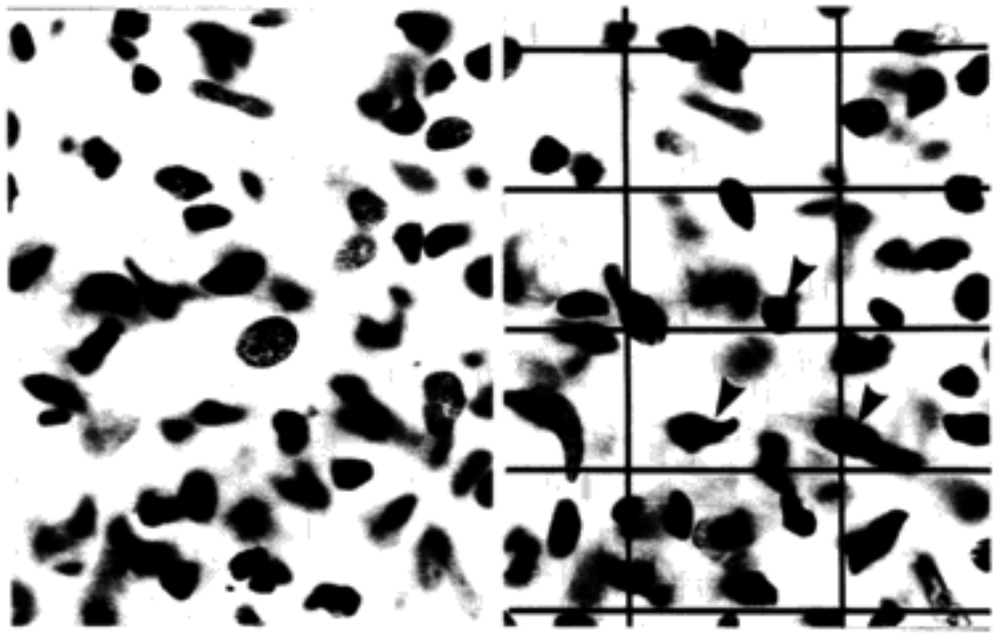


PLATE 10.2: A disector sample formed by two optical section planes. Human renal glomerulus; dark blobs are nuclei. At left is the look-up section; at right the counting section, with a tessellation of rectangular counting frames superimposed (randomly translated). Arrows indicate nuclei counted by the disector/tiling rule. Optical microscope, Hematoxylin-Giemsa stain, magnification $\times 1140$, section separation 4 μm. By kind permission of Dr. Niels Marcussen, University of Aarhus, Denmark.