GEOGRAPHIC INFORMATION ANALYSIS

second edition

DAVID O'SULLIVAN

DAVID J. UNWIN

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

GEOGRAPHIC INFORMATION ANALYSIS

GEOGRAPHIC INFORMATION ANALYSIS

Second Edition

David O'Sullivan and David J. Unwin

WILEY

JOHN WILEY & SONS, INC.

This book is printed on acid-free paper. ♾

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the Publisher and the author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the Publisher nor the author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information about our other products and services, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

O'Sullivan, David, 1966-
Geographic information analysis / David O'Sullivan, David Unwin. – 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-28857-3 (cloth)
1. Geographic information systems. 2. Spatial analysis (Statistics)
I. Unwin, D. (David John) II. Title.
G70.212.O88 2010
910.285–dc22
2009042594

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

David O'Sullivan: for Mum and Dad
David Unwin: for Polly, winner RHS Gold Medal Chelsea
Flower Show 2009

Contents

Preface to the Second Edition

The first edition of this text (O'Sullivan and Unwin, 2003) was written in the first two years of the twenty-first century, but in its basic framework it relied on two key ideas that have a long history in geographic information science: cartography and statistics. Perceptive (and aged?) readers will perhaps have noted sections that owe their origins to *Introductory Spatial Analysis* (Unwin, 1981), a little book one of us wrote almost 30 years ago. The first key idea was the use of a framework for describing geographic objects by their dimension of length into points, lines, areas, and continuous surfaces (fields); the second was to regard mapped distributions as realizations of some spatial stochastic process.

Like any overarching framework, it's not perfect. For example, heavy reliance on fixed geometric entities and on close attention to statistical hypothesis testing may seem dated in the light of developments in spatial representation and statistical inference. In developing this second edition, we thought for some time about moving with the times and adopting one of a variety of alternative frameworks. Our eventual decision to stick to the original blueprint was not taken lightly, but we are sure we have done the right thing. Since the first edition appeared, we have taught classes with curricula built on this framework at senior undergraduate and beginning graduate levels in the United States, the United Kingdom, New Zealand, and globally over the Internet, and we have found it to be pedagogically clear and resilient. For many students, either spatial data or statistical reasoning (and, not infrequently, both) are new or fairly new concepts, and it is important that we provide a "way in" to more advanced topics for that large segment of readers. For those few readers happily familiar with both topics, we hope that the book is broad enough in its coverage and makes enough nods in the direction of more advanced material to remain useful.

CHANGES

In spite of broad continuities, a chapter-by-chapter comparison will show substantial updates in our treatment of point pattern analysis, spatial autocorrelation, kriging, and regression with spatial data, and we have made a number of larger changes, some of which are of emphasis and some of which are more substantial. Those familiar with the first edition will notice that formal hypothesis testing has receded further into the background in favor of greater emphasis on Monte Carlo/randomization approaches that, in most practical work, using the usually messy data that we have to handle, seem to us to offer ways around many of the well-known problems related to spatial data. Even in the seven years since the first edition was published, computing power has made this approach more practical and easier to implement. Second, readers will find that throughout the revised text there is a greater emphasis on essentially *local* descriptions. We believe this also reflects an important methodological change within the science, arguably made possible by access to today's computing environments.

The two most substantial changes follow from and relate back to these changes of emphasis, in that we have added entirely new chapters (Chapters 3 and 8) on *geovisualization* and *local statistics*. In fact, a chapter on maps and mapping was written for the first edition but was not included. This we justified to ourselves by noting the need to keep the length of the book down and by considering that most of our readers would be familiar with some of the central concepts of the art and science of cartography. Subsequent experiences teaching courses on geographic information analysis to what one of us has called "accidental geographers" (Unwin, 2005) have shown that this omission was a mistake. By *accidental geographer*, we mean those new to the analysis of spatial data, whose understanding of geographic science is based largely on the operations made possible by geographic information system (GIS) software. Cartography, or, if you prefer, geovisualization, has added relevance for three reasons. First, even with the enormous range of statistical methods that are available, some form of mapping remains perhaps the major analytical strategy used. Second, an emphasis on local statistics that are then mapped has increased the need for understanding basic cartographic principles. Third, as a walk around almost any GIS trade exhibition will show, otherwise sophisticated GIS users continue to make quite basic cartographic errors. Our new Chapter 3 bears little resemblance to the one originally drafted. The new materials rely heavily on the use of an Internet search engine to find and critique examples, something the senior author was taught almost half a century ago in a student class on map appreciation. That said, we have also tried to locate much of the chapter in the long, and regrettably sometimes neglected, cartographic tradition.

The second major addition is a chapter on local statistics. Again, this is not without its organizational problems, since, as we have discovered, a considerable proportion of materials originally developed in different contexts can plausibly be brought into this framework. Examples include all the materials associated with concepts of distance, adjacency, and neighborhood that go into the definition of geographic structure (**W**) matrices; estimation of the mean height of a field from control point data (spatial *interpolation*); identification of local peaks in the estimated intensity of a point process (*clustering*); and the identification of groups of similar zones by decomposition of a global Moran's I measure of spatial autocorrelation (*Moran scatterplot*). Readers will doubtless find other examples, a sign of the centrality of the concept in much spatial analysis. This chapter provides a more explicit treatment of various local indicators of spatial association and allows us to include an introduction to the ideas behind geographically weighted regression (GWR). Although kernel density estimation (KDE) might easily be placed in this same chapter, we believe that it is most often used in a geovisualization context, and we have moved it to Chapter 3 from its original home with materials on point pattern analysis in Chapter 5. We recognize that these changes make it necessary for the reader from time to time to refer back to previous materials, and we have attempted to signal when this is wise by use of boxed thought exercises.

These additions have been balanced by the removal of some materials. First, for entirely pragmatic reasons, we have removed almost all the text on the analysis of line objects. Although it dealt with some of the basic ideas, neither of us was happy with the original chapter, which for reasons of length did not, and could not, reflect the increasing importance of network analysis in almost every branch of science. As readers of that chapter would have recognized, when dealing with linear objects we struggled to maintain our basic stochastic process approach. Somebody, somewhere, someday will write what is necessary—a major text book on geographic information analysis in a network representation of geography—but the task is well beyond what can be covered in a single chapter of the present book. A chapter on multivariate statistics, which sat a little uncomfortably in the first edition on the pretext of treating n-dimensional data as spatial, has been omitted. We have retained some of that material in the new chapter on geovisualization under the heading of "spatialization." In addition to these larger-scale adjustments, we have removed the extended treatment of the joins count approach to characterizing spatial autocorrelation, which, although pedagogically useful, seemed increasingly irrelevant to contemporary practice. Finally, in the interests of keeping the size of the book manageable, we have dropped an appendix introducing basic statistical concepts, assuming that readers can work from one of the many fine introductory text books available.

Since the publication of the first edition, much has changed and the general field has grown enormously, with developments in computing, statistics, and geographic information science. In updating the materials, we have tried as best we can to reflect this new work and the increasingly "location-aware" scientific and social environment in which it is placed, but we are aware of numerous things that we have omitted. If you look for something and are disappointed, we can only apologize.

SOFTWARE

One major change that we have tried to reflect is the increasing gap between methods used by academic spatial analysts and the functionality embedded in most commercial GIS. It is true that, if you know what you are doing and don't always rely on default settings, many of the methods we describe can be used within such a system, but such use is not ideal. Over the past decade, it has become increasingly obvious that most of today's leading researchers have developed their work in the public domain R programming environment (see Ihaka and Gentleman, 1996). Readers looking to implement the methods we describe should note that almost all of them, and many more, have been implemented in this environment (Baddeley and Turner, 2005; Bivand et al., 2008). Readers wishing to develop new and innovative approaches to geographic information analysis would be well advised to join this community of scholars.

David O'Sullivan
University of Auckland
Te Whare Wānanga o Tāmaki Makaurau

David Unwin
Birbeck, University of London

Matariki 2009

REFERENCES

Baddeley, A. and Turner, R. (2005) Spatstat: an *R* package for analyzing spatial point patterns. *Journal of Statistical Software*, 12: 1–42.

Bivand, R. S., Pebesma, E. J., and Gomez-Rubio, V. (2008) *Applied Spatial Data Analysis with R* (New York: Springer).

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5: 299–314.

O'Sullivan, D. and Unwin, D. J. (2003) *Geographic Information Analysis* (Hoboken, NJ: Wiley).

Unwin, D. J. (1981) *Introductory Spatial Analysis* (London: Methuen).

Unwin, D. J. (2005) Fiddling on a different planet. *Geoforum*, 36: 681–684.

Acknowledgments

DAVID O'SULLIVAN

Students and colleagues in courses at Penn State (GEOG 454, 455, 586) and Auckland (GEOG 318, 771) have helped to refine materials in this edition in many ways, large and small. I am grateful for the supportive input (in a variety of forms) of David Dibiase, Mark Gahegan, Frank Hardisty, Jim Harper, and George Perry. I also acknowledge a University of Auckland Study Leave award and the Spatial Literacy in Teaching (SpLinT) consortium (Universities of Leicester and Nottingham, and University College London), particularly Paul Longley and Nick Tate, for assisting with a Fellowship award in 2008. Study Leave and the SpLinT Fellowship enabled Dave and I to work together on the book for a pleasant and productive couple of days and helped to shape this edition.

I also thank Gill for her constant forbearance and support. Finally, the Saturday morning exploits of Fintan and Malachy, on cricket and football pitches, not to mention farflung virtual LEGO^{TM} galaxies, have reminded me throughout of where the real center of things lies.

DAVE UNWIN

The Washington "Space cadets" fired my graduate student enthusiasm for spatial analysis, and it has been a pleasure and a privilege eventually to meet several of them. The Departments of Geography at Aberystwyth, Leicester, and Birkbeck London have provided supportive bases and knowledgeable colleagues. Particular thanks for numerous discussions go to my former colleagues at the Leicester Midlands Regional Research Laboratory: Alan Strachan, Mike Worboys, David Maguire, Jo Wood, Jason Dykes, and, more recently, Pete Fisher. All teachers should learn from their students, and it is a pleasure to acknowledge the contributions to my thought process made, sometimes unwittingly, sometimes directly, by graduate classes in the Universities of Leicester, Birkbeck/University College London, Waikato (New Zealand), Canterbury (New Zealand), and Redlands (the United

States). A final and particularly formative experience has been that of teaching and tutoring classes of "accidental geographers" over the Internet, which has confirmed my convictions about the value of using geography in science. Like my colleague, I am grateful to the SpLinT Initiative for its 2008 support.

The errors that remain, are, of course, our own.

David O'Sullivan
Tāmaki-makau-rau, Aotearoa

Dave Unwin
Maidwell, England

Preface to the First Edition

Like Topsy, this book "jes growed" out of a little book one of us wrote in the period from 1979 to 1981 (*Introductory Spatial Analysis*, London: Methuen). Although that was fully a decade after the appearance of the first commercial geographical information systems (GIS) and more or less coincided with the advent of the first microcomputers, that book's heritage was deep in the quantitative geography of the 1960s, and the methods discussed used nothing more sophisticated than a hand calculator. Attempts to produce a second edition from 1983 onward were waylaid by other projects—almost invariably projects related to the contemporary rapid developments in GISs. At the same time, computers became available to almost everyone in the developed world, and in research and commerce many people discovered the potential of the geography they could do with GIS software. By the late 1990s, it was apparent that only a completely new text would do, and it was at this point that the two of us embarked on the joint project that resulted in the present book.

The materials we have included have evolved over a long period of time and have been tried and tested in senior undergraduate and postgraduate courses we have taught at universities in Leicester, London (Birkbeck and University Colleges), Pennsylvania (Penn State), Waikato, Canterbury (New Zealand), and elsewhere. We are passionate about the usefulness of the concepts and techniques we present in almost any work with geographic data and we can only hope that we have managed to communicate this to our readers. We also hope that reservations expressed throughout concerning the overzealous or simpleminded application of these ideas do not undermine our essential enthusiasm. Many of our reservations arise from a single source, namely the limitations of digital representations of external reality possible with current (and perhaps future?) technology. We feel that if GIS is to be used effectively as a tool supportive of numerous approaches to geography and is not to be presented as a one-size-fits-all "answer" to every geographical question, it is appropriate to reveal such uncertainty, even to newcomers to the field.

Although it was not planned this way, on reflection we progress from carefully developed basics spelled out in full and very much grounded in the intellectual tradition of *Introductory Spatial Analysis*, to more discursive accounts of more recent computationally intensive procedures. The early material emphasizes the importance of fundamental concepts and problems common to any attempt to apply statistical methods to spatial data and should provide a firm grounding for further study of the more advanced approaches discussed in more general terms in later chapters. The vintage of some of the references we provide is indicative of the fact that at least some of the intellectual roots of what is now called *geographical information science* are firmly embedded in the geography of the 1960s and range far and wide across the concerns of a variety of disciplines. Recent years have seen massive technical innovations in the analysis of geographical data, and we hope that we have been able in the text and in the suggested reading to capture some of the excitement this creates.

Two issues that we have struggled with throughout are the use of mathematics and notation. These are linked, and care is required with both. Mathematically, we have tried to be as rigorous as possible, consistent with our intended audience. Experience suggests that students who find their way to GIS analysis and wish to explore some aspects in more depth come from a wide range of backgrounds with an extraordinary variety of prior experience of mathematics. As a result, our "rigor" is often a matter merely of adopting formal notations. With the exception of a single "it can be shown" in Chapter 9, we have managed to avoid use of the calculus, but matrix and vector notation is used throughout and beyond Chapter 5 is more or less essential to a complete understanding of everything that is going on. Appendix B provides a guide to matrices and vectors that should be sufficient for most readers. If this book is used as a course text, we strongly recommend that instructors take time to cover the contents of this appendix at appropriate points prior to the introduction of the relevant materials in the main text. We assume that readers have a basic grounding in statistics, but to be on the safe side we have included a similar appendix outlining the major statistical ideas on which we draw, and similar comments apply.

Poor notation has a tremendous potential to confuse, and spatial analysis is a field blessed (perhaps cursed) by an array of variables. Absolute consistency is hard to maintain and is probably an overrated virtue in any case. We have tried hard to be as consistent and explicit as possible throughout. Perhaps the most jarring moment in this respect is the introduction in Chapters 8 and 9 of a third locational coordinate, denoted by the letter z. This leads to some awkwardness and a slight notational shift when we deal with regression on spatial coordinates in Section 9.3. On balance we prefer to use (x,y,z) and put up with accusations of inconsistency than to have too many pages bristling with subscripts (a flip through the pages should reassure the less easily intimidated that many subscripts remain). This pragmatic approach should serve as a reminder that, like its predecessor, this book is about the practical analysis of geographic information rather than being a treatise on spatial statistics. First and foremost, this is a geography book!

No book of this length covering so much ground could ever be the unaided work of just two people. Over many years one of us has benefited from contacts with colleagues in education and the geographic information industry far too numerous to mention specifically. To all he is grateful for advice, for discussion, and for good-natured argument. It is a testament to the open and constructive atmosphere in this rapidly developing field that the younger half of this partnership has already benefited from numerous similar contacts, which are also difficult to enumerate individually. Suffice it to say that supportive environments in University College London's Centre for Advanced Spatial Analysis and in the Penn State Geography Department have helped enormously. As usual, the mistakes that remain are our own.

David O'Sullivan
The Pennsylvania State University
(St Kieran's Day, 2002)

Dave Unwin
London, England
(St Valentines' Day, 2002)

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 1

Geographic Information Analysis and Spatial Data

CHAPTER OBJECTIVES

In this first chapter, we:

- Define *geographic information analysis* as it is meant in this book
- Distinguish geographic information analysis from *GIS-based spatial data manipulation* while relating the two
- Review the *entity-attribute model* of spatial data as consisting of *points*, *lines*, *areas,* and *fields*, with associated *nominal*, *ordinal*, *interval,* or *ratio* data
- Note some of the complications in this view, especially *multiple representation* at different scales, *time*, objects with *uncertain boundaries*, objects that are *fuzzy*, and objects that may be *fractal*
- Review *spatial data manipulation operations* and emphasize their importance
- Examine the various *transformations* between representations, noting their utility for geographic information analysis

After reading this chapter, you should be able to:

- List four different approaches to spatial analysis and differentiate between them
- Give reasons why modern methods of spatial analysis are not well represented in the tool kits provided by the typical GIS
- Distinguish between spatial objects and spatial fields and discuss why the vector-versus-raster debate in GIS is really about how we choose to represent these entity types

- Differentiate between point, line, and area objects and give examples of each
- List the fundamental data properties that characterize a field
- Provide examples of real-world entities that do not fit easily into this scheme
- Maintain a clear distinction between a real-world entity, its representation in a digital database, and its display on a map
- Differentiate between nominal, ordinal, interval, and ratio attribute data and give examples of each
- Give examples of at least 12 resulting types of spatial data
- List some of the basic geometrical data manipulations available in the typical GIS
- Outline methods by which the representations of entities can be transformed and explain why this is useful for geographic information analysis

1.1. INTRODUCTION

Geographic information analysis is not an established discipline. In fact, it is a rather new concept. To define what we mean by this term, it is necessary first to define a much older term—*spatial analysis*—and then to describe how we see the relationship between the two. Of course, a succinct definition of spatial analysis is not straightforward either. The term comes up in various contexts. At least four broad areas are identifiable in the literature, each using the term in different ways:

1. *Spatial data manipulation,* usually in a geographic information system (GIS), is often referred to as *spatial analysis*, particularly in GIS companies' promotional material. Your GIS manuals will give you a good sense of the scope of these techniques, as will the texts by Tomlin (1990) and Mitchell (1999).
2. *Spatial data analysis* is descriptive and exploratory. These are important first steps in all spatial analysis, and often are all that can be done with very large and complex data sets. Books by geographers such as Unwin (1982), Bailey and Gatrell (1995), and Fotheringham et al. (1999) are very much in this tradition.
3. *Spatial statistical analysis* employs statistical methods to interrogate spatial data to determine whether or not the data can be represented by a statistical model. The geography texts cited above touch on theses issues, and there are a small number of texts by statisticians interested in the analysis of spatial data, notably those by Ripley (1981, 1988), Diggle (1983), and Cressie (1991).

4. *Spatial modeling* involves constructing models to predict spatial outcomes. In human geography, models are used to predict flows of people and goods between places or to optimize the location of facilities (Wilson, 2000), whereas in environmental science, models may attempt to simulate the dynamics of natural processes (Ford, 1999). Modeling techniques are a natural extension of spatial analysis but are beyond the scope of this book.

In practice, it is often difficult to distinguish between these approaches, and most serious research will involve all four. First, data are collected, visualized, and described. Then exploratory techniques might raise questions and suggest theories about the phenomena of interest. These theories are then subjected to statistical testing using spatial statistical techniques. Theories of what is going on might then be the basis for computer models of the phenomena, and their results, in turn, may be subjected to more statistical investigation and analysis.

It is impossible to consider geographic information without considering the technology that is increasingly its home: geographical information systems (GISs). Although GISs are not ubiquitous in the way that (say) word processors are, they have infiltrated more and more businesses, government agencies, and other decision-making organizations. Even if this is the first time you've read a geography textbook, chances are that you will have already used a GIS without knowing it, perhaps when you used a website to generate a map of a holiday destination or to find driving directions to get you there.

In the above list, current GISs typically include item 1 as standard (since a GIS without these functions would be just a plain old IS!) and have some simple data analysis capabilities, especially exploratory analysis using maps (item 2). GISs have recently begun to incorporate some of the statistical methods of item 3 and only rarely include the capability to build spatial models and determine their likely outcomes (item 4). In fact, it can be hard to extend GIS to perform such analysis, which is why many geographic information analysts use other software environments for work that would be classified as belonging to items 3 and 4. In this book, we focus mostly on items 2 and 3. In practice, you will find that, in spite of rapid advances in the available tools, statistical testing of spatial data remains relatively rare. Statistical methods are well worked out and understood for some types of data but less so for many others. As this book unfolds, you should begin to understand why this is so.

If spatial analysis is so necessary— even worth writing a book about—then why isn't it a standard part of the GIS toolkit? We suggest a number of reasons, among them the following:

- *The GIS view of spatial data and that of spatial analysis are different*. The spatial analysis view of spatial data is more concerned with *processes* and *patterns* than it is with database management and manipulation, whereas the basic requirement for a *spatial database* is far more important to most large GIS buyers (government agencies, utilities) than the ability to perform complex and (sometimes) obscure spatial analysis.
- *Spatial analysis is not widely understood*. Spatial analysis is not obvious or especially easy, although we aim to address that issue in this book. The apparent difficulty means that it is difficult to convince software vendors to include spatial analysis tools as standard products. Spatial analysis tools are a possible addition to GIS that is frequently left out. This rationale has become less significant in recent years as software engineering methods enable GIS vendors to supply "extensions" that can be sold separately to those users who want them. At the same time, third-party vendors can supply add-on components more easily than previously, and open source software has become an increasingly important alternative in some quarters.
- *The spatial analysis perspective can sometimes obscure the advantages of GIS*. By applying spatial analysis techniques, we often raise awkward questions: "It looks like there's a pattern, but is it significant? Maybe not." This is a hard capability to sell!

Despite this focus, don't underestimate the importance of the *spatial data manipulation* functions provided by GIS such as buffering, point-in-polygon queries, and so on. These are essential precursors to generating questions and formulating hypotheses. To reinforce their importance, we review these topics in Section 1.5 and consider how they might benefit from a more statistical approach. More generally, the way spatial data are stored—or *how geographical phenomena are represented in GIS*—is becoming increasingly important for analysis. We therefore spend some time on this issue in Sections 1.2 and 1.3.

For all of these reasons, we use the broader term *geographic information analysis* for the material we cover. A working definition of this term is that it is concerned with investigating the *patterns* that arise as a result of *processe*s that may be operating in space. Techniques and methods to enable the representation, description, measurement, comparison, and generation of spatial patterns are central to the study of geographic information analysis. Of course, at this point our definition isn't very useful, since it raises the question of what we mean by *pattern* and *process*. For now, we will accept whatever intuitive notion you have about the meaning of the key terms. As we work through the concepts of point pattern analysis in Chapters 4 and 5, it will become clearer what is meant by both terms. For now, we will concentrate on the general spatial data types you can expect to encounter.

1.2. SPATIAL DATA TYPES

Thought Exercise: Representation

Throughout this book, you will find thought exercises to help you follow the text in a more hands-on way. Usually, we ask you to do something and use the results to draw some conclusions. You should find that these exercises help you remember what we've said. This first exercise is concerned with how we represent geography in a digital computer:

1. Assume that you are working for a road maintenance agency. Your responsibilities extend to the roads over a county-sized area. Your GIS is required to support operations such as surface renewal, avoiding clashes with other agencies—utility companies, for example—that also dig holes in the roads and make improvements to the road structure.

   Think about and write down how you would record the geometry of the network of roads in your database. What road attributes would you collect?
2. Imagine that you are working for a bus company in the same area. Now the GIS must support operations such as time-tabling, predicting the demand for existing and potential new bus routes, and optimizing where stops are placed.

   How would the recording of the geometry of the road network and its attributes differ from your suggestions in step 1 above?

What simple conclusion can we draw from this? It should be clear that how we represent the same geographic entities differs according to the purpose of the representation. This is obvious, but it can easily be forgotten.

Quite apart from the technical issues involved, social critiques of geographic information analysis often hinge on the fact that analysis frequently confines itself to those aspects of the world that can be easily represented digitally (see Fisher and Unwin, 2005).

When you think of the world in map form, how do you view it? In the early GIS literature, a distinction was often made between two kinds of system characterized by how the geography is represented digitally:

1. One type of system provides a *vector* view, which records locational (x, y) coordinates of the features that make up a map. In the vector view, we list features and represent each as a point, line, or area *object*. Vector GIS originated in the use of computers to draw maps based on digital data and were particularly valued when computer memory was an expensive commodity. Although the fit is inexact, the vector model is closest to an *object* view of the world, where space is thought of as an empty container occupied by different sorts of objects.
2. Contrasted with vector systems are *raster* systems. Instead of starting with objects on the ground, a grid of small units, called *pixels*, of the Earth's surface is defined. For each pixel, the value, or presence or absence of something of interest, is then recorded. Thus, we divide a map into a set of identical, discrete elements and list the contents of each. Because every location in space has a value (even if it is zero or null), a raster approach generally uses more computer memory than a vector one. Raster GIS originated mostly in image processing, where data from remote sensing platforms are often encountered.

In this section, we hope to convince you that at a higher level of abstraction the vector/raster distinction isn't very useful, and that it obscures a more important division between what we call an *object* and a *field* view of the world.

The Object View

In the object view, we consider the world as a series of *entities* located in space. Entities are (usually) real: you can touch them, stand in them, perhaps even move them around. An *object* is a digital representation of all or part of an entity. Objects may be classified into different object types—for example, *point objects*, *line objects,* and *area objects*—and in specific applications, these types are *instantiated* by specific objects. For example, in an environmental GIS, woods and fields might be instances of area objects. In the object view of the world, places can be occupied by any number of objects. A house can exist in a census tract, which may also contain lampposts, bus stops, road segments, parks, and so on.

Because it is also possible to associate *behavior* with objects, the object view has advantages when well-defined objects change over time—for example, the changing data for a census area object over a series of population censuses. Note that we have said nothing about *object orientation* in the computer science sense. Worboys et al. (1990) give a straightforward description of this concept as it relates to spatial data.

The Field View

In the *field* view, the world consists of properties continuously varying across space. An example is the surface of the Earth itself, where the field variable is elevation above sea level. Similarly, we can code the ground in a grid cell as either having a house on it or not. The result is also a field, in this case of binary numbers where 1 = "house" and 0 = "no house". If a single house is large enough or if its outline crosses a grid cell boundary, it may be recorded as being in more than one grid cell. The key ideas here are spatial *continuity* and *self-definition*. In a field, every location has a value (including "not here" or zero) and sets of values taken together define the field. This is in contrast with the object view, in which it is necessary to attach further attributes to represent an object fully—a rectangle is just a rectangle until we attach descriptive attributes to it.

The raster data model is one way to record a field. In this model, the geographic variation of the field is represented by identical, regularly shaped pixels. Earth's surface is often recorded as a regular grid of height values (a *digital elevation matrix*). An alternative is to use area objects in the form of a mesh of nonoverlapping triangles (called a *triangulated irregular network* or TIN) to represent the same field variable. In a TIN, each triangle vertex is assigned the value of the field at that location. In the early days of GIS, especially in cartographic applications, values of the field given by land height were often recorded using digital representations of the contours familiar from topographic maps. This is a representation of a field using overlapping area objects, the areas being the parts of the landscape enclosed within each contour. The important point is that a field can be coded digitally using either a raster or a vector approach.

Finally, another type of field is one made up of a continuous cover of assignments for a *categorical variable*. Every location has a value, but the "values" are the names given to phenomena. Consider a map of soil type. Every location has a soil, so we have spatial continuity, and we also have self-definition by the soil type involved, so this is a field view. Other examples might be land use maps, even a simple map of areas suitable or unsuitable for some development. In the literature, these types of field variable have gone under a number of different names, among them *k-color maps* and *binary maps*. A term that is gaining ground is *categorical coverage*, indicating a field made up of a categorical variable. The important point is that such categorical coverage can be coded digitally using either a vector or a raster approach.

Choosing the Representation to Be Used

In practice, it is useful to think about the elements of reality modeled in a GIS database as having two types of existence and to keep *both* distinct from the way the same entities might be displayed on a map. First, there is the element in reality, which we call the *entity*. Second, there is the element as it is *represented* in the database. In database theory, this is called the *object* (confusingly, this means that a field is a type of object). Clearly, what we see as entities in the real world depends on the application, but to make much sense, an entity must be

- *Identifiable*. If you can't see it, then you can't record it.
- *Relevant*. It must be of interest.
- *Describable*. It must have attributes or characteristics that we can record.

Formally, an *entity* is defined as a phenomenon of interest in reality that is not further subdivided into phenomena of *the same kind*. For example, a road *network* could be considered an entity and subdivided into component parts called *roads*. These might be further subdivided, but these parts would not be called roads. Instead, they might be considered *road segments* or something similar. Similarly, a *forest* entity could be subdivided into smaller areas called *stands*, which are in turn made up of individual *trees*.

The relationship between the object and field representations is a deep one, which, it can be argued, goes back to philosophical debates in ancient Greece about the nature of reality: a continuously varying field of phenomena or an empty container full of distinct objects? You should now be able to see that the key question, from the present perspective, is not which picture of reality is *correct* but which we choose to adopt for the task at hand. A GIS-equipped corporation concerned with the management of facilities such as individual buildings, roads, or other infrastructure would almost certainly consider an object view most appropriate. In contrast, developers of a system for the analysis of hazards in the environment may adopt a field view. Most theory in environmental science tends to take this approach, using, for example, fields of temperature, wind speed, height, and so on. Similarly, data from remote sensing platforms are collected as continuous rasters, so a field view is the more obvious approach.

It is also a good idea to make a clear distinction not only between the entity and its representation in a GIS database, but also between both of these and the way the same entity is displayed on a map. Representing the content of a map is not the same as representing the world. The objectives of map design are visual—to show map users something about the real world—whereas the objectives of a database are concerned with management, measurement, analysis, and modeling. It pays to keep these objectives distinct when choosing how to represent the world in digital form.

Types of Spatial Object

The digital representation of different entities requires the selection of appropriate spatial object types, and there have been a number of attempts to define general spatial object types. A common approach—reinvented many times—is based on the spatial *dimensionality* of the object concerned. Think about how many *types* of object you can draw. You can mark a *point*, an object with no length, which may be considered to have a spatial dimension or length, L, raised to the power zero, hence L^0. You can draw a line, an object having the same spatial dimension as any simple length, that is, L^1. You can also shade an area, which is an object with spatial dimension length squared, or L^2. Finally, you can use standard cartographic or artistic conventions to represent a volume, which has spatial dimension length cubed, or L^3. The U. S. *National Standard for Digital Cartographic Databases* (DCDSTF, 1988) and Worboy's generic model for planar spatial objects (Worboys, 1992, 1995) both define a comprehensive typology of spatial objects in terms similar to these.

An Exercise: Objects and Fields Decoded

1. Obtain a topographic map at a scale of 1:50,000 or larger of your home area. Study the map, and for at least 10 of the types of entity the map represents—remember that the map is already a representation—list whether they would best be coded as an object or a field. If the entity is to be represented as an object, state whether it is a point, line, or area.
2. If you were asked to produce an initial specification for a data model that would enable a mapping agency to ''play back'' this map from a digital version held in a database, how many specific instances of objects (of all kinds) and fields would you need to record?

Hint: Use the map key. There is, of course, no single correct answer to this question.

1.3. SOME COMPLICATIONS

The view of the world we have presented so far is deceptively simple, and deliberately so. There are a number of complications, which we now examine. In each case, our perspective is that of a geographic information analyst, and the key question to be asked is the extent to which the complication impacts on any analytical results obtained.

Objects Are Not Always What They Appear to Be

Students often confuse the various cartographic conventional representations with the fundamental nature of objects and fields. For example, on a map, a cartographic line may be used to mark the edge of an area, but the entity is still an area object. Real line objects represent linear entities such as railways, roads, and rivers. On topographic maps, it is common to represent the continuous field variable of height above sea level using the lines we call *contours*; yet, as we have discussed, fields can be represented on maps in many different ways.

Objects Are Usually Multidimensional

Very frequently, spatial objects have more than the single dimension of variability that defines them. We might, for example, locate a point object by its (x, y) coordinates in two spatial dimensions, but in many applications it would be much better to record it in three spatial dimensions (x, y, z), with depth or height as a third dimension. A volume of rock studied by a geologist exists at some depth at a location but also has attributes such as it porosity or color; the interest will be in how this attribute varies in (X, Y, Z) space. Many GISs do not cope easily with such data, so frequently it is necessary to record the additional coordinate as another attribute of an object fixed at a location in (x, y). This can make perfectly natural queries and analyses that require the full three spatial dimensions awkward or even impossible. Raper (2000) provides numerous illustrations of the multidimensional nature of geographic objects.

Objects Don't Move or Change

The view of the world presented so far is a static one, with no concept of time except possibly as an attribute of objects. This is fine for some problems, but in many applications our major interest is in how things change over time. Standard GISs do not easily handle location in time as well as location in space. A moment's thought will reveal the problems that incorporating an object's location in time might generate. The idea of change over time, what we call *process*, is of particular importance to most sciences, yet handling it in a digital environment that does not readily incorporate it in any object's description will always be difficult. The problem has been discussed for many years (see Langran, 1992, and O'Sullivan, 2005, for a review of progress), but as far as we are aware, no commercial *temporal GIS* has yet been produced. In research there have been many attempts, such as *PC-Raster*™ (see Wesseling et al., 1996), to create one, almost all of which involve the definition of some generic language for describing change in both space and time.

Objects Don't Have Simple Geometries

Some aspects of geographic reality that we might want to capture are not well represented in either the raster/vector or object/field views. The obvious example here is a transport or river *network*. Often, a network is modeled as a set of line objects (routes) and a set of point objects (transport nodes), but this data structure is very awkward in many applications, and it is difficult to get the representation just right (think of one-way streets, restricted turns, lanes, and so on). Another example is becoming increasingly important: *image* data. An image in a GIS might be a scanned map used as a backdrop or it might be a photograph encoded in a standard format. At the nuts and bolts level, images are coded using a raster approach, but the key to understanding them is that, other than being able to locate a cursor on an image, the values of the attributes themselves are not readily extracted, nor, for that matter, are the individual pixel values important: it is the image *as a whole* that matters. In the most recent revision of the *ArcGIS,* some of these complexities are recognized by having *five* representations of geography, called *locations*, *features* (made up of locations), *surfaces* (fields), *images,* and *networks* (see Zeiler, 1999). As geographic information analysis becomes increasingly sophisticated and is extended to embrace applications that hadn't even been considered when the basic framework we adopt was developed, this issue is likely to be of greater importance.

Objects Depend on the Scale of Analysis

Different object *types* may represent the same real-world phenomenon at different scales. For example, on his daily journey to work, one of us used to arrive in London by rail at an entity called Euston Station. At one scale this is best represented by a dot on a map, which in turn is an instance of a point object that can be represented digitally by its (x, y) co-ordinates. Zoom in a little, and Euston Station becomes an area object, best represented digitally as a closed string of (x, y) coordinates defining a polygon. Zooming in closer still, we see a network of railway lines (a set of line objects) together with some buildings (area objects), all of which would be represented by an even more complex data description. Clearly, the same entity may be represented in several ways. This is an example of the *multiple representation* problem in geographic information analysis. Its main consequence is to make it imperative that in designing a geographic information database and populating it with objects of interest, it is vital that the type of representation chosen will allow the intended analyses to be carried out. As the next exercise illustrates, this is also true for any maps produced from the same database.

Scale and Object Type

We can illustrate this idea using a convenient example from Great Britain accessed via your Web browser. The same exercise can easily be done using paper maps or Web-delivered map extracts from another national mapping agency.

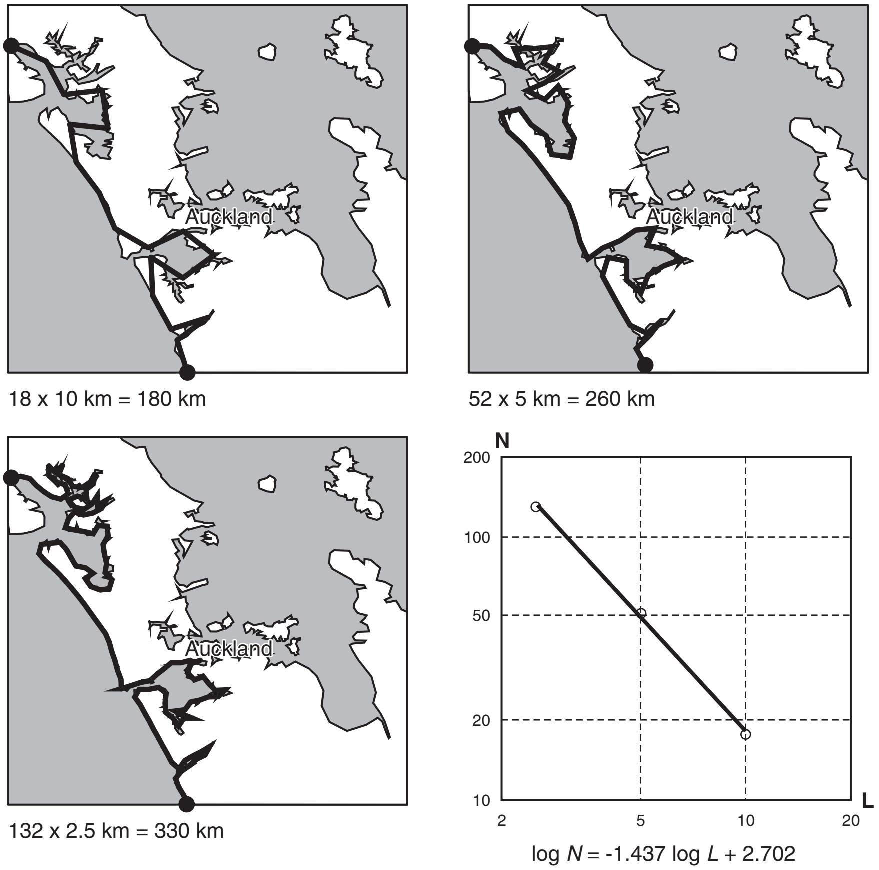
1. Go to www.ordnancesurvey.co.uk. This will bring you to a screen with an option to ''Get-a-map.'' At the window, enter ''Maidwell'' (without quotation marks) which is the name of a small village in the English Midlands and hit GO.
2. You arrive at a screen with an extract from the 1:50,000 topographic map of the area around this village.
3. To the left of the map are some balloons labeled ''+'' and ''−''. If you run the mouse over them, you will see that each balloon corresponds to a map of the area at scales of 1:25,000 (zoom level 5), 1:50,000 (zoom level 4), and 1:250,000 (zoom level 3 in two versions, a ''Miniscale'' and a ''Simplified Miniscale'').
4. The exercise is simple. Make a table in which columns represent each of the five mapping scales and the rows are entities of interest—we suggest ''roads'', ''houses'', ''public house'' (there is a rather good one), ''rivers,'' and ''land-height''; enter a code into each cell of this table to indicate how the feature is represented. It will help if you use codes such as P (point feature), L (line feature), A (area object), F (field), and X for features that are absent at that scale.
5. What does this tell you about multiple representation?

CEOs of national mapping agencies know to their cost that it is almost impossible to produce maps from a single digital database at scales other than that for which the original database design was intended.

Objects Might Have Fractal Dimension

A further complication is that some entities are *fractals*, having the same or similar level of detail no matter how closely we zoom in. Fractals are difficult to represent digitally unless we accept that the representation is only a snapshot at a particular resolution. The classic example is a linear feature such as a coastline whose "crinkliness" remains the same no matter how closely we examine it. No matter how accurately we record the spatial coordinates, it is impossible to capture all the detail. A rather unexpected consequence is that when dealing with irregular lines, their length appears to increase the more "accurately" we measure it!

Imagine measuring the coastline of (say) a part of the North Island of New Zealand using a pair of dividers set to 10 km, as in the top left panel of Figure 1.1. We will call the dividers' separation the *yardstick*. With a yardstick



18 x 10 km = 180 km

52 x 5 km = 260 km

132 x 2.5 km = 330 km

log N = -1.437 log L + 2.702

Figure 1.1 Determining the fractal dimension of part of the New Zealand coastline.

of 10 km, we count 18 segments and get a coastline length of 18 × 10 km = 180 km. Repeating the process with a 5 km yardstick, as in the top right panel, we count 52 segments and get a length of 52 × 5 = 260 km. Finally, using a yardstick of 2.5 km, we obtain a total length of 132 × 2.5 = 330 km. The coastline appears longer the closer we look. What would happen if we had a 1-km or a 100-m yardstick? What about 100 mm? What is the "real" length of this coastline?

Fractal dimension is a mathematical idea for dealing with this difficulty. Although it was popularized by Mandelbrot (1977), the idea that the length of lines varies with the scale of their representation was spotted long before the fractal concept was widespread in mathematics (Richardson, 1961). *Fractal* is a compression of *fraction* and *dimensional* and expresses the idea that a line may be somewhere *between* one- and two- dimensional, with a fractal dimension of, say, 1.2 or 1.5. One understanding of an object's dimension is that it expresses how its apparent size (in this case length), measured by counting smaller elements (in this case yardsticks), changes as the linear size (or *resolution*) of the smaller elements changes. A simple nonfractal one- or two- dimensional entity's size, as measured by counting subelement yardsticks, increases with the power of its dimension, so that the number of shorter segments in a simple straight line doubles if we halve the yardstick size. The count of small cube yardsticks in a large volume increases eightfold if we halve the dimension of the small cubes. If the linear dimensions of the yardsticks for two measurements are L_1 and L_2, and the respective counts of yardsticks are N_1 and N_2, then the fractal dimension D is given by

$$D = \frac{\log(N_1/N_2)}{\log(L_1/L_2)} \tag{1.1}$$

This definition does not restrict D to whole number values, and it can be used to estimate the fractal dimension of irregularly shaped entities. The lower right panel of Figure 1.1 shows each yardstick length and yardstick–count combination on a log-log *Richardson plot*. The three points lie roughly on a straight line, whose negative slope, fitted by simple linear regression, gives the fractal dimension. In the example, we arrive at an estimate for the fractal dimension of 1.44. In fact, we can properly make this measurement only on the coastline itself, because any stored representation, such as the map we started with, has a limiting resolution at which the length of the line will become fixed as we set our dividers smaller. However, the yardstick length–count relationship is often stable over several orders of magnitude, and we can estimate the fractal dimension of an entity from a large-scale object representation.

Some More Work: Do It for Yourself

As so often is the case, the best way to understand this is to do it yourself.

1. Find a reasonably detailed topographic map at a scale of about 1:25,000 or 1:50,000 and select a length of river as your object of study. Obviously, since we are interested in the sinuosity of linear objects, it makes sense to choose a river that shows meandering behavior. The equivalent of a 20-km length is about right and will not involve too much work.
2. Now set a pair of dividers at a large equivalent distance on the ground, say 1 km, and ''walk'' them along the river counting the number of steps. Record the yardstick length and the number of segments.
3. Repeat using a halved yardstick, equivalent to 500 m on the ground.
4. Repeat the process again and again until the yardstick becomes so short that the experiment is impractical. You should now have a table of values.
5. Convert both the number of steps and the yardstick length to their logarithms and plot the resulting numbers with log(number of steps) on the vertical axis and log(yardstick length) on the horizontal axis. If you have access to a spreadsheet, you should be able to do this easily. The points should fall roughly along a straight line, although success isn't guaranteed.
6. Finally, use the spreadsheet, (or a straight edge and a good eye) to fit a best-fit line to your data and estimate the fractal dimension of your river.

So, what does this all *mean*? The simplest interpretation of the fractal dimension of a line is as a measure of its "wiggliness" or "crinkliness." The fractal dimension of an entity expresses the extent to which it "fills" up the next dimension from its *topological dimension*. A line has a single topological dimension of length L^1. A line with fractal dimension 1.0 (note the decimal point) is an idealized line and takes up no space in two dimensions. However, a line with fractal dimension 1.1 or 1.15 begins to "fill up" the two dimensions of the plane in which it is drawn. Many linear features in geography have a fractal dimension somewhere between about 1.1 and 1.5.

Variants of the Richardson plot can be used to estimate the fractal dimension of area and volume objects (or surfaces) by counting the number of elements they contain at different linear resolutions. A considerable amount of work has been done on measuring the fractal dimension of the developed area of cities (Batty and Longley, 1994). While a perfectly smooth surface has fractal dimension 2.0, a rough one might have fractal dimension 2.3. Often, the fractal dimension of surface topography can be related to other characteristics. Other examples of fractal dimension in geographic phenomena are provided by soil pH profiles and river networks (see Burrough, 1981; Goodchild and Mark, 1987; Lam and De Cola, 1993; Turcotte, 1997).

The fractal concept is strongly related to the notion of scale. Some researchers have tried to make use of this in cartographic generalization, with mixed results. It should be clear that fractal dimension is indicative of how measures of an object will change with scale and generalization, and, in general, scale-varying properties of phenomena can be related to their fractal dimension (Goodchild, 1980). More recently, it has been suggested that measuring the fractal dimension of digitized data can help to determine the scale of the source map line work from which it is derived (Duckham and Drummond, 2000).

Objects Can Be Fuzzy and/or Have Indeterminate Boundaries

The preceding discussion has assumed that the objects we deal with are what is technically called *crisp*, and that if they have a spatial extent, their boundaries can in principle be recognized exactly. Many spatial entities that we might want to describe and analyze aren't crisp, and some may also have uncertain boundaries. The archetypal example is soil. On a map, the soil type will be represented by nonoverlapping area polygons, with hard and fast lines separating the various soil types that the surveyor recognizes. This is an example of a *k*-color map discussed in Section 3.7, but as any soil surveyor knows, it is really a fiction—for two possible reasons.

First, although the soil can change very abruptly, such that a line can be drawn on a map to separate different soil types, soil types may also grade almost imperceptibly from one to another, such that there is no certain boundary between them. Honest soil surveyors often recognize the uncertain nature of such a boundary by marking the transition with a dotted line. In a GIS and in subsequent spatial analysis, this uncertainty is often simply erased by assuming that such lines are in fact certain. The same issue arises, for example, in geology, where rock types can change imperceptibly; in marketing, where the boundaries of some trade area might be uncertain; and when describing mental maps. Every Londoner knows that a part of the city is called "Soho," but this has no legislative basis and most people would have difficulty deciding where it begins and ends. In the same city, "Westminster" has a legislative basis and so, in principle, has a certain boundary, but we doubt that many Londoners would know this and instead think of it in much the same uncertain way that they would think of Soho. To add to the complexity, and as with Soho and some soil types, some parts of an object's boundary might be uncertain but other parts of the same boundary might be certain. One possible way to handle this boundary uncertainty is to assign some probability of membership of the defined type to each location, so that instead of saying "Here we are on soil type such and such," our maps and data would say "Here there is a probability of (say) 0.7 that we are on soil type such and such."

Second, and again best illustrated using soils, objects might be *fuzzy*. In saying that a given soil belongs to a specific soil type, we are asserting that this type (or *set*) is itself *crisp*, by which we mean that we can unambiguously state whether or not the soil really is of that type. Yet, some sets might defy such assignment, so all we can say is that the given soil is more or less of a given sort, replacing our certainty with a value that expresses the extent to which it might belong to the given type. This isn't the same as the boundary uncertainty discussed above, where we are certain about the types but uncertain about the given soil. Here we are uncertain about the type of soil itself but, in a sense, certain about the soil belonging to that uncertain type. Practically, the extent to which any given entity is a member of such a fuzzy set is often recorded using a "membership" value ranging from 0 to 1. This can give rise to confusion with the probabilities associated with uncertain boundaries. Thinking through the exercise below might help you distinguish the two sorts of uncertainty.

Thought Exercise: Certainty and Uncertainty

Consider the following hierarchy of statements:

> John is over 1.8 m tall.

This is a certain statement about John being a member of the crisp set of all people over 1.8 m tall.

> I think John is over 1.8 m tall.

This set is still crisp, but we aren't sure about John's membership of it and might assign a probability to our uncertainty. This is analogous to the uncertain boundary issue.

(*continues*)

(*box continued*)

John is tall

''Tall'' is a fuzzy set, but we are certain that John belongs to it. It is the fuzzy category ''tall'' that now encapsulates the uncertainty. It is also possible that John is in the fuzzy set ''really tall'' or ''of average height'' (although probably not in all three of these fuzzy sets, since they now cover quite a range of circumstances). For any particular height, we might assign a membership value for these sets that records how ''tall,'' ''really tall,'' or ''of average height'' it is. If John's actual height is 1.9 m, then his membership in the set ''tall'' might be 1.0, ''really tall'' 0.6, and for ''of average height'' 0.05. This is the fuzziness issue.

I think John is really tall.

This combines the two types of uncertainty into one statement about John

Now think about how these same statements can be translated into spatial examples. What are the implications of both types of uncertainty for allegedly simple measures such as the total area of a given soil that could be extracted from a GIS in a matter of seconds?

1.4. SCALES FOR ATTRIBUTE DESCRIPTION

In addition to point, line, and area object types, we need a means of assigning attributes to spatially located objects. The range of possible attributes is huge, since the number of possible ways we can describe things is limited only by our imagination. For example, we might describe buildings by their height, color, age, use, rental value, number of windows, architectural style, ownership, and so on. Formally, an *attribute is any characteristic of an entity selected for representation*. In this section, we explore a simple way of classifying attributes into types based on their *level of measurement*. The level of measurement is often a constraint on the choice of method of analysis and, ultimately, on the inferences that can be drawn from a study of that attribute's spatial structure.

It is important to clarify what is meant by measurement. When information is collected, *measurement* is the process of assigning a class or value to an observed phenomenon according to some set rules. It is not always made clear that this definition does not restrict us to assignments involving numbers. The definition also includes the classification of phenomena into types or their ranking relative to one another on an assumed scale. You are reading a work that you assign to the general class of objects called books.

You could rank it relative to other books on some scale of merit as good, indifferent, or bad. It is apparent that this general view of measurement describes a process that goes on in our minds virtually all our waking lives as we sense, evaluate, and store information about our environment.

If this everyday process is to yield useful measurements, it is necessary to insist that measurements are made using a *definable process*, giving *reproducible* outcomes that are as *valid* as possible. The first requirement implies that the measurer knows what he or she is measuring and is able to perform the necessary operations; the second is that repetition of the process yields the same results and gives similar results when different data are used; the third implies that the measurements are true or accurate. If any of these requirements are not met, the resulting measurements will be of limited use in any GIS, or at any rate, we will need good information about the ways in which the measurements fail to comply with these requirements in order to make effective use of them. In short, we need to know what we are measuring, there must be a predefined scale on which we can place phenomena, and we must use a consistent set of rules to control this placement.

Sometimes what we need to measure to produce attribute data is obvious, but at other times, we are interested in analyzing concepts that are not readily measured and for which no agreed-upon measurement rules exist. This is most common when the concept of interest is itself vague or has a variety of possible interpretations. For example, it is easy to use a GIS to map the population density over a region, but because it involves people's reactions, standard of living, and available resources, the concept of *overpopulation* cannot be measured simply by the population density. Note that these ideas do not prevent us from creating measures based on opinions, perceptions, and so on, and therefore admit the development of GIS dealing with qualitative data, provided that attention is paid to the difficulties.

The rules defining the assignment of a name, rank, or number to phenomena determine what is called the *level of measurement*, different levels being associated with different rules. Stevens (1946) devised a useful classification of measurement levels that recognizes four levels: *nominal*, *ordinal*, *interval,* and *ratio.*

Nominal Measures

Because no assumptions are made about relative values being assigned to attributes, *nominal* measures are the lowest level in Stevens's scheme. Each value is a distinct *category*, serving only to label or name the phenomenon. We call certain buildings "shops," and there is no loss of information if these are called "category 2" instead. The only requirement is that categories are *inclusive* and *mutually exclusive*. By *inclusive*, we mean that it must be possible to assign all objects to some category or other ("shop" or "not a shop"). By *mutually exclusive*, we mean that no object should be capable of being placed in more than one class. No assumption of ordering or of distance between categories is made. In nominal data, any numbers used serve merely as symbols and cannot be manipulated mathematically in a meaningful way. This limits the operations that can be performed on them. Even so, we can count category members to form frequency distributions. If entities are spatially located, we may map them and perform operations on their (x, y) locational coordinates.

Ordinal Measures

For nominal measures, there are no implied relationships between classes other than their mutual exclusivity. If it is possible to rank classes consistently according to some criterion, then we have an *ordinal* level of measurement. An example is the classification of land into capability classes according to its agricultural potential. We know the order, but not the differences, along an assumed scale. Thus, the difference between the first and second classes may be very different from that between the ninth and tenth classes. Like nominal data, not all mathematical operations are clearly meaningful for ordinal data, but some statistical manipulations that do not assume regular differences are possible.

Attributes measured on the nominal and ordinal scales are often collectively referred to as *categorical data*.

Interval and Ratio Measures

In addition to ordering, the *interval* level of measurement has the property that differences or distances between categories are defined using fixed equal units. Thermometers typically measure on an interval scale, ensuring that the difference between, say, 25°C and 35°C is the same as that between 75.5°C and 85.5°C. However, interval scales lack an inherent zero and so can be used only to measure *differences*, not absolute or relative magnitudes. *Ratio* scales have an inherent zero. A distance of 0 m really does mean no distance, unlike the interval scale 0°C, which does not indicate no temperature. By the same token, 6 m is twice as far as 3 m, whereas 100°C is not twice as hot as 50°C.

The distinction is clarified by examining what happens if we calculate the ratio of two measurements. If place A is 10 km (6.2137 miles) from B and 20 km (12.4274 miles) from C, then the ratio of the distances is

$$\frac{\text{distance}\,AB}{\text{distance}\,AC}=\frac{10}{20}\equiv\frac{6.2137}{12.4274}\equiv\frac{1}{2} \tag{1.2}$$

whatever units of distance are used. Distance is fundamentally a ratio-scaled measurement. Interval scales do not preserve ratios in the same way. If place B has a mean annual temperature of 10°C (50°F) and place C is 20°C (68°F), we cannot claim that C is twice as hot as B because the ratio depends upon our units of measurement. In Celsius it is $20/10 = 2$, but in Fahrenheit it is $68/50 = 1.36$. In spite of this difference, interval and ratio data can usually be manipulated arithmetically and statistically in similar ways, so it is usual to treat them together. Together, they are called *numerical measures*.

Although data may have been collected at one measurement level, it is often possible and convenient to convert them into a *lower* level for mapping and analysis. Interval and ratio data can be converted into an ordinal scale, such as high/low or hot/tepid/cold. What is *generally* not possible is to collect data at one level and attempt to map and analyze them as if they were at a higher level, as, for example, by trying to add ordinal scores.

It is important to note that not everybody is convinced by Stevens's scheme for classifying levels of measurement. Velleman and Wilkinson (1993) have pointed out that it may be unnecessarily restrictive to rule out various types of analysis because the level of the attribute measurement seems not to support it (they also point out that this was not Stevens's intention). A good example is where a nominal attribute—say, a county ID number—seems to have some relationship with another variable of interest. Often in spatial numbering schemes there is a spatial pattern to the numbering—perhaps from east to west or north to south, or from an urban center outward. In such cases, relationships might very well be found between a theoretically nominal attribute (the ID number) and some other variable. Of course, in this case it would be important to determine what is responsible for the relationship and not simply to announce that zip codes are correlated with crime rates!

Later, Stevens himself added a *log interval* scale to cover measures such as earthquake intensity and pH in which the interval between measures rises according to a power rule. Later still, Chrisman (1998) pointed out that there are many types of attribute data in GISs that don't fit this scheme. For example, many types of line objects are best represented by both their magnitude and direction as *vector* quantities, and we often refer measures to cyclical scales such as angles that repeat every 360°. Such criticism of the measurement level approach emphasizes the important principle that it is always good to pursue investigations with an open mind. Nevertheless, the nominal, ordinal, interval, ratio scheme remains useful in considering the possibilities for analysis.

Dimensions and Units

Apart from their level of measurement, attributes have the property of *dimensionality* and are related to some underlying *scale of units*. If we describe a stream as a line object, variables we might consider important include its velocity, cross-sectional area, discharge, water temperature, and so on. These measurable variables are some of its so-called *dimensions* of variability. The choice of dimensions depends on the interests of the researcher, but in many problems in science it can often be reduced to combinations of the three fundamental dimensions of mass, length, and time, indicated by the letters M, L, and T. For example, a velocity dimension is distance L divided by time T, or L/T. This is true regardless of whether velocity is recorded in miles per hour or meters per second. LT^{-1} is another way of writing length divided by time.

Similarly, cross-sectional areas can be reduced to the product of two length dimensions, or L^2, discharge is a volume L^3 per unit of time T with dimensions L^3T^{-1}, and so on. Nondimensional variables are an important class whose values are independent of the units involved. For example, an angle measured in *radians* is the ratio of two lengths—arc length and circle radius—whose dimensions cancel out ($LL^{-1} = L^0$) to give no dimension. An important source of nondimensional values is observations recorded as proportions of some fixed total. For example, the proportion of the population that is white in some census district is a nondimensional ratio.

Dimensional analysis is an extremely valuable method in any applied work. Because equations must balance dimensionally as well as numerically, the method can be used to check for the existence of variables that have not been taken into account and even to help in suggesting the correct form of functional relationships. Surprisingly, geographers have shown little interest in dimensional analysis, perhaps because in a great deal of human geographic work no obvious fundamental dimensions have been recognized. Yet, as Haynes (1975, 1978) has shown, there is nothing to stop the use of standard dimensions such as P (= number of people) or $ (= money), and this usage may often suggest possible forms of equations.

Finally, interval and ratio attributes are related to a fixed scale of *units,* the standard scales used to give numerical values to each dimension. Throughout history, many systems of units have been used to describe the same dimensions. For example, in distance measurement, use has been made of "British" or Imperial units (inches, feet, miles), metric units (meters, kilometers), and other traditional systems (hands, rods, chains, nautical miles), giving a bewildering and confusing variety of fundamental and derived units. Although many systems were used because of their relevance to everyday life and are often convenient, in science they are unsatisfactory and can become confusing. This is something that NASA found out to enormous cost in 1998 when confusion over the system of units used to measure the gravitational acceleration of Mars spelled disaster for the Mars Climate Orbiter mission.

Thought Exercise: Spatial Data Types in Everyday Life

Look at Figure 1.2, which attempts to cross-tabulate measurement level with the geometric object types we have discussed to arrive at 12 possible spatial data types.

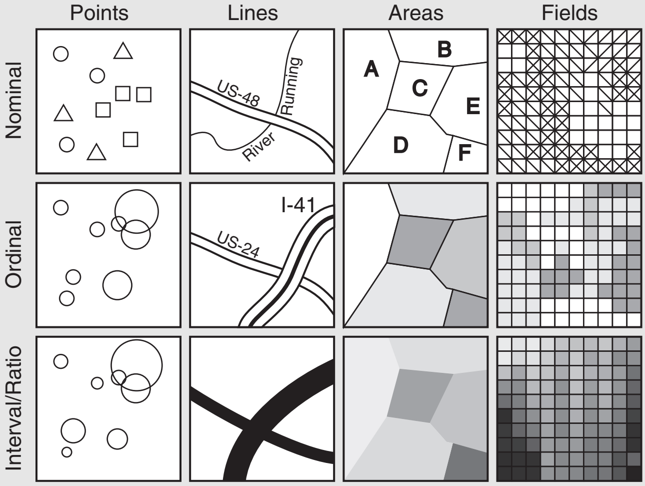


Figure 1.2 A schematic representation of entity-attribute spatial data types.

Now we want you to think about the rather abstract ideas we have been discussing.

What types of spatial object do you move among in your day-to-day life? For example:

- Is your house a point, an area, or both?
- Is your route to work, school, or college a line? What attributes might be used to describe it?
- Are you a nominal point data type? Perhaps you are a space–time (hence four-dimensional) line?

(*continues*)

(*box continued*)

- What measurement scales would be suitable for the attributes you would use to describe each of these (and any other) spatial objects you have suggested?

The answers to these questions give a sense of how potentially rich but also how reductive the entity-attribute framework is. As we explore spatial analysis further, remember this point: regardless of the insights that spatial analysis may yield, it is always performed on a representation of reality that may ultimately limit its usefulness.

1.5. GIS AND SPATIAL DATA MANIPULATION

We have noted that it is the ability to perform spatial manipulations on its data that distinguishes a GIS from any standard database management system. In this section, we examine a selection of these spatial data manipulations from our perspective as geographic information analysts. We do not intend to cover all these operations in detail since the number is very large and their implementation varies from system to system. In this section, we develop two spatial analytical perspectives on them. First, we develop the idea that these geometric operations involve some form of *transformation* between spatial data types. Second, we draw attention to the impact of *error* in our coding of the (x, y) coordinates used on the various outcomes.

Sometimes the geometry involved is simple—for example, finding the total length of line objects with a given characteristic (rivers, railways, roads needing repair) or calculating the total area and perimeter of some area objects (woodlands, crops of a certain type). At other times, it is *intersecting* types of spatial units in different ways that is the key. For example, we can easily use a GIS to determine how many cases of a disease occur within various distances of certain kinds of factories or other point objects. We need geo-coded data for cases of the disease and also for the facilities. These are usually available in the form of mailing addresses both for those afflicted by the disease and for the suspect facilities. We can then *buffer* the facilities to some distance (say 1 km) and use *point-in-polygon* operations to determine how many cases of the disease occur in the relevant buffer areas. The end result is a set of numbers recording how many cases of the disease occurred in the vicinity of each factory and how many occurred nowhere near a factory. Having determined these numbers, we could use appropriate statistical methods to determine whether or not the rates exhibit some nonrandom pattern.

Similarly, map *overlay* is where two or more map layers are combined in various ways to produce new combined layers. The classic example involves combining a number of development suitability classifications into a single composite index. This application was one of the original inspirations for GIS technology (see Ian McHarg's 1969 classic book *Design with Nature*). Input map data might include land slope, woodland density, transport accessibility (which might have been generated from buffer operations on the transport system), environmental sensitivity, and geological suitability for building. Map overlay produces a composite map formed from multiple intersections of all the inputs. Areas in the composite map have multiple attributes, derived from the attributes of their "parents," and can be assigned an overall suitability rating for development. The fundamental operation here is *geometric intersection* of the polygon areas in each map. A related operation *merges* polygons in different maps, depending on the similarity of their attributes. Incidentally, both of these operations are examples of the interchangeability of the raster and vector models, since either can readily be performed in a system based on either model. In fact, the two operations are developments—in geographic space—of the intersection and union operations familiar from set theory and Venn diagrams. Because it is so often a part of geographic information analysis, map overlay and the issues it raises are further discussed in Chapter 11.

Whether we are referring to point, line, area, or field entities, these operations (length, area, perimeter, intersection, buffer, merger, point-in-polygon, overlay, etc.) all involve relatively simple geometric manipulations of locational (x, y) coordinates. A useful way to think of them is as transformations between the various spatial data types that we recognized in Section 1.2. For example, if we have a data set made up of point objects, we might be interested in the areas within, say, 5 km of these objects defined by a series of circular buffers centered on each object. The buffered areas form a set of area objects, and we have transformed from points of length dimension L^0 to areas of length dimension L^2 (L^0 to L^2). In fact, such a buffer can also be considered to be a defined isoline on a continuous surface of distances from the points which is L^0 to L^3. Had the original buffer been along a line object, the transformation would have been from line to area (L^1 to L^2), and a buffer around an area object creates a second area object (L^2 to L^2). Reverse operations are also possible. We could start with area objects, and transform them into lines by computing their skeleton network (L^2 to L^1) or find their centroids, thus generating point objects (L^2 to L^0) (see Chapter 7 for more on these operations). Table 1.1 attempts to summarize this transformational view of GIS operations.

Rows of the table represent the data type from which we transform, and columns represent the resulting data type. Each row and column

Table 1.1 Spatial Geometric Operations as Transformations Between Data Types

		TO			
		Point, L^0	*Line, L^1*	*Area, L^2*	*Field, L^3*
FROM	*Point, L^0*	Mean center	Network graphs	Proximity polygons TIN, point buffer	Interpolation. Kernel density estimation Distance surfaces
	Line, L^1	Intersection	Shortest distance path	Line buffer	Distance to nearest line object surface
	Area, L^2	Centroid	Graph of area skeleton	Area buffer, Polygon overlay	Pycnophylatic interpolation and other surface models
	Field, L^3	Surface specific points VIPs	Surface network	Watershed delineation, Hill masses	Equivalent vector field

intersection defines one or more possible geometric operations. Whether you are a novice or expert user of a GIS, it is worth while spending a little time on the following thought exercise.

Thought Exercise: Geometry and GIS

If you already know enough to navigate your way around a GIS, we invite you to see how many of these operations can be achieved using it. If you are new to geographic information analysis, it is worthwhile to check your favorite GIS textbook for examples of each of these operations. We would be the first to admit that our matrix is certainly incomplete! We also recognize, following Chrisman (1999), that it oversimplifies the transformations involved.

Why is the ability to change the way we represent spatial entities important? First, as our example of cases of a disease around a suspect facility indicates, we might have hypotheses that can only be tested if we make use of appropriate geometric transformation. Second, it is very often the case that changing the way we represent our spatial data types allows us to gain a new perspective on our problem and may in itself lead easily and directly to a solution. The next exercise gives an example.

Changing the Representation: Delimiting Town Centers

A town center (or downtown) is a good example of an area object with uncertain boundaries. We all know when we are in one, but precisely where we enter it is more difficult and the criteria we use vary from place to place. In the United Kingdom, the need to monitor the economic health of town centers led in the 1990s to a desire in government to develop a consistent set of town center boundaries relevant to all towns in the country. A paper by Mark Thurstain-Goodwin and David Unwin (2000) reports on the method that was adopted. It gives a good illustration of how changing the representation by a geometric transformation led to the development of a working system.

In the United Kingdom, the increasing availability of high-resolution spatial data using the so-called unit post code as its georeference not only makes this transformation useful, it also make it essential for analysis. At this very high level of spatial resolution in which data on a series of urban functions (retail, entertainment, commercial, and so on) are known as nearly exact (*x*, *y*) locations, use of these data as either point or area objects is not easy. What happens is that the intrinsic spatial ''granularity'' of these functions makes it difficult to apply any of the traditional point- or area-based methods. The alternative that was developed used kernel density estimation (see Section 3.6) to transform the data from point or area objects into continuous surfaces of spatial densities. Town centers could then be delineated by choosing appropriate contours on the density surfaces as their boundaries.

Viewed from our perspective as geographic information analysts, these transformations share a characteristic that can be worrying but that is often forgotten. With some exceptions, such as kernel density estimation and spatial interpolation using kriging, all are deterministic operations assuming that, since the input data are exact (x, y) coordinates and the processes are simple arithmetic manipulations performed by computer using many significant digits, the outputs must similarly be, to all intents and purposes, also exact. What this view forgets is that in any GIS these same coordinates are themselves a digital representation of the real world and that this representation cannot be exact. Locations are often found using error-prone semiautomatic digitizing or are badly recorded in some original field survey. The result of any geometric operations on such data will to a greater or lesser extent carry forward this uncertainty into any outputs. If there is further uncertainty introduced by the algorithms we apply to the data, as in kernel density estimation (Section 3.6), interpolation (Chapters 9 and 10), and overlay (Chapter 11), then the situation becomes even more complex. In Section 1.3 we encountered one consequence of such errors when discussing the true length of a line object, such as a coastline, thought to be of fractal dimension. Even if we know any line connecting two digitized locations to be straight, what is the impact of uncertainty in these coordinate locations on the true position of the line? Similarly, if we have a digitized outline of a wooded area, what is the impact of a similar error on our estimate of the true wooded area? And how are such errors propagated through into the results of a complex series of spatial geometric manipulations?

These questions have been addressed in the research literature (see, for example, Heuvelink et al., 1989; Heuvelink, 1993; Heuvelink, Burrough, 1993), but there is little evidence that their implications are being carried forward into routine work with spatial data and GISs. Over a decade ago, one of us (Unwin, 1995) reviewed the literature on error and uncertainty in GISs and concluded that systems need to be sensitive to these issues. The simple truth is that they (and the great majority of their users) are not.

1.6. THE ROAD AHEAD

In the remainder of this book, we take you on a tour of the field of geographic information analysis. We have organized the tour in what we hope you will find is a logical way. The next chapter looks at some of the big problems of spatial analysis—what makes spatial statistical analysis different from standard statistical analysis and the pitfalls and potential therein. Chapter 3 looks at methods by which spatial data can be visualized. Chapter 4 describes some fundamental issues in the analysis of spatial data, defining the important concepts of pattern and process, and Chapter 5 deals with the description and statistical analysis of point patterns. Chapter 6 looks at more recent approaches to this important topic. The critical property of spatial autocorrelation is introduced in Chapter 7, which deals with analysis of area objects. Chapter 8 brings together these ideas with some of the visualization materials from Chapter 3 to look at the relatively recently developed idea of local statistics. Chapters 9 and 10 deal with the analysis of continuous fields. In Chapter 11, we look at map overlay operations from a spatial analytic perspective. Finally, Chapter 12 describes some newer directions and developments in spatial analysis.

Throughout, we have tried to keep the level of mathematics as low as possible, but there are places where we have to draw on what to some may be unfamiliar matrix algebra. To help you, we have included an Appendix that summarizes the basics you need to know. If your mathematics is a little rusty, we suggest that you have a look at this appendix now.

CHAPTER REVIEW

- *Spatial analysis* is just one of a whole range of analytical techniques available in geography. It should be distinguished from *spatial data manipulations*, on the one hand, and *spatial modeling*, on the other.
- For the purposes of this book, *geographic information analysis* is the study of techniques and methods to enable the representation, description, measurement, comparison, and generation of spatial patterns.
- *Exploratory*, *descriptive*, and *statistical* techniques may be applied to spatial data to investigate the patterns that may arise as a result of processes operating in space.
- Spatial data may be of various broad types: *points*, *lines*, *areas,* and *fields*. Each type typically requires different techniques and approaches.
- The relationship between *real geographic entities* and spatial data is complex and *scale-dependent*.
- Representing geographic reality as points, lines, areas, and fields is *reductive*, and this must be borne in mind in all subsequent analysis.
- These objects are frequently not as simple as this geometric view leads one to assume. They may exist in *three spatial dimensions*, move and change over *time*, have a representation that is strongly *scale-dependent*, relate to entities that are themselves *fuzzy* and/or have *indeterminate boundaries*, or even be *fractal.*
- Although we have emphasized the difference between spatial analysis and GIS operations, the two are interrelated, and *most current spatial analysis is carried out on data stored and prepared in GISs.*
- Simple geometric *transformations* enable us to change the way entities are represented, and this might be useful for analysis.
- Finally, in any analysis of geographic information, we need to develop a sensitivity to the likely *sources of error* in our results.

REFERENCES

Bailey, T. C. and Gatrell, A. C. (1995) *Interactive Spatial Data Analysis* (Harlow, England: Longman).

Batty, M. and Longley, P. A. (1994) *Fractal Cities: A Geometry of Form and Function* (London: Academic Press).

Burrough, P. A. (1981) Fractal dimensions of landscapes and other environmental data, *Nature*, 294: 240–242.

Chrisman, N. R. (1998) Rethinking levels of measurement of cartography. *Cartography and GIS*, 25: 231–242.

Chrisman, N. R. (1999) A transformational approach to GIS operations. *International Journal of Geographical Information Science*, 13: 617–637.

Cressie, N. (1991) *Statistics for Spatial Data* (Chichester, England: Wiley).

DCDSTF—Digital Cartographic Data Standards Task Force (1988) The proposed standard for digital cartographic data. *The American Cartographer*, 15: 9–140.

Diggle, P. (1983) *Statistical Analysis of Spatial Point Patterns* (London: Academic Press).

Duckham, D. and Drummond, J. (2000) Assessment of error in digital vector data using fractal geometry, *International Journal of Geographical Information Science*, 14: 67–84.

Fisher, P. F. and Unwin, D. J., eds. (2005) *Re-presenting GIS* (Chichester, England: Wiley).

Ford, A. (1999) *Modeling the Environment: An Introduction to System Dynamics Models of Environmental Systems* (Washington, DC: Island Press).

Fotheringham, S., Brunsdon, C., and Charlton, M. (1999) *Quantitative Geography: Perspectives on Spatial Data Analysis* (London: Sage).

Goodchild, M. F. (1980) Fractals and the accuracy of geographical measures, *Mathematical Geology*, 12: 85–98.

Goodchild, M. F. and Mark, D. M. (1987) The fractal nature of geographic phenomena. *Annals of the Association of American Geographers*, 77: 265–278.

Haynes, R. M. (1975) Dimensional analysis: some applications in human geography. *Geographical Analysis*, 7: 51–67.

Haynes, R. M. (1978) A note on dimensions and relationships in human geography, *Geographical Analysis*, 10: 288–292.

Heuvelink, G. B. M. (1993) Error propagation in quantitative spatial modelling: applications in geographical information systems. *Netherlands Geographical Studies.* Utrecht: University of Utrecht.

Heuvelink, G. B. M. and Burrough, P. A. (1993) Error propagation in logical cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Systems*, 7: 231–246.

Heuvelink, G. B. M., Burrough, P. A., and Stein, A. (1989) Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information* Systems, 3: 303–322.

Hearnshaw, H. and Unwin, D. J., eds. (1994) *Visualisation in Geographical Information Systems*, (London: Wiley).

Lam, N. and De Cola, L., eds. (1993) *Fractals in Geography* (Englewood Cliffs, NJ: Prentice Hall).

Langran, G. (1992) *Time in Geographic Information Systems* (London: Taylor & Francis).

Mandelbrot, B. M. (1977) *Fractals: Form, Chance and Dimension* (San Francisco: Freeman).

McHarg, I. (1969) *Design with Nature* (Philadelphia: Natural History Press).

Mitchell, A. (1999) *The ESRI Guide to GIS Analysis* (Redlands, CA: ESRI Press).

O'Sullivan, D. (2005) Geographical information science: time changes everything. *Progress in Human Geography*, 29(6): 749–756.

Raper, J. F. (2000) *Multidimensional Geographic Information Science* (London: CRC Press).

Richardson, L. F. (1961) The problem of contiguity. *General Systems Yearbook*, 6: 139–187.

Ripley, B. D. (1981) *Spatial Statistics* (Chichester, England: Wiley).

Ripley, B. D. (1988) *Statistical Inference for Spatial Processes* (Cambridge: Cambridge University Press).

Stevens, S. S. (1946) On the theory of scales of measurements. *Science*, 103: 677–680.

Thurstain-Goodwin, M. and Unwin, D. J. (2000) Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations, *Transactions in GIS*, 4(4): 305–317.

Tomlin, C. D. (1990) *Geographic Information Systems and Cartographic Modelling* (Englewood Cliffs, NJ: Prentice Hall).

Turcotte, D. L. (1997) *Fractals and Chaos in Geology and Geophysics*, 2nd ed. (Cambridge: Cambridge University Press).

Unwin, D. J. (1982) *Introductory Spatial Analysis* (London: Methuen).

Unwin, D. J. (1994) Visualization, GIS and cartography, *Progress in Human Geography*, 18: 516–522.

Unwin, D. J. (1995) Geographical information systems and the problem of error and uncertainty. *Progress in Human Geography*, 19: 549–558.

Velleman, P. F. and Wilkinson, L. (1993) Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 47: 65–72.

Wesseling, C. G., Karssenberg, D., Van Deursen, W. P. A., and Burrough, P. A. (1996) Integrating dynamic environmental models in GIS: the development of a Dynamic Modelling language. *Transactions in GIS*, 1: 40–48.

Wilson, A. G. (2000) *Complex Spatial Systems* (London: Pearson Education).

Worboys, M. F. (1992) A generic model for planar spatial objects. *International Journal of Geographical Information Systems*, 6: 353–372.

Worboys, M. F. (1995) *Geographic Information Systems: A Computing Perspective* (London: Taylor & Francis).

Worboys, M. F., Hearnshaw, H. M., and Maguire, D. J. (1990) Object-oriented data modeling for spatial databases. *International Journal of Geographical Information Systems*, 4: 369–383.

Zeiler, M. (1999) *Modeling Our World: The ESRI Guide to Geodatabase Design* (Redlands, CA: ESRI Press).

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 2

The Pitfalls and Potential of Spatial Data

CHAPTER OBJECTIVES

In this chapter, we attempt to:

- Justify the view that spatial data are in some sense special
- Identify a number of problems in the statistical analysis of spatial data associated with *spatial autocorrelation*, *modifiable areal units*, the *ecological fallacy*, and *scale*, what we call the "bad news"
- Outline the ideas of *distance*, *adjacency*, *interaction,* and *neighborhood* —the "good news" central to much spatial analysis
- Show how *proximity polygons* can be derived for a set of point objects
- Introduce the idea that these relations can be summarized using matrices and encourage you to spend some time getting used to this way of organizing geographic data

After reading this chapter, you should be able to:

- List four major problems in the analysis of geographic information
- Outline the geographic concepts of distance, adjacency, interaction, and neighborhood and show how these can be recorded using matrix representations
- Explain how proximity polygons and the related Delaunay triangulation can be developed for point objects

2.1. INTRODUCTION

It may not be obvious why spatial data require special analytic techniques, distinct from standard statistical analysis that might be applied to ordinary data. As the spatial aspect of more and more data has been recognized due to the diffusion of GIS technologies and the consequent enthusiasm for mapping, this is an important issue to understand clearly. The academic literature is replete with essays on why space either is or is not important, or on why we should take the letter *G* out of GIS. In this chapter, we consider some of the reasons why adding a spatial location to some attribute data changes them in fundamental ways.

There is bad news and good news here. Some of the most important reasons why spatial data must be treated differently appear as problems or pitfalls for the unwary. Many of the standard techniques and methods documented in statistics textbooks are found to have significant problems when we try to apply them to the analysis of spatial distributions. This is the bad news, which we deal with first in Section 2.2. The good news is presented in Section 2.3. This boils down to the fact that geospatial referencing inherently provides us with a number of new ways of looking at data and the relations among them. The concepts of *distance*, *adjacency*, *interaction,* and *neighborhood* are used extensively throughout this book to assist in the analysis of spatial data. Because of their all-encompassing importance, it is useful to introduce these concepts early on in a rather abstract and formal way so that you begin to get a feel for them immediately. Our discussion of these fundamentals also includes *proximity polygons*, which appear repeatedly in this book and are increasingly being used as an interesting way of looking at many geographical problems. We hope that by introducing these concepts and ideas early on, you will begin to understand what it means to *think spatially* about the analysis of geographic information.

2.2. THE BAD NEWS: THE PITFALLS OF SPATIAL DATA

Conventional statistical analysis frequently imposes a number of conditions or assumptions on the data it uses. Foremost among these is the requirement that samples be random. The most fundamental reason that spatial data are special is that they almost always violate this requirement. The technical term describing this problem is *spatial autocorrelation*, which must therefore come first on any list of the pitfalls of spatial data. Other closely related tricky problems frequently arise, including the *modifiable areal unit problem*, associated issues of *scale* and *edge effects,* and the *ecological fallacy*.

Spatial Autocorrelation

Spatial autocorrelation is a complicated name for the obvious fact that *data from locations near one another in space are more likely to be similar than data from locations remote from one another*. If you know that the elevation at point X is 250 m, then you have a good idea that the elevation at point Y, 10 m from X, is probably in the range 240 to 260 m. Of course, there could be a huge cliff between the two locations. Location Y *might* be 500 m above sea level, although it is highly unlikely. It will almost certainly not be 1000 m above sea level. On the other hand, location Z, 1000 m from X, certainly could be 500 m above sea level. It could even be 1000 m above sea level or even 100 m *below* sea level. We are much more uncertain about the likely elevation of Z because it is farther away from X. If Z is instead 100 km away from X, almost anything is possible, because knowing the elevation at X tells us very little about the elevation 100 km away.

If spatial autocorrelation were not commonplace, then geographic analysis would be of little interest and geography would be irrelevant. Again, think of spot heights: we know that high values are likely to be close to one another and in different places from low values—in fact, these are called *mountains* and *valleys*. Many geographic phenomena can be characterized in these terms as local similarities in some spatially varying phenomenon. Cities are local concentrations of population—and much else besides: economic activity and social diversity, for example. Storms are local foci of particular atmospheric conditions. Climate consists of the repeated occurrence of similar spatial patterns of weather in particular places. *If geography is worth studying at all, it must be because phenomena do not vary randomly through space*. The existence of spatial autocorrelation is therefore a given in geography. Unfortunately, it is also an impediment to the application of conventional statistics.

The nonrandom distribution of phenomena in space has various consequences for conventional statistical analysis. For example, the usual parameter estimates based on samples that are not randomly distributed in space will be biased toward values prevalent in the regions favored in the sampling scheme. As a result, many of the assumptions we are required to make about data before applying statistical tests become invalid. Another way of looking at this is that spatial autocorrelation introduces *redundancy* into data, so that each additional item of data provides less new information than is indicated by a simple assessment based on n, the sample size. This affects the calculation of confidence intervals and so forth. Such effects mean that there is a strong case for assessing the degree of autocorrelation in a spatial data set before doing any conventional statistics at all. Diagnostic measures for the autocorrelation present in data are available, such as *Moran's I,* and *Geary's C*, and these will be described in Chapter 7. Later, in Chapter 10, we introduce the *variogram cloud*, a plot that also helps us understand the autocorrelation pattern in a spatial data set.

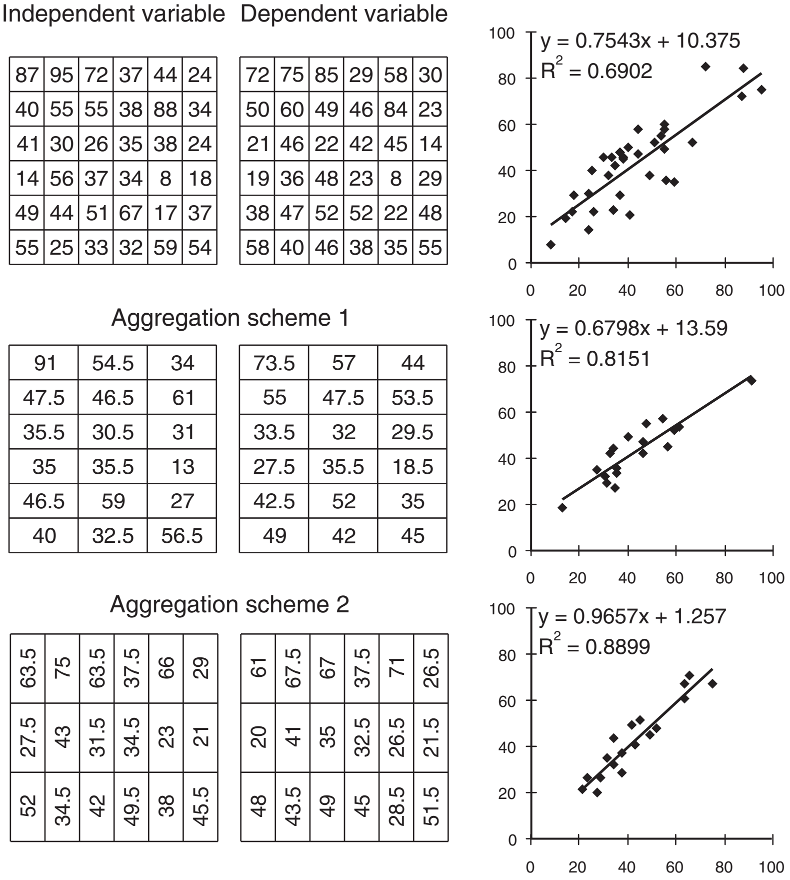
These techniques help us to describe how useful knowing the location of an observation is if we wish to determine the likely value of an attribute measured at that location. There are three general possibilities: *positive autocorrelation*, *negative autocorrelation,* and *noncorrelation* or *zero autocorrelation*. Positive autocorrelation is the most commonly observed case and refers to situations where nearby observations are likely to be similar to one another. Negative autocorrelation is much less common and occurs when observations from nearby locations are likely to be different from one another. Zero autocorrelation is the case where no spatial effect is discernible and observations seem to vary randomly through space. It is important to be clear about the difference between negative and zero autocorrelation, as students frequently confuse the two.

Describing and modeling patterns of variation across a study region, effectively *describing the autocorrelation structure*, is of primary importance in spatial analysis. Again, in general terms, spatial variation is of two kinds: first- and second-order. *First-order* spatial variation occurs when observations across a study region vary from place to place due to changes in the underlying properties of the local environment. For example, the rates of incidence of crime might vary spatially simply because of variations in the population density, such that they increase near the center of a large city. In contrast, *second-order* variation is due to interaction effects between observations, such as the occurrence of crime in an area making it more likely that there will be crimes surrounding that area, perhaps in the shape of local "hotspots" in the vicinity of bars and clubs or near local street drug markets. In practice, it is difficult to distinguish between first- and second-order effects, but it is often necessary to model both when developing statistical methods for handling spatial data. We discuss the distinction between first- and second-order spatial variation in more detail in Chapter 4 when we introduce the idea of a spatial process.

Although autocorrelation presents a considerable challenge to conventional statistical methods and remains problematic, quantitative geographers have made a virtue of it by developing a number of autocorrelation measures into powerful descriptive tools. It would be wrong to claim that the problem of spatial autocorrelation has been solved, but considerable progress has been made in developing techniques that account for its effects and in taking advantage of the opportunity it provides for useful geographic descriptions.

The Modifiable Areal Unit Problem

Another major difficulty with spatial data is that they are often aggregates of data originally compiled at a more detailed level. The best example is a national census, which is collected at the household level but reported for



Independent variable Dependent variable

87	95	72	37	44	24
40	55	55	38	88	34
41	30	26	35	38	24
14	56	37	34	8	18
49	44	51	67	17	37
55	25	33	32	59	54

72	75	85	29	58	30
50	60	49	46	84	23
21	46	22	42	45	14
19	36	48	23	8	29
38	47	52	52	22	48
58	40	46	38	35	55

Aggregation scheme 1

91	54.5	34
47.5	46.5	61
35.5	30.5	31
35	35.5	13
46.5	59	27
40	32.5	56.5

73.5	57	44
55	47.5	53.5
33.5	32	29.5
27.5	35.5	18.5
42.5	52	35
49	42	45

Aggregation scheme 2

63.5	75	63.5	37.5	66	29
27.5	43	31.5	34.5	23	21
52	34.5	42	49.5	38	45.5

61	67.5	67	37.5	71	26.5
20	41	35	32.5	26.5	21.5
48	43.5	49	45	28.5	51.5

Figure 2.1 An illustration of MAUP.

practical and privacy reasons at various levels of aggregation such as city districts, counties, and states. The problem is that the aggregation units used are *arbitrary* with respect to the phenomena under investigation, yet the units used will affect statistics determined on the basis of data reported in this way. This difficulty is referred to as the *modifiable areal unit problem* (MAUP). If the spatial units in a particular study were specified differently, we might observe very different patterns and relationships. The problem is illustrated in Figure 2.1, where two different aggregation schemes applied to a spatial data set result in two different regression results. There is a clear impact on the regression equation and the coefficient of determination, R^2. This is an artificial example, but the effect is general and is not widely understood, even though it has been known for a long time (see Gehlke and Biehl, 1934). Usually, as shown here, regression relationships are strengthened by aggregation. In fact, using a simulation approach, Openshaw and Taylor (1979) showed that with the same underlying data, it is possible to aggregate units together in ways that can produce correlations *anywhere* between −1.0 and +1.0!

The effect is not altogether mysterious, and two things are happening. The first relates to the scale of analysis and to aggregation effects such that combining any pair of observations will produce an outcome that is closer to the mean of the overall data, so that after aggregation, the new data are likely to be more tightly clustered around a regression line and thus to have a stronger coefficient of determination. This effect is shown in our example by both the aggregation schemes used producing better fits than the original disaggregated data. Usually this problem persists as we aggregate up to larger units. A second effect is the substantial differences between the results obtained under different aggregation schemes. The complications are usually referred to separately as the *aggregation* effect and the *zoning* effect.

MAUP is of more than academic or theoretical interest. Its effects have been well known for many years to politicians concerned with ensuring that the boundaries of electoral districts are defined in the most advantageous way for them. It provides one explanation for why, in the 2000 U.S. presidential election, Al Gore, with more of the popular vote than George Bush, still failed to become president. A different aggregation of U.S. counties into states could have produced a different outcome (in fact, it is likely that in this very close election, switching just one or two northern Florida counties to Georgia or Alabama would have produced a different outcome.)

The practical implications of MAUP are immense for almost all decision-making processes involving GIS technology, since with the now ready availability of detailed but still aggregated maps, policy could easily focus on issues and problems which might look very different if the aggregation scheme used were changed. The implication is that our choice of spatial reference frame is itself a significant determinant of the statistical and other patterns we observe. Openshaw (1983) suggests that a lack of understanding of MAUP has led many to choose to pretend that the problem does not exist in order to allow *some* analysis to be performed so that we can “just get on with it.” This is a little unfair, but the problem is a serious one that has had less attention than it deserves. Openshaw’s suggestion is that the problem be turned into an exploratory and descriptive tool, as has been done with spatial autocorrelation. In this view, we might postulate a relationship between, say, income and crime rates. We would then search for an aggregation scheme that maximizes the strength of this relationship. The output from such an analysis would be a spatial partition into areal units. The interesting geographic question then becomes “Why do these zones produce the strongest relationship?” Perhaps because of the computational complexities and the implicit requirement for very detailed individual-level data, this idea has not been widely taken up.

The Ecological Fallacy

MAUP is closely related to a more general statistical problem: the *ecological fallacy*. This arises when a statistical relationship observed at one level of aggregation is assumed to hold because the same relationship holds at a more detailed level. For example, we might observe a strong relationship between income and crime at the county level, with lower-income counties being associated with higher crime rates. If from this we conclude that lower-income individuals are more likely to commit a crime, then we are falling for the ecological fallacy. In fact, it is only valid to say exactly what the data say: that lower-income counties tend to experience higher crime rates. What causes the observed effect may be something entirely different—perhaps lower-income families have less effective home security systems and are more prone to burglary (a relatively direct link); or lower-income areas are home to more chronic drug users who commit crimes irrespective of income (an indirect link); or the cause may have nothing at all to do with income.

It is important to acknowledge that a relationship at a high level of aggregation *may* be explained by the same relationship operating at lower levels. For example, one of the earliest pieces of evidence to make the connection between smoking and lung cancer was presented by Doll (1955; cited by Freedman et al., 1998) in the form of a scatterplot showing per capita national rates of cigarette smoking and the rate of death from lung cancer for 11 countries. A strong correlation is evident in the plot. However, we would be wrong to conclude, based on this evidence alone, that smoking is a cause of lung cancer. It turns out that it is, but this conclusion is based on many other studies conducted at the individual level. Data on smoking and cancer at the country level can still only support the conclusion that countries with larger numbers of smokers tend to have higher death rates from lung cancer.

Having now been made aware of the problem, if you pay closer attention to the news, you will find that the ecological fallacy is common in everyday and media discourse. Crime rates and (variously) the death penalty, gun control or imprisonment rates, and road fatalities and speed limits, seat belts, or cycle helmet laws are classic examples. Unfortunately, the fallacy is almost as common in academic discourse! It often seems to arise from a desire for simple explanations, but in human geography things are rarely so simple. The common thread tying the ecological fallacy to MAUP is that statistical relationships may change at different levels of aggregation.

Scale

This brings us neatly to the next point. The *geographic scale* at which we examine a phenomenon can affect the observations we make, and this must always be considered prior to spatial analysis. We have already encountered one way that scale can dramatically affect spatial analysis since the object type that is appropriate for representing a particular entity is scale dependent. For example, at the continental scale, a city is conveniently represented by a point. At the regional scale, it becomes an area object. At the local scale, the city becomes a complex collection of point, line, area, and network objects. The scale we work at affects the representations we use, and this in turn is likely to have effects on spatial analysis; yet, in general, the correct or appropriate geographic scale for a study is impossible to determine beforehand, and due attention should be paid to this issue.

Nonuniformity of Space and Edge Effects

A final significant issue distinguishing spatial analysis from conventional statistics is that *space is not uniform*. For example, we might have data on crime locations gathered for a single police precinct. It is very easy to see patterns in such data, hence the *pin maps* (see Chapter 3) seen in any self-respecting movie police chief's office. Patterns may appear particularly strong if crime locations are mapped simply as points without reference to the underlying geography. There will almost certainly be clusters simply as a result of where people live and work, and apparent gaps in (for example) parks or at major road intersections. These gaps and clusters are not unexpected but arise as a result of the nonuniformity of the urban space with respect to the phenomenon being mapped. Similar problems are encountered in examining the incidence of disease, where the location of the at-risk population must be considered. Such problems also occur in point patterns of different plant types, where underlying patterns in soil types, or simply the presence of other plant types, might lead us to expect variation in the spatial density of the plants we're interested in.

A particular type of nonuniformity problem, which is almost invariably encountered, is due to *edge effects*. These arise where an artificial boundary is imposed on a study, often just to keep it manageable. The problem is that sites in the center of the study area can have nearby observations in all directions, whereas sites at the edges of the study area only have neighbors toward the center of the study area. Unless the study area has been very carefully defined, it is unlikely that this reflects reality, and the artificially produced asymmetry in the data must be accounted for. In some specialized areas of spatial analysis, techniques for dealing with edge effects are well developed, but the problem remains poorly understood in many cases.

2.3. THE GOOD NEWS: THE POTENTIAL OF SPATIAL DATA

It should not surprise you to find out that many of the problems outlined in Section 2.2 have not been solved satisfactorily. Indeed, in the early enthusiasm for the quantitative revolution in geography (in the late 1950s and the 1960s), many of these problems were glossed over. Unfortunately, dealing with these problems is not simple, so that only more mathematically oriented geographers and relatively small numbers of statisticians have paid much attention to the issues. More recently, with the advent of GIS and a much broader awareness of the significance of spatial data, there has been considerable progress, with new techniques appearing all the time. Regardless of the sophistication and complexity of the techniques adopted, the fundamental characteristics of spatial data are critical to unlocking their potential. To give you a sense of this, we continue our overview of what's special about spatial data, not focusing on the problems that consideration of spatial aspects introduces, but instead examining some of the potential for additional insight provided by an examination of the locational attributes of data.

The important spatial concepts that appear throughout this book are *distance*, *adjacency*, and *interaction*, together with the closely related notion of *neighborhood*. These appear in a variety of guises in most applications of statistical methods to spatial data. Here we point to their importance, outline some of their uses, and indicate some of the contexts where they will appear. The reason for the importance of these ideas is clear. In spatial analysis, while we are still interested in the distribution of values in observational data (classical descriptive statistical measures like the mean, variance, and so on), we are now *also* interested in the distribution of the associated entities *in space*. This *spatial distribution* can only be described in terms of the relationships between spatial entities, and spatial relationships are usually conceived in terms of one or more of the relationships we call distance, adjacency, interaction, and neighborhood.

Distance

Distance is usually (but not always) described by the simple crow's flight distance between the spatial entities of interest. In small study regions, where the Earth's curvature effects can be ignored, simple *Euclidean* *distances* are usually adequate and may be calculated using Pythagoras's familiar formula, which tells us that

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \qquad (2.1)$$

is the distance between two points located by their spatial coordinates (x_i, y_i) and (x_j, y_j). Over larger regions, more complex calculations may be required to take account of the curvature of the Earth's surface.

Euclidean, straight-line, or crow's flight distances are the simplest, but there are many other mathematical measures of distance that we could adopt. It might, for example, be necessary to consider distances measured over an intervening road, rail, river, or air transport network, and such notions of distance significantly extend the scope of the idea. It is a short step to go from distance over a road network to expected driving time. Distance is then no longer measured in kilometers, but rather in units of time (hours and minutes). Such broader concepts of distance can be nonintuitive and contradictory. For example, we might have a distance measure based on the *perceived travel time* among a set of urban landmarks. We might collect such data by surveying a number of people and asking them how long it takes to get from the museum to the railway station, for instance. These alternative distances can exhibit some very odd properties. It may, for example, be generally perceived to take longer to get from A to B than from B to A. Such effects are not absent from real distances, however. In a city, the structure of the transport network can affect distances, making them *actually* vary at different times of the day or in different directions. As another example, transatlantic flight times at northern latitudes (from the U.S. Eastern Seaboard to Western Europe) generally vary considerably, being shorter flying east, with the prevailing winds, than flying west against the same winds.

Sadly, in most of this book, we ignore these complexities and assume that simple Euclidean distance is adequate. If you are interested, Gatrell's book *Distance and Space* (1983) explores some of these intricacies and is highly recommended. You might also undertake the research suggested in the thought exercise that follows.

Thought Exercise: Conceptions of Distance

In *Distance and Space*, Gatrell (1983) shows that, however defined, distance is an example of a relationship between elements of a set, in this case of spatial locations. This exercise uses a journey in London (England) but could easily be modified for any other city with which you are familiar.

Our objective is to get from the Euston main-line railway station to the Waterloo main-line station using different notions of the distance between them, according to the methods of transport (and the size of our wallet!).

1. First of all, what is the straight-line distance between these stations? A suitable map can be found at the Transport for London Web site: tfl.gov.uk/assets/downloads/Central-London-Day-Bus-Map.pdf. You will need to use a ruler and knowledge of the scale of the map to answer this question.
2. Second, suppose you were to hire a taxicab for the same journey. These cabs usually take the shortest route, but the distance is obviously governed by the roads followed, and in any case, you would almost certainly be interested in the cost. There is a schedule of taxi fares for London that you can use to estimate this cost at www.tfl.gov.uk/gettingaround/taxisandminicabs/taxis/1140.aspx.
3. Finally, of course, many Londoners would make the same journey by the Underground (the famous ''Tube'') at a standard fare for a journey in what's called Zone 1. The distance of concern would almost certainly be the time the journey takes. The same Web site has a trip choice aid at www.tfl.gov.uk/gettingaround that will estimate this for you, and you can see which Tube line is involved by looking at www.tfl.gov.uk/assets/downloads/standardtube-map.pdf.

So, in an applied problem—getting from station to station—which distance is appropriate?

Adjacency

Adjacency can be thought of as the nominal, or binary, equivalent of distance. Two spatial entities are either adjacent or they are not. Of course, how adjacency should be determined is not necessarily clear. The most obvious case is a set of polygons, in which we consider any two polygons that share an edge to be adjacent. An equally simple formulation is to decide that any two entities within some fixed distance of one another (say 100 m) are adjacent to one another. Alternatively, we might decide that the six nearest entities to any particular entity are adjacent to it. We might even decide that only the single *nearest neighbor* is adjacent.

Thought Exercise: A Map Based on Adjacency

If you did the previous exercise, the final map you looked at was a representation of the London Underground Railway network based on one of the most famous maps of all time: Harry Beck's iconic map dating from 1933. Note that its ''distances'' are really determined by whether or not stations are adjacent to each other. Similar maps based on conceptions of distance such as time, cost, or even human perception can be drawn, and these same ideas extended to create what in Chapter 3 we call *spatialization* of information that is not intrinsically spatial at all.

As with distance, we can play with the adjacency concept, and two entities that are adjacent may not necessarily be near each other. A good example of this is provided by the structure of scheduled air transport connections between cities. In the British Isles, it is possible to fly between London and Belfast, or between London and Dublin, but not between Belfast and Dublin. If adjacency is equated with connection by scheduled flights, then London is adjacent to Belfast and Dublin (both roughly 500 km away), but the two Irish cities (only 136 km apart) are not adjacent to each other. Adjacency is an important idea in the measurement of autocorrelation effects when a region is divided into areal units (Chapter 7) and in spatial interpolation schemes (Chapters 9 and 10).

Interaction

Interaction may be considered as a combination of distance and adjacency, and rests on the intuitively obvious idea that nearer things are more closely related than distant things—a notion often referred to as the "first law" of geography (see Tobler, 1970). Mathematically, we often represent the degree of interaction between two spatial entities as a number between 0.0 (no interaction) and 1.0 (a high degree of interaction). If we represent adjacency in the same way, it can be measured on the same scale with only 0 (nonadjacent) or 1 (adjacent) allowed, because adjacency is binary. Typically, in spatial analysis, the interaction between two entities is determined by some sort of *inverse distance weighting*. A typical formulation is

$$w_{ij} \propto \frac{1}{d^k} \tag{2.2}$$

where w_{ij} is the interaction *weight* between the two entities i and j that are a distance d apart in space. The distance exponent, k, controls the rate of decline of the weight. An *inverse power law* for interaction like this ensures that entities close together have stronger interactions than those farther apart. Often, the interaction between two entities is positively weighted by some attribute of those entities. A common formulation uses some measure of the size of the entities, such as the populations, p_i and p_j. This gives us a modified interaction weight

$$w_{ij} \propto \frac{p_i p_j}{d^k} \tag{2.3}$$

Working with purely spatial characteristics of entities, we might positively weight the interaction between two areal units by their respective areas and divide by the distance between their centers.

As with distance, measures other than simple geographic distance may be appropriate in different contexts. For example, we might think of the trade volume between two regions or countries as a measure of their degree of interaction. Interaction of the simple geometric kind is important to the study of simple interpolation methods discussed in Chapters 9 and 10.

Neighborhood

Finally, we may wish to employ the concept of *neighborhood*. There are a number of ways of thinking about this. We might, for example, define the neighborhood of a particular spatial entity as the set of all other entities adjacent to the entity we are interested in. This clearly depends entirely on how we determine the adjacencies. Alternatively, the neighborhood of an entity may also be defined not with respect to sets of adjacent entities, but as a region of space associated with that entity and defined by distance from it. An approach closer than either of these to the common use of the word neighborhood is the idea that regions in a spatial distribution that are alike are neighborhoods distinct from other regions which are also internally similar, but different from surrounding regions. This notion of neighborhood is very general indeed. For example, many geographic objects may be thought of as local neighborhoods in numerical field data. What we call a *mountain* is a neighborhood in a field of elevation values that is distinguished by its consisting of higher values than those in surrounding regions.

Figure 2.2 illustrates versions of these four fundamental concepts. In the upper left panel, the distance between the central point object A and the others in the study region has been measured and is indicated. Generally speaking, it is always possible to determine the distance between a pair of objects. In the

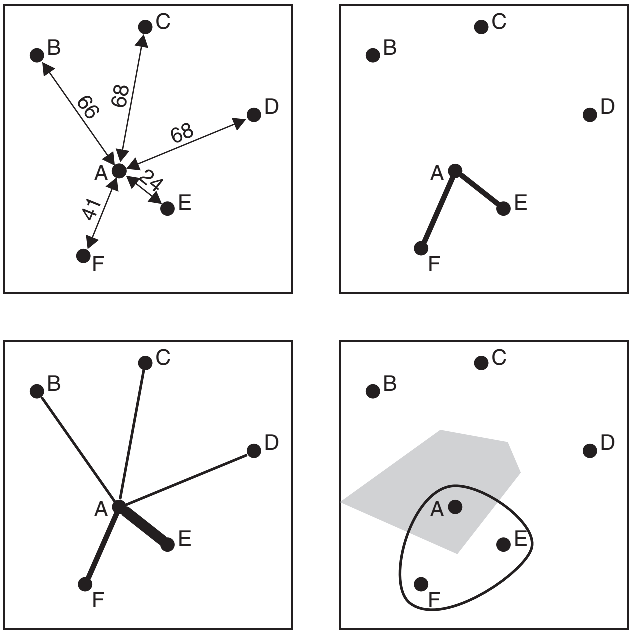


Figure 2.2 A schematic representation of the distance, adjacency, interaction, and neighborhood concepts.

second panel, adjacency between object A and two others (E and F) is indicated by the lines joining them. In this case, objects E and F are the two that are closest to A in terms of the distances shown in the first panel. This definition of adjacency might have been arrived at by a number of methods. For example, we might have decided that pairs of objects within 50 m of one another are adjacent. Notice that this definition would mean that the object labeled D has no adjacent objects. An alternative definition might be that the two objects closest to each object are adjacent to it. This would guarantee that all the objects have two other adjacent objects, although it would also mean that adjacency was no longer a *symmetrical* relationship. For example, on this definition, E is adjacent to D (whose two nearest neighbors are C and E), but D is *not* adjacent to E (whose two nearest neighbors are A and F). In the third panel at the lower left, an interaction measure is indicated by the line thickness drawn between A and every other object. The interaction weight here is inversely related to the distances in the first panel, so that interaction between A and E is strongest, and is weak between A and each of B, C, and D. In the final panel, two possible ideas of the neighborhood of object A are illustrated. The outlined curved area is the set of objects adjacent to A, which includes A, E, and F. An object is usually considered to be adjacent to itself, as here. Another possible interpretation is the shaded polygon, which is the region of this space that is closer to A than to any other object in the region.

Summarizing Relationships in Matrices

One way of pulling all these concepts together is to note that they may all be represented conveniently in *matrices*. If you know nothing at all about matrices, we advise you to read the Appendix, where some of the mathematics of matrices is introduced. Simply put, a matrix is a table of numbers organized in rows and columns; for example,

$$\begin{bmatrix} 2 & 1 \\ 5 & 3 \end{bmatrix} \tag{2.4}$$

is a *two-by-two* (2×2) *matrix* with two *rows* and two *columns*. Matrices are normally written this way with square brackets. Now we can summarize the information on distances in any spatial data using a *distance matrix* such as

$$\mathbf{D} = \begin{bmatrix} 0 & 66 & 68 & 68 & 24 & 41 \\ 66 & 0 & 51 & 110 & 99 & 101 \\ 68 & 51 & 0 & 67 & 91 & 116 \\ 68 & 110 & 67 & 0 & 60 & 108 \\ 24 & 99 & 91 & 60 & 0 & 45 \\ 41 & 101 & 116 & 108 & 45 & 0 \end{bmatrix} \tag{2.5}$$

where the uppercase, boldface letter **D** denotes the entire table of numbers. The distances in this matrix are all the distances between objects A, B, C, D, E, and F in Figure 2.2. Notice that the first row represents object A, with its series of distances to objects B, C, D, E, and F, respectively of 66, 68, 68, 24, and 41 m. A number of things are important to note:

- The row and column orders in the matrix are the same: both are in the order ABCDEF.
- This means that because it contains the “distance” from each object to itself, the *main diagonal* of the matrix from top left to bottom right has all zeros.
- The matrix is *symmetrical* about the main diagonal, so that (for example) the number in the third row, fourth column is equal to the number in the fourth row, third column (equal to 67). This is because these elements record the distance from C to D and from D to C, which are identical.

All the distance information for the data set is contained in the matrix. Therefore, any analysis based on these distances alone can be performed using the matrix.

In much the same way, although matrix elements are now 1s or 0s, we can construct an *adjacency matrix*, **A**, for the same set of objects:

$$\mathbf{A}_{d\leq 50} = \begin{bmatrix} * & 0 & 0 & 0 & 1 & 1 \\ 0 & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & 0 & 0 & 0 \\ 0 & 0 & 0 & * & 0 & 0 \\ 1 & 0 & 0 & 0 & * & 1 \\ 1 & 0 & 0 & 0 & 1 & * \end{bmatrix} \tag{2.6}$$

This is the matrix we get if the rule for adjacency is that two objects must be less than 50 m apart. Again, the matrix is symmetrical. Notice that if we sum the numbers in any row or column, we get the number of objects adjacent to the corresponding object. Thus, the row total for the first row is 2, which corresponds to the fact that under this definition, object A has two adjacent objects. Notice that we have put a $*$ symbol in the main diagonal positions, because it is not clear if an object is adjacent to itself or not. In specific applications, it may be appropriate to consider objects as adjacent to themselves or not.

Using a different adjacency rule gives us a different matrix. If the rule for adjacency is that each object is adjacent to its three nearest neighbors, then we get a different **A** matrix:

$$\mathbf{A}_{k=3} = \begin{bmatrix} * & 1 & 0 & 0 & 1 & 1 \\ 1 & * & 1 & 0 & 1 & 0 \\ 1 & 1 & * & 1 & 0 & 0 \\ 1 & 0 & 1 & * & 1 & 0 \\ 1 & 0 & 0 & 1 & * & 1 \\ 1 & 1 & 0 & 0 & 1 & * \end{bmatrix} \tag{2.7}$$

Notice that the matrix is no longer symmetrical, because, as already mentioned, a "nearest-neighbors" rule for adjacency makes the relationship asymmetric. Each row sums to 3, as we would expect, but the column totals of 5, 3, 2, 2, 4, and 2 are different. This is because the definition of adjacency is such that E being adjacent to B does not guarantee that B is adjacent to E. We can see from this matrix that object A is actually adjacent to all the other objects. This is due to its central location in the study area.

Finally, we can construct an *interaction* or *weights matrix*, **W**, for this data set. If we use a simple inverse distance (1/*d*) rule, then we get the following matrix:

$$\mathbf{W} = \begin{array}{cc} & \text{row totals:} \\ \begin{bmatrix} \infty & 0.0152 & 0.0147 & 0.0147 & 0.0417 & 0.0244 \\ 0.0152 & \infty & 0.0196 & 0.0091 & 0.0101 & 0.0099 \\ 0.0147 & 0.0196 & \infty & 0.0149 & 0.0110 & 0.0086 \\ 0.0147 & 0.0091 & 0.0149 & \infty & 0.0167 & 0.0093 \\ 0.0417 & 0.0101 & 0.0110 & 0.0167 & \infty & 0.0222 \\ 0.0244 & 0.0099 & 0.0086 & 0.0093 & 0.0222 & \infty \end{bmatrix} & \begin{array}{c} 0.1106 \\ 0.0639 \\ 0.0688 \\ 0.0646 \\ 0.1016 \\ 0.0744 \end{array} \end{array} \tag{2.8}$$

Note that the main diagonal elements have a value of infinity. Often these elements are ignored because infinity is a difficult number to deal with. A common variation on the weights matrix is to adjust the values in each row so that they sum to 1. Row totals for the matrix (discounting the infinity values) are shown above, so we divide each entry in the first row by 0.1106, the second row entries by 0.0639, and so on, to get

$$\begin{array}{cc} \mathbf{W} = & \begin{bmatrix} \infty & 0.1370 & 0.1329 & 0.1329 & 0.3767 & 0.2205 \\ 0.2373 & \infty & 0.3071 & 0.1424 & 0.1582 & 0.1551 \\ 0.2136 & 0.2848 & \infty & 0.2168 & 0.1596 & 0.1252 \\ 0.2275 & 0.1406 & 0.2309 & \infty & 0.2578 & 0.1432 \\ 0.4099 & 0.0994 & 0.1081 & 0.1640 & \infty & 0.2186 \\ 0.3279 & 0.1331 & 0.1159 & 0.1245 & 0.2987 & \infty \end{bmatrix} \\ \begin{array}{c}\text{column} \\ \text{totals:}\end{array} & \begin{array}{cccccc} 1.4161 & 0.7949 & 0.8949 & 0.7805 & 1.2510 & 0.8626 \end{array} \end{array} \tag{2.9}$$

In this matrix, each row sums to 1. Column totals now reflect how much interaction effect or influence the corresponding object has on all the other objects in the region. In this case, column 1 (object A) has the largest total, again reflecting its central location. The least influential object is D, with a column total of only 0.7805 (compared to A's total of 1.4161).

The important point to take from this section is not any particular way of analyzing the numbers in a distance, adjacency, or interaction weights matrix, but the fact that this organization of the spatial data is helpful in analysis. Matrix-based methods become increasingly important as more advanced techniques are applied. Sometimes this is because various mathematical manipulations of matrices produce a new perspective on the data. Often, however, it is simply because concise description of rather complex mathematical manipulations is possible using matrices, and this helps us to develop techniques further. You will be seeing more of matrices elsewhere in this book.

Thought Exercise

If you still haven't read the Appendix, then do so now. A few hours of effort won't make you a great mathematician, but it will pay enormous future dividends as you explore the world of geographic information analysis.

Proximity Polygons

Another very general tool in specifying the spatial properties of a set of objects is the partitioning of a study region into *proximity polygons*. This procedure is most easily explained by starting with the proximity polygons of a simple set of point objects. The proximity polygon of any entity is that region of the space which is closer to the entity than it is to any other. This is shown in Figure 2.3 for a set of point entities. Proximity polygons are also known as *Thiessen* or *Voronoi polygons* and have been rediscovered several times in many disciplines (for a *very* thorough review, see Okabe et al., 2000). Although this is a computationally inefficient approach, for point objects the polygons are surprisingly easy to construct using the perpendicular bisectors of lines joining pairs of points, as shown in Figure 2.4.

More complex constructions are required to determine the proximity polygons for line and area objects. However, it is always possible to partition a region of space into a set of polygons, each of which is nearest to a specific object of any kind—point, line, or area—in the region. This is true even for

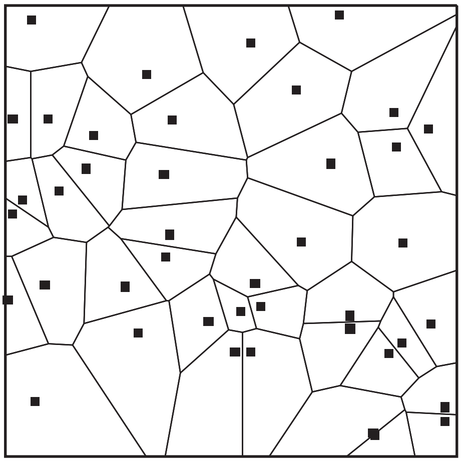


Figure 2.3 The proximity polygons for a set of point events.

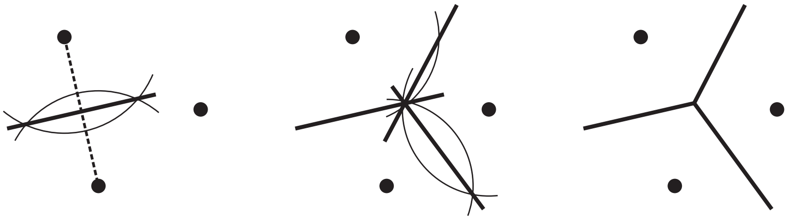


Figure 2.4 Construction of proximity polygons. Polygon edges are all perpendicular bisectors of lines joining pairs of points.

mixed sets of objects, where some are points, some are lines, and some are areas. The idea of proximity polygons is therefore very general and powerful. In fact, it can also be applied in three dimensions when the polygons become like bubbles. Note that the polygons always fill the space without overlapping, since any particular location must be closest to only one object, or, if it is equidistant from more than one object, it lies on a polygon boundary.

From a set of proximity polygons we can derive at least two different concepts of neighborhood. The first is the obvious one. The proximity polygon associated with an entity is its neighborhood. This idea has some geographically useful applications. For example, the proximity polygons associated with a set of (say) post offices allow you quickly to decide which is the closest—it's the one whose polygon you're in! The same idea may be applied to other types of buildings, such as schools, hospitals, supermarkets, and so forth. The proximity polygon of a school is often a good first approximation to its catchment area, for example.

A second concept of neighborhood may also be developed from proximity polygons. Thinking again of point objects, we may join any pair of points whose proximity polygons share an edge. The resulting construction is shown in Figure 2.5 and is known as the *Delaunay triangulation*. A triangulation of a set of points is any system of interconnections between them that forms a set of triangles. The Delauanay triangulation is frequently used, partly because its triangles are as near to equilateral as possible. This is useful, for example, in constructing representations of terrain from spot heights.

One criticism of the other approaches to neighborhood and adjacency that we have examined is that they ignore the nonuniformity of geographic space, because they simplistically apply an idea like "Any two objects less than 100 m apart are adjacent," regardless of the number of other objects nearby. Although proximity polygons do not address this criticism directly, the neighborhood relations that they set up are determined with respect to local patterns in the data rather than by using criteria such as "nearest neighbor" or "within 50 m." This may be an advantage in some cases. The proximity approach is also easily extended to nonuniform spaces if determination of distance is done over (say) a street network rather than a plane area (see

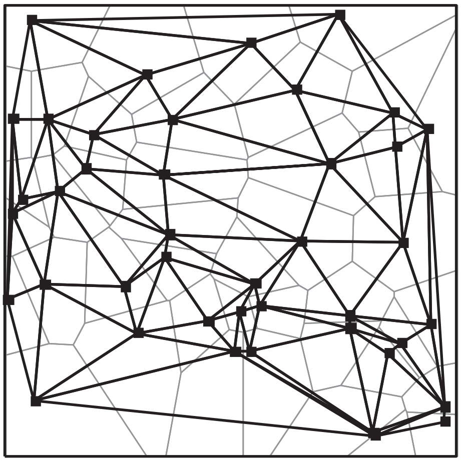


Figure 2.5 Derivation of the Delaunay triangulation from proximity polygons.

Okabe et al., 2000, 2008). Other versions of the idea include defining polygons that are regions where objects are second closest, third closest, or even farthest away. These constructions are generally more complex however, with overlapping regions, and they have less obvious application.

Approaches to analysis based on proximity polygons are still relatively unexplored, because construction of the polygons, although simple, is extremely tedious. Recently, researchers have started to take advantage of the ready availability of computer processing power, and the idea is becoming increasingly widely used in many areas of spatial analysis. We will encounter it again.

CHAPTER REVIEW

- *Autocorrelation* undermines conventional inferential statistics due to redundancy in data arising from similarity in nearby observations.
- The *modifiable areal unit problem* (MAUP) also undermines conventional methods, especially regression.
- As always in geography, *scale* can have a significant impact on spatial analysis, and choosing an appropriate scale is an important first step in all spatial analysis.
- The *nonuniformity* of space is also problematic. *Edge effects* are almost always present and should be considered.
- Although these issues remain problematic, the last 30 years or so have seen progress on many of them, and spatial analysis involving quantitative geographic methods spatial analysis is more sophisticated now

than it was when it was heavily criticized in human geography in the 1970s and 1980s.

- Important concepts in geographic information analysis are *distance*, *adjacency*, *interaction*, and *neighborhood*, and all may be defined in different ways.
- Using *matrices* is a convenient way to summarize these concepts as they apply to any particular distribution of geographic objects.
- *Proximity polygons* and their dual, the *Delaunay triangulation,* are useful constructions in geographic analysis.
- Spatial data really are special!

REFERENCES

Doll, R. (1955) Etiology of lung cancer. *Advances in Cancer Research*, 3: 1–50.

Freedman, D., Pisani, R., and Purves, R. (1998) *Statistics*, 3rd ed. (New York: W. W. Norton).

Gatrell, A. C. (1983) *Distance and Space: A Geographical Perspective* (Oxford: Oxford University Press).

Gehlke, C. E. and Biehl, K. (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29(185): 169–170.

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. (2000) *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd ed. (Chichester, England: Wiley).

Okabe, A., Boots, B., and Sugihara, K. (1994) Nearest neighborhood operations with generalized Voronoi diagrams: a review. *International Journal of Geographical Information Systems*, 8: 43–71.

Okabe, A., Satoh, T., Furuta, T., Suzuki, A., and Okano, K. (2008) Generalized network Voronoi diagrams: concepts, computational methods, and applications. *International Journal of Geographical Information Science*, 22: 965–994.

Openshaw, S. (1983) *The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography* 38, 41 pages (Norwich, England: Geo Books). Available at http://www.qmrg.org.uk/catmog.

Openshaw, S., and Taylor, P. J. (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem in Wrigley, N. (ed.), *Statistical Methods in the Spatial Sciences* (London: Pion), pp. 127–144.

Tobler, W. R. (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46: 234–240.

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 3

Fundamentals—Mapping It Out

CHAPTER OBJECTIVES

In this chapter, we:

- Describe the *changing role of maps* in geographic information analysis
- Justify the use of maps as analytical devices, equivalent to graphics such as histograms and boxplots used in conventional statistical description and analysis
- Define the so-called *graphic variables* and show how *geovisualization* by computer extends them in new ways
- Outline the major mapping options available to display points, areas, and fields
- Introduce the idea of *spatialization,* the production of map-like displays of nonspatial data

After reading this chapter and following the various boxed thought exercises, you should be able to:

- Justify why geographic information analysis should be concerned with maps and mapping
- Outline the major differences between maps drawn on paper and maps drawn on screen
- List and describe Bertin's original set of seven graphic variables and explain why they need to be qualified when we use them to describe ways of mapping
- List and describe various additional graphic variables introduced when we use modern computing hardware and software
- Select an appropriate map type for a given geographic phenomenon being displayed

- Outline the basic rationale and strategy involved in spatialization
- Above all else, look at any maps critically

This chapter contains very little direct advice on explicitly cartographic questions such as the choice of symbolism and color scheme, nor do we provide many illustrative examples; each of these issues is worth a textbook of its own. These issues are covered in the relevant textbooks, for example those by Dent (1990), Robinson et al. (1995), and Krygier and Wood (2005) and, perhaps more accessibly nowadays, on the World Wide Web using appropriate searches. In what follows, we provide pointers to relevant literature, as well as advice on the design and use of maps from the perspective of data exploration and analysis rather than of map design.

3.1. INTRODUCTION: THE CARTOGRAPHIC TRADITION

Maps are the most persuasive GIS output. This is something that GIS vendors recognize, and is obvious at exhibitions where many trade stands display maps produced using their systems. Less obvious, however, is that maps also play a major role in determining the nature of the inputs to GIS. Despite advances in the direct collection of spatial information using technologies such as the global positioning system (GPS), paper maps remain a source of data for GIS and, for better or worse, methods of analysis originally developed using maps have also affected much of the spatial analysis that we undertake with GIS. It follows that the traditions of cartography are of fundamental importance to GIS at data entry, in analysis, and in presentation (see Kraak, 2006).

The past few years have seen a number of popular accounts of specific maps (see, for example, Winchester, 2002; Foxell, 2008; Johnson, 2006; Schwartz, 2008), and as we write, it is obvious that societies worldwide are discovering maps and mapping through media such as in-car navigation, GPS with mapping capabilities, location-aware information served to cell (mobile) phones, and numerous Web sites that serve mapping customized to meet specific locational needs. Even more exciting has been the opening up of mapping capability so that almost anyone with an Internet connection can create personalized maps, perhaps as a way of cataloging photographs or recording a vacation trip. Hudson-Smith (2008) provides an easy-to-follow guide on how to create maps using the facilities and data available from *Google Maps*™ and *Google Earth*™. The term *neogeography* has been coined to refer to a whole raft of activities that have map creation as a central objective but owe little to past cartographic traditions and motivations. Rather than being based on the digital data provided by national mapping agencies, much neogeography is based on data provided by private individuals, called *crowd sourcing* or, more formally, *volunteered geographic information* and through data aggregation companies such as *Google*$^{\text{TM}}$. Paradoxically, while all this activity has been going on, and as Dodge and Perkins (2008) document, academic geographers do not seem to have taken much notice, with maps and mapping becoming less and less their concern.

Maps have been used for centuries as a data storage and access mechanism for topographic and cadastral (land ownership) information, and since the nineteenth century the thematic map has been used to display statistical data. These uses of maps for storage, access, and communication of results are well known and, by and large, are understood and accepted. Less widely accepted is the use of maps as a direct means of analysis where the act of display is itself an analytical strategy. If you are brought up in the Western scientific tradition, reliance on a display as a means of analysis can be difficult to accept. Although visualization has long been used as an informal route to understanding, graphical analysis has usually been subsidiary to mathematics and statistics. Although we say that we "see" a result, almost invariably the preferred form of analysis has been mathematical or statistical.

Suspicion of display as a form of analysis can be justified in three ways. First, *maps do not seem to compress data*. One way to define science is as a search for *algorithmic compressions*, that is, methods that lead to simplification by reducing the incoherence of vast arrays of information to explanations. From this viewpoint, like pictures, maps may well be worth a thousand words, but equations are even more compact descriptions and hence to be preferred. Indeed, the argument often made in favor of maps—that they *increase* the information available by showing spatial relationships and patterns—can be seen as running totally counter to the idea of algorithmic compression. Second, *maps give mixed messages*. Like natural language and figurative imagery, maps are a *polysemic* (many-signed) means of communication. In polysemic communication, the meaning of each symbol is deduced from observation of the *collection* of signs and thus is capable of many different interpretations. It follows that displays are almost always ambiguous and that cartography must concern itself with models of the communication process from data through cartographer to map user. Maps can fail to communicate what is intended or give a false impression. This is in contrast to the *monosemic* nature of mathematics, in which the meaning of symbols is rigorously defined in advance so that each symbol signifies just one thing: alternative interpretations are simply not allowed. Third, *maps are hard to draw*. Until recently, limited data availability and manual drafting methods using specialized tools meant that maps required skill to produce. This discouraged the use of display as a mode of data exploration because the costs in labor, time and money were prohibitive.

3.2. GEOVISUALIZATION AND ANALYSIS

The attitudes to display outlined above changed with the development and use of *scientific visualization*, defined as *exploring data and information graphically as a means of gaining understanding and insight* (Earnshaw and Wiseman, 1992). There are many reasons why visualization has become popular in all the sciences. First, developments in sensor technology and automated data capture have provided data at rates faster than they can be easily converted into knowledge. Second, some of the most exciting discoveries in science have been associated with *nonlinear dynamics*, or *chaos theory*, where apparently simple mathematical equations conceal enormously complex behaviors and structures that are most readily appreciated when displayed graphically. Third, as complex simulation models have become common scientific products, it has become necessary to use visualization as the only practical way to assimilate their outputs. A good example is the display of output from the atmospheric general circulation models used in the investigation of global warming. Last, but emphatically not least, improvements in computing mean that scientific visualization is now routinely possible using standard computing hardware on the desktop.

Visualization is in the tradition of *exploratory data analysis* in statistics. Its worth lies in its emphasis on the use of graphics in the *development of ideas*—not, as in traditional graphics, in their *presentation*. Indeed, visualization often turns traditional research procedures upside down by developing ideas graphically and then presenting them by nongraphic means. Modern visualization and traditional cartography have much in common, and it is hardly surprising that we now see studies that could be called cartography or visualization according to taste. Techniques borrowed from visualization have been used to improve on traditional map design, to visualize quantities such as the errors in interpolation and satellite image classification, and to create entirely new forms of display (Fisher et al., 1993). Since the use of graphics in science has been called *scientific visualization* (see, for example, Hearnshaw and Unwin, 1994), the term *geovisualization* has been coined to describe the fusion of visualization and cartography (see Dykes et al., 2005; Dodge et al., 2008).

The technical changes that underpin geovisualization don't simply mean that we draw our maps in different ways; they lead to a much more important set of changes in the ways that we design and use them. Instead of using map symbols to represent selected features, geovisualization frequently attempts

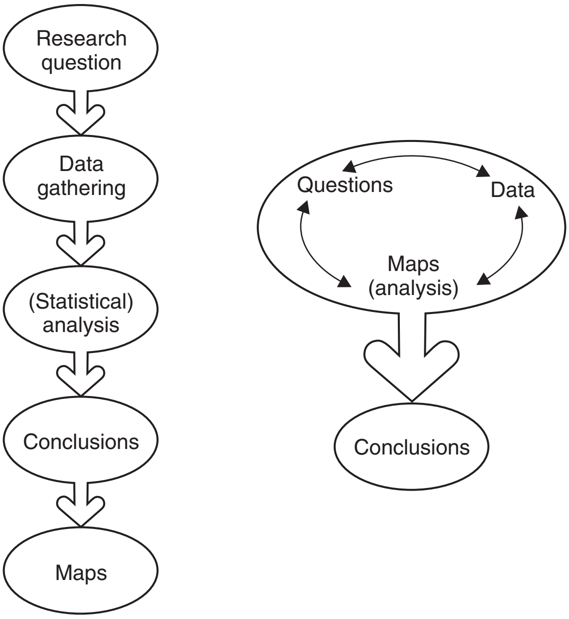


Figure 3.1 The changing role of maps in the analysis process.

to create photo-realistic scenes that display as much data as possible (Fisher and Unwin, 2002). In the world of geovisualization, maps are seldom completed end products of an investigation intended for a general audience. Instead, they are the means to an end and are most often viewed just once by a single person.

So, the main consequence of making maps easy to draw on screen is to extend their role so that they are now visualization and analysis tools. This view is illustrated in Figure 3.1. Traditional research plans are illustrated on the left-hand side, proceeding in a linear fashion from questions to data, analysis, and conclusions, where maps were an important presentation tool. The contemporary GIS research environment is more like that illustrated on the right. Data are readily available in the form of maps, prompting questions. Of course, the researcher may also come to a problem with a set of questions and begin looking for answers using available data by mapping them. Maps produced as intermediate products in this process (and not intended for publication) may prompt further questions and the search for more data. This complex and fluid process continues until useful conclusions are produced. Of course, the traditional approach was never as rigid or linear as portrayed here, and the contemporary approach may be more structured than this description suggests. The important point is that *maps have become tools of analysis*; they are no longer simply tools for the presentation of results.

Two related consequences of the changing role of maps have been the demise of cartography as a distinct discipline and the recognition that maps are just one form of display. Early progress in relating ideas from visualization to GIS and mapping can be found in Hearnshaw and Unwin (1994), MacEachren and Fraser Taylor (1994), and Unwin (1994). Now, in a world where almost everyone can have mapping capabilities on their desktop, it is worth emphasizing that computer scientists and others concerned with visualization also have much to learn from the accumulated wisdom of cartographers. Our reason, then, for devoting this chapter to maps and mapping should now be clear. Dynamic and interactive map displays are powerful analytical tools, but if we are to use them, it is important that the lessons learned by cartographers about making useful and meaningful map products are not forgotten. It is some of those lessons that we discuss in the remainder of this chapter.

3.3. THE GRAPHIC VARIABLES OF JACQUES BERTIN

Maps are drawn on flat sheets of paper, so the ways that a cartographer can display information are limited. An influential analysis of the toolbox of graphic techniques available for the display of information was presented by Jacques Bertin in his book *Semiologie Graphique*, originally published in French in 1967 and translated into English (and updated) by W. J. Berg as *Semiology of Graphics* (Bertin, 1983). *Semiology*, incidentally, means the study of signs, and Bertin tried to develop a theory of the signs used in all graphics. He recognized seven ways in which graphic signs may be varied to indicate different information. These have become known as his *graphic variables*, which are location, value, color/hue, size, shape, spacing/texture, and orientation:

- *Location*, where a symbol is placed on a map, is determined by geography and is the primary means of showing spatial relations. Although the property of location is straightforward for many types of graphic, in geography we must be careful. This is because changing the *map projection* used can change the relative locations of map symbols. It is well known that all projections must distort in some way: the important thing is to understand this point and work within the resulting limitations.

- *Value* refers to the lightness or darkness of a symbol. Typically, differences in the value of a symbol are used to represent differences in interval and ratio scaled variables. In paper mapping, the usual rule is that the darker a symbol, the higher the value it represents, although in computer cartography, where dark backgrounds are common, lighter (brighter) shading may indicate higher values. In fact, Bertin's analysis of value was simplistic, and it is now understood that the relationship between the value of a symbol and the numerical value it is thought to represent is not straightforward.
- *Hue*, or *color*, is an obvious graphic variable and, by contrast with value, it is usually used to represent qualitative variables on nominal or ordinal scales rather than quantitative differences. Color used to be expensive to reproduce on paper, so that maps intended for research publications were often designed to be printed in one color only. Moreover, a continuous grading of color was extremely difficult and expensive to create. Nowadays, color printing is less expensive, and almost all computer displays are capable of generating many more hues than our eyes can reliably distinguish, so color is now a widely used graphic variable.
- Color is also a much-abused graphic variable partly because it is extremely complex. There are at least four reasons for this. First, color theory demonstrates that there is more to color than just *hue*, which is the sensation of color as red, blue, green, and so on. In addition, we must consider a color's *value*—the sensation of lightness or darkness produced—and its *chroma*—its apparent intensity or brilliance. Second, the human eye-brain system does not see hues with equal sensitivity. Sensitivity varies, being highest for green, followed by red, yellow, blue, and purple. Third, colors have cultural associations that affect the way we read a map containing them. You can probably list numerous color associations, and skilful cartographers (not to mention advertisers) make clever use of them. Finally, the appearance of a color is not a simple function of its own hue, value, and chroma; it also depends on the size of area colored and on the surrounding colors. For example, a small area of a brilliant red may be visually acceptable where a large one is not, and any color tends to take on some of the appearance of its background. The golden rule for any use of color in mapping is to be careful, since color may create more problems than it solves. The temptation to create highly colored displays and assume that they will automatically help the visualization process should always be resisted.

Avoiding the Cartographic Quagmire?

Mark Monmonier's classic little book *How to Lie with Maps* (Monmonier, 1991) has a whole chapter titled ''Color: Attraction and Distraction'' that starts with a very direct message: ''Color is a cartographic quagmire'' (p. 147). All too often, we see on the World Wide Web and elsewhere wholly inappropriate and incorrect use of color in mapping.

An obvious solution is to keep it simple. Cindy Brewer and Mark Harrower have together produced ColorBrewer, a very useful online guide to color use on maps to be found at www.colorbrewer.org. It is worth spending time looking at the advice it gives.

- It is obvious that the *size* of a symbol can be used to show quantitative differences. It might be thought that a simple linear function could be used such that increases in the area of a symbol are proportional to the value represented. However, it has been shown that the brain has difficulty inferring quantity accurately from symbol size. We consider this issue in more detail in our discussion of proportional symbol maps in Section 3.6.
- *Shape*, the geometric form of a symbol, may be used to differentiate between different types of objects. Cartographers use this variable often—for example, in the way different types of building are represented, or on road maps, where shape is applied to line objects to distinguish different classes of highway.
- *Spacing*, the arrangement and/or density of symbols in a pattern, may also be used to show quantitative differences. In cartography, a simple example is the use of a pattern of dots to indicate the areal density of a phenomenon in a *dot density map*.
- Finally, the *orientation* of a pattern—of cross-hatching, for example—may be used to show qualitative difference.

Although Bertin's seven-variable scheme appears logical and all-embracing, research on how people perceive the graphic variables has shown that none are as simple as he suggests. A critical point is that each graphic variable should only be used to show types of variation for which it is suited. For example, hue works well if we use it to show differences in qualitative information, but attempts to use it for quantitative variation require great care. Conversely, although value and size work well for displaying quantitative information, they are not easily adapted to show qualitative differences.

A Summary Exercise

Two things you might like to think about here are:

1. In what ways do Bertin's graphics variables oversimplify?
2. Looking back at the distinctions between variables recorded using different levels of measurement (see Section 1.4), which of the graphic variables are best used for nominal, ordinal, and interval/ratio scaled data?

A useful way to address the second issue is to create a cross-tabulation of the seven graphic variables (rows) and the four levels of measurement (columns), noting for each cell whether or not the graphic variables can be used.

3.4. NEW GRAPHIC VARIABLES

Developments in computer technology allow us to create new forms of display that extend the number of graphic variables well beyond the original seven. These new cartographic variables include animation to create sequences of maps, creative use of map projection, the ability to link maps back to data, and the ability to link maps to other graphics.

Animation and Graphics Scripts

The use of animation to produce linked sequences of maps is not particularly new (see Tobler, 1970). However, until recently, producing map sequences was a considerable undertaking that involved laboriously drawing and photographing individual maps for movie projection. The availability of large volumes of digital data and fast computers with good graphics capabilities now enables maps to be animated with little difficulty. A sequence of maps can have time rescaled to display either more quickly (common) or less quickly (unusual) than the real-world phenomenon of interest. This is an obvious way to animate a map, but it is not the only one.

Interactive dynamic maps enable the focus of interest to be varied either spatially, by changing scale and location, with *zoom* and *pan* functionality, or statistically, by selecting subsets from statistical views of the same data. Animation may also use out-of-sequence orderings of time, or sequencing by the value of a selected variable or to emphasize the dynamic nature of a variable portrayed, as in maps of streamlines of a flow. You often see the wind arrows or storm systems on a TV weather map animated in this way.

An early example of transient private maps that embraced dynamic techniques was provided by Ferreira and Wiggins (1990). Their *density dial* allows users to slice through a variable's range at a chosen level and view the resulting classification on a map. Such a dial enables users to determine the sensitivity of a classification scheme. By interactively varying the scheme, they may determine the regions or levels at which a variety of patterns occur. Cartographic visualization researchers at The Pennsylvania State University have developed the use of animation that includes a whole series of new dynamic variables. These include the duration, rate of change, ordering, and phase (or rhythmic repetition of events) of a map sequence (DiBiase et al., 1992).

Although animation is attractive for detecting patterns at the first stages of an investigation, difficulties remain. First, although an investigator may gain insights from animation, it is difficult to publish such insights in the usual way. Instead, it is slowly becoming routine for visualization sequences to be made available as videos on the Web. Second, as with many new graphic variables, there are few established design rules for its effective use.

Linking and Brushing

Transient symbolism may be used to highlight symbols on a computer map display when they have been selected for investigation. This idea has been extended to the process of *brushing*, where corresponding symbols in two or more views are identified by similar distinct symbolism. Monmonier (1989) extended the technique by linking choropleth maps of area-valued data (see Section 3.7) with statistical plots of the same data, adding a geographic component to scatterplots, and vice versa. Just as transient maps eliminate the constraints associated with producing single representations, linking removes the restriction of having to use the spatial dimensions of the page to show geographic locations. Maps use the spatial arrangement of symbols on a display to reflect geographic relationships while simultaneously using the screen location on linked views to show statistical variation. This can be a powerful exploratory technique.

Projection

The projection used for a map is usually considered fixed and is often chosen from a limited number of standard forms, each suited to particular applications. It is also well understood that nonstandard forms of standard projections (such as the Mercator) can be excellent graphic devices in themselves, and that more obscure projections that are seldom drawn also have useful properties. Computerization enables maps to be drawn easily in almost any projection and so should encourage more innovative use of projection as an effective graphic variable. In fact, systematic, exploratory use of map projection in this way remains rare. In Section 3.7 we will deal with cartograms, which can be regarded as a particular—some would say peculiar—map projection.

3.5. ISSUES IN GEOVISUALIZATION

Geovisualization is not without problems. The change from maps on paper to maps on screen enables the analyst to use many new graphic variables, and the temptation is to create highly colorful, dynamic, and linked displays. A major problem is that we still know very little about good design using these variables. Sometimes they can be effective, but just because technology allows you to do something clever doesn't mean that you should! Similarly, just because paper and screen differ, this does not mean that the GI analyst should lightly discard the accumulated wisdom of well over a century and a half of thematic mapping.

Geovisualization is not a single analytical strategy. We can identify at least three different approaches, relating to the interplay between the data, their geography, and the technology used. The first is the pure *geovisualization* route, attempting to enable interactive exploration of the data using object linking, brushing, and so on, but by and large leaving the data intact. The second is the *spatial analytical* route, which modifies the numbers to be mapped mathematically—for example, by converting them into density estimates, probabilities against some hypothesized process, or the derivation of "local" statistics to isolate areas of specific research interest. Much of the remainder of this book develops this approach. The third approach involves some *transformation*, often by reprojecting the data into a space, such as an area cartogram, in which some aspects of geographic reality are more apparent. Currently, work tends to be channeled down one or another of these routes, but it is clear that most progress will be made by combining them. Mennis (2006) provides an example of careful visualization of the results of the local statistical operation known as *geographically weighted regression* (see Chapter 8). Similarly, a classical spatial analytical tool, the *Moran scatterplot* (Anselin, 1996), seldom makes much sense unless it is linked back to a choropleth map of the individual values.

As yet, there is little well-established theory to enable us to answer basic visualization questions such as "What works?", "Why does it work?" or even "What is likely to be the best way of displaying these data?" All we have are some moderately well-articulated design rules, some interesting speculation based in, for example, communication theory or semiotics, some results from usability and perception experiments, and appeals to our instincts. The result is that geovisualization sometimes seems to consist of finding new ways of turning one complex, hard-to-understand graphic and its associated data into another that is also complex and hard to understand. It may be that a basis for useful theory exists and that only the required synthesis is lacking, but it may also be that geovisualization cannot be formalized.

A final issue concerns the relationship of geovisualization to the derivation, testing, and presentation of theory. Although it is frequently claimed that visualization provides a way to develop theory, we doubt that theory generation using just graphics is possible. The interplay between graphics, theory, and prior domain knowledge is often more complex than geovisualizers recognize. There are many examples of geovisualization as a way to *test* existing hypotheses. This is the map as *proposition*. A simple example is provided by recent accounts of John Snow's iconic 1854 map of the cholera epidemic in Soho and its demonstration that a single polluted water supply pump was its cause, not the then popular notion of a "miasma" in the air. Statistical analysis has verified Snow's visual association (see Koch and Denke, 2004), and many authors cite Snow's map as a classic example of geovisualization yielding the hypothesis that cholera is water borne. The recent debate (see Brodie et al., 2000; Koch, 2004, 2005) makes it clear that Snow already had his hypothesis and that the map was a specific test of it against alternatives.

3.6. MAPPING AND EXPLORING POINTS

Dot or Pin Maps

The simplest map we can draw has a number of dots, one at each point where a specified object is located. In the terminology of Chapter 1, each dot represents a nominal-level attribute of an entity at a point location; in Bertin's typology (Section 3.3), we are using location as our graphic variable. If you are accustomed to using the Web and GIS, you will probably think of these maps as *pin maps*.

Creating a Pin Map Using Google Maps

Point your Web browser at www.google.com and click on the ''Maps'' option. If you live in a reasonably large city, enter text such as ''coffee shops in'' <name of your town>. (Failing this, have a look at the center of Northampton (England) by typing ''coffee shops in NN1''.) The result will be a pin map of coffee shops in the town concerned. Note that each pin has an exact one-to-one correspondence with an ''event'' (the existence of such a shop) and that there is also a link to additional data, such as the shop's name and other particulars. However, you may well find that the list is flawed in some way and that the locations at which the pins are placed are not exact. This is good enough, perhaps, for you to find a shop, but not for any formal statistical analysis of the pattern of locations revealed.

The theory of dotting is straightforward: simply place a point symbol at each location where an instance of the entity being mapped occurs. Because there is only one symbol for each entity, this is a *one-to-one mapping*, and the number of symbols is the same as the number of entities represented. The only design considerations are the shape, size, and hue of the symbol. The simplest symbol we can imagine is a small circular black dot, of a size large enough for dots to be individually visible on the final map but not so large that adjacent dots coalesce. The overall visual impression depends on relative areas of black and white, so that, ideally, as more dots are placed in an area, the result should seem denser in proportion. Unfortunately, experiment shows that we do not perceive dot density this way. Mackay (1949) suggested that at wide dot spacing changes in apparent density are rapid, but as more dots are added, the perceived change is less, until dot coalescence occurs when there is another large perceptual change.

In a simple dot map, the cartographer has no control over the number of dots. However, practicality often requires that each dot represent a defined number of objects, thus giving a *many-to-one mapping*. The result is a *dot density map*, and instead of location, we are now reliant on spacing as the graphic variable. The advantage of this technique is the control it gives over the overall visual impression. Unfortunately, there are no clear rules on the choice of the per-dot value. Some authors suggest that, for a given dot size, a value should be chosen so that dots coalesce only in the densest areas. Others argue that the original data should be recoverable from the map, implying that each dot must be distinct and countable. Dot density maps have the disadvantage that the location of dots is arbitrary. In the absence of other information about the distribution, the cartographer has little choice but to apply dots evenly over an area. When other information is available, dots may be located to reflect this. In a computer environment it would ideally be possible to experiment with dot size and value settings, but since no algorithm for satisfactory automated placement of dots has been developed, this is not usually possible.

Looking at Some Pin Maps

Go to Google (not Google Maps) and search for dot map. We are willing to bet that most of the images returned are dot density maps of the type described above. When we tried this, we got some fairly nice examples from the USDA 2002 Census of Agriculture, but the only true dot maps that had a one-to-one mapping with the events were for the locations of crime incidents in Sri Lanka! Searching for pin maps improves things a bit, but not much. Can you do any better?

Kernel Density Maps

A key property of any pattern of point-located events is its overall areal *density*, given by the number of point events per unit of area. As we will see in Chapter 5, in spatial analysis we prefer to think of this as an *estimate* of the *intensity* of the process, λ, given by

$$\hat{\lambda} = \frac{n}{a} = \frac{\#\,(S \in A)}{a} \tag{3.1}$$

where $\#(S \in A)$ is the number of events in pattern S found in study region A and a is the area of the region. This overall measure is often of only limited use; instead, many statistical techniques use some estimate of the *local* density of points. This idea underlies *kernel density estimation* (KDE) methods. The concept is that the pattern has a density at any location in the study region—not just at locations where there is an event. This density is estimated by counting the number of events in a region, or *kernel*, centered at the location where the estimate is to be made. In terms of the geovisualization strategies suggested in Section 3.2, KDE uses a transformation approach, in this case from point objects to a field of density estimates that can be easily visualized.

The simplest approach is to use a circle centered at the location for which a density estimate is required, count the number of point events falling into this circle, and divide by the circle's area. Then we have an intensity estimate at point $\mathbf{p}$

$$\hat{\lambda}_{\mathrm{p}} = \frac{\#\,(S \in C(\mathbf{p}, r))}{\pi r^2} \tag{3.2}$$

where $C(\mathbf{p},r)$ is a circle of radius r centered at the location of interest $\mathbf{p}$, as shown in Figure 3.2. If we make estimates for a series of locations throughout

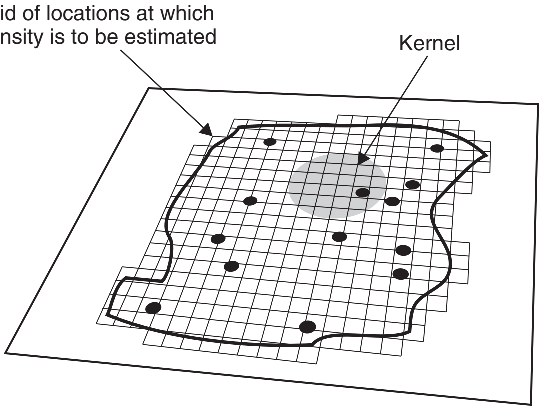


Figure 3.2 Simple, or naive, density estimation.

the study region, then it is possible to map the values produced, and this gives us an impression of the point pattern.

More sophisticated variations on the basic KDE idea make use of *kernel functions*, which weight nearby events more heavily than distant ones in estimating the local density. If the kernel functions are properly designed, KDE produces a surface that encloses a volume equivalent to the total number of events n in the pattern. A quartic kernel function, often used, is shown schematically in Figure 3.3.

Other functional forms, based on the distance of the point to be estimated from events in the pattern, are possible and are specified with a parameter that is equivalent to the simple bandwidth r. This means that the procedure is arbitrary to some degree, but distance-weighted kernel-fitting procedures ensure that the resulting surfaces of density estimates will be continuous. A typical output is shown in Figure 3.4. The resulting map is a surface, and

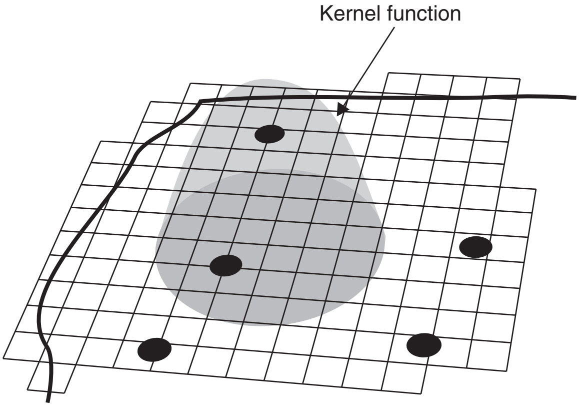


Figure 3.3 KDE using a quartic distance-weighted function.

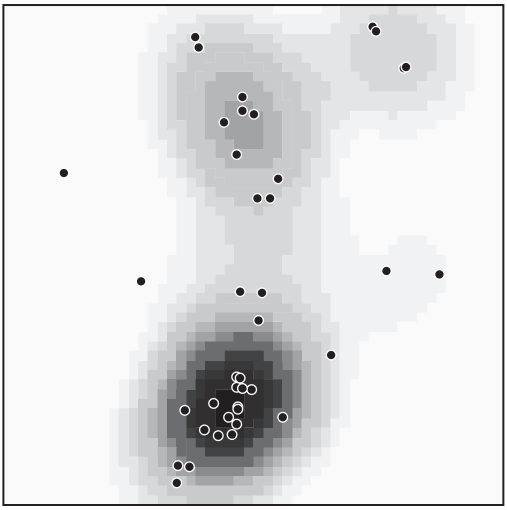


Figure 3.4 A typical output surface from KDE and its original point pattern.

contours can be drawn on it to give an indication of regions of high and low point density.

It should be clear that the choice of r, the so-called kernel *bandwidth*, strongly affects the resulting estimated density surface. If the bandwidth is large, then estimated densities $\hat{\lambda}_p$will be similar everywhere and close to the average density for the whole pattern. When the bandwidth is small, then the surface pattern will be strongly focused on individual events, with density estimates of zero in locations remote from any events. In practice, this problem is reduced by focusing on kernel bandwidths that have some meaning in the context of the study. For example, in examining point patterns of reported crime, we might use a bandwidth related to patrol vehicle response times. Generally, experimentation is required to arrive at a satisfactory density surface.

An important variant on KDE allows events in the pattern to be *counts* allocated to points. For example, points might correspond to places of employment, with associated counts of the number of employees. The resulting KDE surface shows "employment density" across the study region and may be a useful way of visualizing otherwise very complex distributional information. Some care is required in using this variation of the method to avoid confusion with *interpolation* techniques, which are discussed in Chapters 9 and 10.

The kernel density transformation is one of the most useful in applied GIS analysis. It provides a very good way to visualize a point pattern to detect "hot spots" where the local density is estimated to be high. In addition, it provides a good way of linking point objects to other geographic data. For example, if we have mortality data across a region as a set of point locations and want to relate these to a spatially continuous variable such as atmospheric pollution, a simple approach is to transform the mortality data on deaths into a density surface and compare this with a surface of atmospheric pollutant concentrations. Density estimation tools are provided in many commercial GISs, although often no detailed information is provided about the kernel functions used, so care should be exercised in using them. Public domain software to create surface estimates using a variable bandwidth that adapts automatically to the local density of observations has been published by Brunsdon (1995).

Located Proportional Symbol Maps

If we systematically vary the size of a point symbol, it is possible to show attribute data measured at ordinal, interval, and ratio levels on a *located proportional symbol map*. Examples might be the outputs of a series of factories, the number of people they employ, or the population of cities over a wide area. In each case, the data are attributes of point objects, and are not samples drawn from an underlying continuous field. The most common style of proportional symbol map employs circles whose areas are varied according to the value to be represented.

The simplicity of this approach hides troublesome technical details. If a map is intended to allow users to estimate numerical data based on symbol sizes, problems arise, and the choice of functional form is not straightforward. The human perceptual system does not register increases in circular area well. Most people underestimate the size of larger circles relative to small ones, so this approach tends to cause underestimation of larger values. Various solutions have been suggested (see Robinson et al., 1995).

Another point symbol map that makes use of geolocated proportional symbols is a *pie chart map*. Such maps show data that make up proportions of a whole—for example, the proportions of different products in the total output of a factory. Each pie symbol is scaled to reflect the total as before, but is subdivided into two or more differently shaded parts, one for each of the components. Almost invariably, sectors of a circle are used.

Before we leave the visualization of point event data, one further warning is in order. Frequently, all the devices we have discussed (dots, proportionate symbols, pie charts) are used to visualize data that are some form of areal aggregate, such as the total population of an area or its mix of industries. Such maps are perfectly legitimate, but from the viewpoint of geographic information analysis, they are not maps of point events; rather, they are maps that use similar symbolism to display area–value information. If you see such maps, then ask the two simple questions:

- Are the symbols placed at the exact locations of the point events, or are they at some arbitrary location within an area object?
- What is the relationship between the number of symbols and the number of events? Is it one to one, many to one, or one to many?

If the symbols are arbitrarily located and the relationship isn't one to one, then you are really dealing with some form of area object display.

Finding Some Examples

Visit your library or surf the Web to see if you can find examples of these types of point symbol maps:

- Dot density
- Proportional point symbols
- Pie chart maps

In each case, ask yourself the two questions that we ask above.

3.7. MAPPING AND EXPLORING AREAS

Color Patch Maps

Many different types of maps can be drawn to represent area data. The simplest is the color patch map, which has been give the fancy name of *chorochromatic* (*choro* = relating to area, *chroma* = relating to color), where graphic symbolism (not necessarily color) indicates the presence of a named attribute over a natural or imposed area. One might, for example, shade all the areas of the United Kingdom classified as urban in character, or all rock outcrops of Silurian age, or rainfall in excess of a specified threshold. The simplest maps of this type portray one nominal category in one color, giving a *two-phase mosaic* or *binary map*. Sometimes this kind of map is referred to as a *two-color map*, because one color is used to indicate the presence of an attribute and, by implication, the white areas are a second color, indicating where the attribute is not found. Chorochromatic maps can also use several colors to show a number of nominal classes simultaneously. In general, we talk of a *k-color map*, where *k* is the number of colors, and therefore also the number of nominal categories involved.

Some Color Patch Maps

Use Google Images^{TM} or a similar search engine to find examples of simple color patch maps. They are most likely to deal with phenomena such as geology and land cover. You might search for a copy of the map that Winchester (2002) claims ''changed the world,'' which is William Smith's superb first geology map dating from 1815. How many ''colors'' did Smith use? Having looked at Smith's map, next find a map of the political ''color'' of the U.S. states in the 2008 presidential election. In what important ways does this map differ from Smith's? There is an obvious and important difference in the way the areas are defined.

Color maps using natural areas are unlikely to be misinterpreted, but the same cannot be said for those based on data collected over a set of imposed areas. Without further information, all that the second type of map tells us is that the attribute is present somewhere in the area or, in aggregate, that the area is of the type implied. Problems may also occur if the nominal categories used, although mutually exclusive, are not spatially exclusive. On a map, the distributions will overlap. To solve this problem, several solutions can be adopted. An obvious one is to choose and map only those categories that are spatially exclusive; others are to intermingle sets of point symbols, to use special symbols or colors for the mixed distributions, or, even more simply, to draw a separate map for each category.

Choropleth Maps

The second major type of map for area objects is probably the most widely used, and also the most misunderstood and incorrectly produced of all the map types discussed in this chapter. This is a map that displays interval or ratio scaled attribute data collected over imposed or, less commonly, natural areas. Such maps are called *choropleths*, or area-value maps (*choro* = relating to area, *pleth* = relating to value). Because geographic data, especially from census sources, are almost always agglomerated to areal units, choropleth maps are common, and the technique is available in virtually all GISs.

Figure 3.5 provides an example. It attempts to show the density of population across approximately 250 census area units in the Auckland region of New Zealand. The map has several elements with two distinct types of data. First, there are the actual head counts of people living in each area, except those shown in outline only where no data have been mapped. Second,

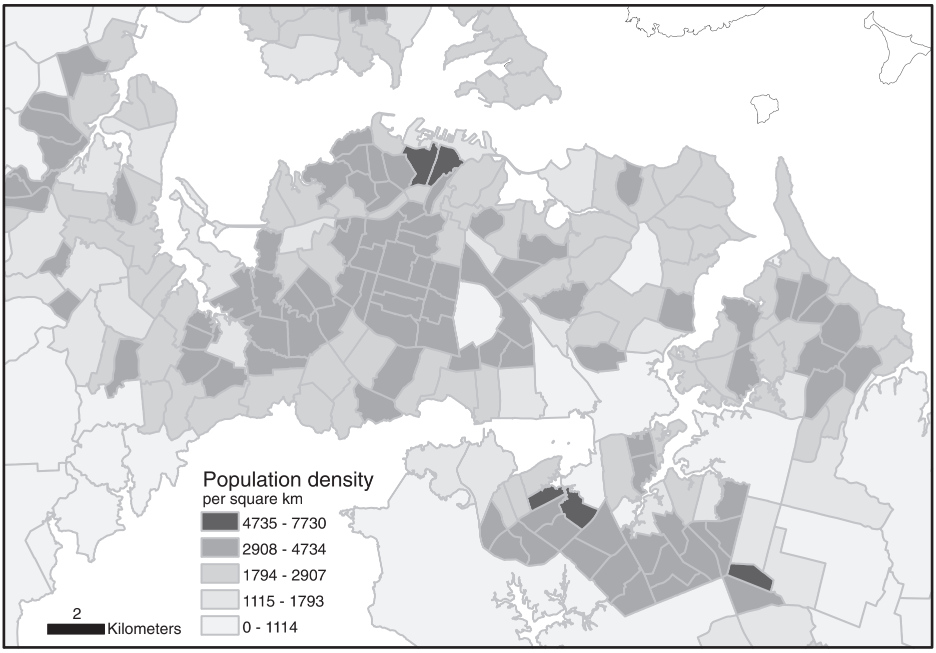


Figure 3.5 A choropleth map showing, the 2006 density of population over census area units of the Auckland region of New Zealand.

there are the data that describe the outlines of the areas over which the counts have been aggregated. These have been used to find the area, based on which population densities have been calculated. To create the map, the density values have been classified into five bands and a shade style has been assigned to each, from light (lowest population density) to dark (highest).

Statisticians will recognize that what we have created is in some respects a two-dimensional version of a histogram in which the individual small areas have the same role as the *bins* (classes) and the mapped density values are similar to the heights of the histogram bars. Choropleth maps are honest in the sense that they are true to the data, but they can be very poor representations of the underlying geographic patterns they purport to represent. Thinking of them as two-dimensional versions of the standard histogram helps us see why. In statistics, histograms are used to describe the distribution of sample data and to provide a graphical estimate of an unknown underlying probability density function. In such work, there are two sources of variation—the value range along the x-axis and the frequencies shown by the bar heights on the y-axis—and so the two dimensions provided by a sheet of paper or a screen are sufficient. When compiling a histogram, the analyst has complete control over the size of the *bins* used and can show each height exactly using just the two dimensions given by a sheet of paper or a screen.

Neither is possible when drawing a choropleth, which has unequally sized zones over which the analyst has no control and whose heights have to be classified in some way, such that symbolism can be used to show the third dimension. It follows that choropleth maps can often give a poor visualization of any underlying patterns:

- *The areas used.* Are these natural or imposed? If the former, how were they defined and by whom? If the latter, are the areas used appropriate? Do large areas dominate the way the map looks? Are the "steps" at the edges of each block of color likely to reflect variation in the underlying phenomenon?
- *The data.* Are they *counts* of some sort? If so, there is a built-in tendency for larger areas to have bigger values, and the map may be worthless. Choropleth maps make sense only if the numbers being mapped are ratios, either areal densities (such as the number of people per unit of area) or population rates (such as the number of births per thousand population in the area). If the data are ratios, are these based on low numbers? If so, the mapped values may be very unstable to small changes. If we add one person to an area where there is already just one person, we double the population density, whereas adding one person to several thousand people makes virtually no difference at all.
- *The classification used.* Prior to the use of computers, almost all choropleth maps were "classed," that is, each data value was assigned to one of a small number of classes. Experience suggests that five to seven classes is appropriate, but it is easy to show that the appearance of a map can be changed dramatically by varying the number of classes. Figure 3.6 illustrates the effect. These maps show the percentage of

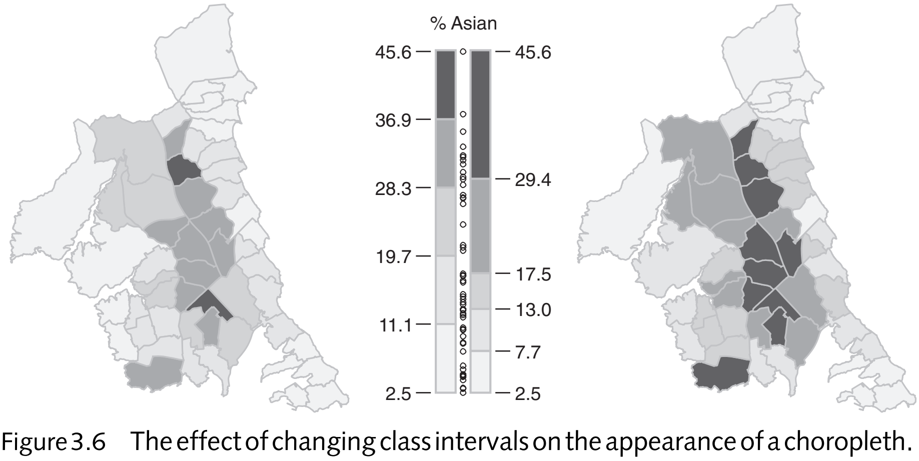


Figure 3.6 The effect of changing class intervals on the appearance of a choropleth.

population of Asian ethnicity in 53 census area units in North Shore City, Auckland, in 2006. The left-hand map uses five equal intervals, while the right-hand one uses five "quantiles." In a classic paper, Evans (1977) describes a large number of possible schemes. His conclusion is that you must examine the statistical frequency distribution of the data before deciding on a classification scheme.

- *The symbolism used.* This is the most obvious attribute of a choropleth. Traditionally, choropleths were created by shading each area using a "screen" pattern of lines such that the more lines there were, the darker the area looked and the higher the value. It is now more usual to use gradations of a single color to show the increase in intensity. Either way, how you choose to shade the map greatly affects its final appearance.

These issues are discussed and illustrated in more detail in Dykes and Unwin (2001), which shows how the visual appearance of any choropleth is extremely sensitive to the choices made in its construction.

Some Choropleth Maps

Use Google Images or a similar search engine to find examples of choropleth color maps. In each case, make a summary commentary on:

- The suitability of the areas used
- The appropriateness of the data used (for example, are they ratios or absolute numbers?)
- The classification scheme used, if any
- The method by which these numbers are displayed
- The overall effectiveness of the map

We suspect that after completing this exercise, you might well agree with one of our students, who said that he would ''never look at this sort of map again without taking a large pinch of salt.''

One result of this sensitivity of the appearance of a choropleth map to the choices made in its construction has been a number of experiments by statisticians and cartographers to develop alternatives. The list of suggestions is long. The strategies adopted have used both geovisualization and spatial analysis as a way of improving on the basic map type.

Classless Choropleths

Many years ago, Tobler pointed out that modern displays don't actually need to class choropleths since they are capable of showing as many shades of color as needed (Tobler, 1973). His point was debated at the time (see Dobson, 1973), but nowadays it is relatively common to see *classless choropleths*, and the approach is available as an option in most GISs. Although this approach has some attractive features, it is by no means a solution to the problem of choosing appropriate class intervals, because the human visual system is poor at judging exact color shades relative to a range of possible values. This makes it hard for a map reader to estimate values in particular units in a classless map.

Maps of Relative Rates

Another enhancement of choropleth maps adopts a different strategy, modifying the numbers to be mapped rather than the cartography. It is common, particularly in epidemiology, to encounter data sets for which many of the area counts are very small numbers, many of them zeros. The frequency distribution of such data is far from normal, making selection of class intervals difficult, and computed rates will be unstable to small changes in the data, leading to absurd maps when the counts in some zones are close to zero. A simple way to cope with the presence of many zeros due to Cressie and his co-workers (Cressie, 1993) is to add 1 to each value before calculating the rates of incidence relative to the area population totals. This has the effect of discriminating (slightly) among the zero-valued areas, but the problem of instability remains. An alternative is to map area scores relative to some assumed distribution. In epidemiology, use is often made of standardized mortality ratios, which are the ratios of the deaths in each zone relative to those expected on the basis of some externally specified (typically national) age/sex-specific rates. Almost any approach along similar lines will produce more sensible choropleth maps. For example, in their census atlas of population, the Census Research Unit (CRU, 1980) mapped a series of variables using what they called the *signed chi-square statistic*. This was defined as the square of the difference between the actual number in each zone and the number expected under the assumption of an evenly distributed population, divided by the expected value itself, much as in the conventional chi-square statistic. Use of the square required each value to have its positive (more than expected) or negative (less than expected) signs added after the calculation, with the mapped values displayed using a bipolar scale centered on zero. A simpler alternative, illustrated by Dykes and Unwin (2001), is to use the simpler square root relation $(O_i - E_i)/\sqrt{E_i}$, which automatically takes care of the sign. Another approach, suggested by Choynowski (1959), is to map the probabilities of getting values more or less extreme than those observed, assuming that the underlying distribution is Poisson. This type of map has been frequently used in medical geography and spatial epidemiology. Finally, several authors have suggested that it is possible to take a Bayesian approach to the problem, adjusting estimated ratios in each zone either away from or toward an overall global value for the rate according to some prior measure of confidence in the ratios (see Marshall, 1991; Langford, 1994).

Dasymetric Mapping

There have been a number of attempts to use transformation of the base map to produce better visualizations. If, for example, the data to be visualized relate to some aspect of the population and the zones come from some census subdivision, then it makes sense in each zone to exclude any part of the land that does not have any residential housing, such as parks, water bodies, and commercial premises, when calculating areal density estimates. Such maps are called *dasymetric*, and the approach has been suggested many times. Usually, authors attribute the idea to Wright (1936), but it was also suggested in Russia in the 1920s (Fabrikant, 2003). In his study, Wright used a manual analysis of a standard topographic map to define the settled area, but within a GIS it is relatively easy to use remote sensed imagery to obtain the same information (Langford and Unwin, 1994; Mennis, 2003).

Surface Models for Area Objects

A second alternative that also transforms the areal base is to estimate a continuous surface of rates from the irregular pattern of zones that make up the area objects and to visualize this as a surface display. As spatial data have become available at increasingly high spatial resolution, the size of the zones used has become smaller, making it possible to use the basic choropleth data in much the same way as a statistician would use a histogram, as estimates of some underlying continuous field of spatial densities using either a variant of KDE (Thurstain-Goodwin and Unwin, 2000; Donnay and Unwin, 2001) or interpolation onto a fine raster (Martin, 1989). Once such a surface transformation has been obtained, it is possible to explore the data further using standard surface processing (Dykes et al., 1997).

Area Cartograms

A final approach to improving the visualization of area-aggregated data is to reproject the data onto an area cartogram base. This is a radical but justifiable solution to the choropleth problem that has been used for many years (see one of the geography literature's all-time classics: Tobler, 1963). On an area cartogram, the zones themselves are drawn such that the area of each is set proportionate to some other variable—typically, but not always, the population of the zone. Meeting area cartograms for the first time, people often think that they are "funny-looking maps" and, perhaps for that reason, find them difficult to "read", but they have considerable use in geographic analysis. As extensive reviews by their major current exponents show (Dorling, 1996; Tobler, 2004), they have been drawn for many years, and if we view them from the perspective of our third visualization strategy, as transformations, they are not as unusual as people sometimes think.

Getting the Idea

A very comprehensive set of cartogram world maps of single variables produced by Daniel Dorling's team can be found at www.worldmapper.org/. It is interesting to see how cartograms change one's view of the planet. Cartograms of world population show the importance of India and China, but the remaining displays for a whole series of variables associated with social and economic matters highlight global inequalities in a way that conventional choropleth mapping of the same data cannot. These maps are also available as an atlas (Dorling et al., 2008). Alternatively, you can use your favorite search engine to find some of the many cartograms produced after the 2008 U.S. presidential election.

Cartograms are a form of map projection. What an area cartogram does is systematically to expand/contract areas locally as some function of another variable for every zone on the map. There is no single or simple way of effecting an area cartogram transformation, although there is a long history of attempts (for reviews, see Dorling, 1994; Tobler, 2004). Early attempts used a variety of mechanical methods (see Hunter and Young, 1968; Skoda and Robertson, 1972) and the results were not often used, but by the 1980s, work was well underway to develop computer algorithms capable of producing useful results. Probably the most popular algorithms currently in use are those by Dorling (1992, 1995) and, more recently, those by Gastner and Newman (2004), but others have been reported (Gusein-Zade and Tikunov, 1993; Keim et al., 2004).

3.8. MAPPING AND EXPLORING FIELDS

Point Values: Spot Heights, Benchmarks, and Bubble Plots

Spatially continuous fields can be mapped in various ways. The simplest method is to plot actual data values at a number of points. The points chosen could be significant ones on the surface, such as peaks and valleys, the result of a random sampling, or values on a systematic grid. Topographic maps often show *spot heights* as point symbols (usually a dot or small open circle), with the value written alongside. Both location and value are displayed accurately, but there is no entity on the ground as on the map. In contrast, the *benchmarks* and *triangulation points* also shown on topographic maps are also usually marked on the ground in some way. Such point height information has the advantages of accuracy and honesty. Only those data that are known are displayed, and the map user is not offered any additional interpretation of them. The major disadvantage is that no impression of the overall shape of the field—its spatial structure—is given, but such an overall impression depends upon some form of *interpolation* of the data values to create a complete surface model. That said, a clear advance on marked spot heights using basic visualization as its strategy is their display using symbols varying in size or color according to the field value located at each data point location. These plots are called *bubble plots* and are very useful as a first look at some surface data.

Contours and Isolines

If we are prepared to use interpolation to create a model of the surface, then numerous display options are available. The *isoline*, of which the familiar relief *contour* is the most familiar example, is the standard way to represent a continuous field of data. In isoline mapping, we make an imaginary connection of all locations in the field that are of equal height (z) value to form a three-dimensional curve. These curves are then projected onto a two-dimensional surface, usually (but not always, see below) the x–y plane, to produce an isoline or contour map. Isolines show the absolute value of the field and, by their spacing, also provide information on its gradient. The resulting map is a view from an infinite distance above the field, that is, a perpendicular or *orthogonal* view with no perspective effects. The biggest difficulty with isoline or contour mapping is that we need to know a great deal about variation in the values of the underlying phenomenon and, through the method of interpolation used, we are departing from the honesty of a simple map of spot heights. Numerous interpolation methods are available, which we discuss in Chapters 9 and 10, and most GISs have an isoline mapping capability.

Usually isolines are depicted as fine, continuous lines of appropriate color, broken in places to allow labeling. To aid interpretation, every fourth or fifth line may be thickened to act as a marker. The most important factor governing the appearance of the map is the number of isolines and hence the interval used. A large number of isolines gives a detailed picture of the field but may obscure other map details. A small number requires less data and will not mask other detail but may give a poor sense of the spatial structure. Deciding on the isoline interval is a problem similar to that of choosing class intervals for choropleth maps. However, if isolines are to show surface slope by their proximity to each other, there must be a standard, equal interval. The choice of the interval is not simple; it depends on the scale and use of the map and also on the nature of the surface being mapped. For example, the current Ordnance Survey of Great Britain 1:50,000 sheets have an interval of 10 m, which is very much a compromise. In mountainous areas, such as the Scottish Highlands, it produces a clear representation of the relief and leads to excessive crowding of contours only on the steepest slopes. The same 10-m interval applied to the flattest areas of the English lowlands often fails to pick up significant features in the landscape. This is a general problem, because it is possible for significant features to be "filtered out" by being located entirely within a single isoline interval. Familiarity with contour maps can lead us to forget this unfortunate property of isolines and also to underestimate the difficulty of interpreting contour maps for the less experienced user.

Looking at Contours

Look at an example of the topographic mapping provided by your national mapping agency (NMA) at a scale of around 1:50,000. Do these maps have contours and, if so, how are the contour lines represented? If your library resources allow, it is instructive to compare this map with a product of a different NMA.

Enhancing the Isoline

Spot heights and isolines are often supplemented by other methods to improve the overall visual impression of variation across a field. In *layer coloring*, the field value range is divided into a series of bands, and areas in each band are shaded an appropriate color. In a raster GIS data structure, where a field is recorded over a regular grid of values, layer coloring is easy to apply. Topographic maps are often layer colored, using a sequence from green (lowland) through yellow and brown to blues and whites (high mountains). Similarly, many newspapers color temperature bands in weather maps from blue (cold) to red (hot). The major drawback of this method is a tendency to generate an impression of steps or sharp boundaries in field values; it is imperative that the range of colors chosen show a gradation in apparent intensity. Again, inspection of maps in the media and on the Web shows that this is often not the case.

Another way to represent field data concentrates attention on the shape deduced from *gradients* and introduces the idea of a *vector field*. A *scalar* is a quantity characterized solely by its magnitude, which remains unchanged no matter how we project the data. A *vector* quantity is specified by both a magnitude and a direction, and so depends also on how the data are projected. Examples of vectors include the wind, migration flows, the flow of water in a river, and so on. Just as it is possible to examine scalar fields, so we can have vector fields, that is, continuous fields of measurable vectors changing from point to point in space. *Every scalar field has a vector field associated with it given by its gradient or slope*. We can plot the magnitude and direction of the gradient as a single downslope arrow. A map of a scalar field's associated vector field provides a reasonably good visualization of the structure of the scalar field and was the essence of *hachures* used in early relief maps, but these are rarely seen today. Nowadays, vector fields may be displayed using arrows, each with its head giving the direction and length of the magnitude, or using two maps, one for the down-gradient magnitude, the other for its direction.

Another supplementary technique, often used on relief maps but rarely on other isoline maps, is *hill shading*. A continuous shading variation from light to dark is used to indicate surface slope, with darker shades corresponding to steeper slopes. In a GIS, the required intensity of shading can be calculated from a grid of field values and may be stored as a new field, to be draped over other displays (such as a contour map). Relief is presented as if seen under illumination from a distant light source, the precise effect varying according to the source position. If this position is vertically above the surface, no shadows are created, but the amount of illumination varies in relation to the slope. Many maps use an oblique source in the northwest, which gives light northwest-facing and dark southeast-facing slopes, so that with the map viewed with north "at the top," the shadows run toward the viewer. This can give a strong visual impression of relief but is not without problems. First, whether or not a slope is in shadow depends as much on its aspect (direction) as on its angle of slope. For example, a gentle southeast-facing slope may be illuminated as if it were a steeper one facing in some other direction. Second, for reasons that are poorly understood, viewing the map with the shadows running away from the viewer produces a startling and potentially very confusing relief inversion: valleys become hills and vice versa.

Of our three visualization strategies, hill shading is effectively the second option, transforming the field values in some way so as better to reveal the *shape* of the surface rather than its height. In conventional hill shading, we compute the first derivative of height in order to generate the gradient data, which are then displayed. This approach can be taken further by computing the curvature of the surface, which is the rate of change of the gradient. In a series of papers, Wood and his collaborators have used maps based on the local curvature of surfaces of relief for the objective identification of landscape features such as peaks, passes, and valleys (see, for example, Fisher et al., 2004). Wood et al. (1999) show that such analysis can usefully be applied to socioeconomic data such as population density.

Given technological developments that enable the creation of sophisticated surface displays whose production could not be contemplated otherwise, the distinction between what is clearly and unequivocally a map and what is more properly thought of as a virtual reality display has become very blurred.

Other Ways of Displaying Surfaces

A useful distinction that can be made when considering maps of surface data is whether or not they are *planimetrically correct*. A planimetrically correct representation is one where the viewpoint is from infinity, with the surface relief projected onto a plane at right angles to the viewing direction. Such vertical views preserve plan distance and do not hide parts of the surface from view. Traditional isoline maps conform to this approach.

Many other methods of displaying relief are not planimetrically correct. In fact, a planimetrically correct perspective is a highly artificial view, one that humans never experience in reality. We look across landscapes, or down onto them from a single point. It might be inferred that our ability to visualize the relief of a surface would be improved if it were displayed using a more natural viewpoint that incorporated perspective effects. Many GISs can plot surfaces in three-dimensional projections. In its simplest form, two sets of intersecting parallel lines trace out elevation or some other quantitative attribute. Known as a *fishnet*, this form of rendering is closer to what we might regard as a real view of topography. It does not suffer from the quantization of elevation into bands, as happens with contours. Three-dimensional projections can be enhanced by *draping* images of other attribute values over the surface. The drape may take the form of the same elevation information or additional related information, such as the surface slope or average rainfall. Overlaying images with some degree of realism, such as shaded relief, or remotely sensed colour composites, exploits the human tendency to make sense of visual images and provides a *virtual reality* view of the land surface (Fisher and Unwin, 2002). However, the information conveyed by such images can be difficult to control, and much remains to be learned about design rules for these displays. Arguably, they introduce yet more graphic variables for consideration, such as vertical exaggeration, line frequency, sequencing, and viewing direction.

Landserf for Surface Display

Most GISs have sophisticated surface display options, but Landserf is a public-domain system designed specifically for the visualization and analysis of surfaces by Jo Wood of London's City University. Landserf can be accessed at www.landserf.org.

The ''image gallery'' at this Web site has examples of numerous techniques for the display of surface information, including three-dimensional perspective views, gradient and curvature mapping, ''fly throughs'' synthesized from digital elevation data, and so on. Visit the site and examine each of the images in this gallery.

3.9. THE SPATIALIZATION OF NONSPATIAL DATA

In everyday life, we often use the words *map* and *mapping* as a metaphor for an organizational framework for some nongeographic objects of interest. We talk, for example, of “maps” of the brain, of DNA sequences, of galaxies, and even of the computer hard drive (see Hall, 1992). In this usage, the term *map* is often used to mean that the information is organized such that each object of interest can be located by coordinates (x, y) in two dimensions. If these happen to be spatial coordinates, the result is a geographic map, but clearly, the same general approach can used with any other variables and the result will be a map in some other space. The visualization gain is that we can comprehend the relationships between objects by simple visual examination of our map. Whether or not this is sensible, or leads to better understanding, is a matter for the scientific discipline concerned, but we refer to this display of data in two dimensions as the general process of *spatialization* of nonspatial data.

Some Examples of Spatialization

The best way to understand spatialization is to examine some examples. Doing a Web search for spatialization will bring up a lot of examples, such as:

1. ''Genomes as Geography.'' This article by Dolan et al. (2006) uses the mapping and search tools developed within a GIS to display genomes spatially in a system called GenoSIS.
2. Indexing a book. In an interesting collaboration, Dykes, Fabrikant and Wood (2005) have produced a surface map display based on spatialization of the papers in an edited volume on geovisualization (Dykes, MacEachren, and Kraak, 2005); see www.soi.city.ac.uk/~jwo/landserf/gallery/image4.html.
3. Organizing an information space such as your hard drive directory.
4. One of the most prolific researchers into the issues around spatialization has been Sarah Fabrikant. Some of her work can be seen at www.geog.ucsb.edu/~sara/html/research/diss/spatialization.html.

Without worrying about map projection issues, when we plot objects on a geographic map, we use Cartesian coordinates with two axes, each of which is a distance, north/south (*northings*) and east/west (*eastings*). These distances are real ones that we can measure.

The key to understanding spatialization is to retain the same notion of locating objects by distances in a Cartesian system but to relax our definition of distance. By far the best basic reference on these ideas is the book by Gatrell (1983), which deals with numerous concepts of distance as an example of a *relation between elements of a defined set*. We might, for example, measure the distance between two places by the time taken to travel between them or the total travel cost involved. Using these definitions of distance greatly alters our view of the planet: it is much quicker (but rather more expensive) for one of us to get from his home in England to New York than it is to get to the Isle of Skye off the west coast of Scotland, but the real-world distance to New York is much greater than it is to Skye. The "time-distance" is less, and the "cost-distance" is more, but which is analytically more useful?

We can push this analysis further by thinking of distance as similarity/dissimilarity (see Fabrikant et al., 2004)—for example, between counts of plant species in a series of ecological samples, the number of shared keywords in some document, and even the number of times one player passes the ball to another in a game of soccer (Gatrell and Gould, 1979). In a multivariate data space, these similarities could also be the standard Pearsonian correlation coefficients between the variables. The relationship is that of statistical correlation, and the distance is in correlation space.

The problem in spatialization is to take a matrix of observed similarities/dissimilarities and display the information in a two- or three-dimensional Cartesian space that can be represented by a standard map type. Space does not allow us to go into detail on this issue but, as Skupin and Fabrikant (2003) show, workers have experimented with virtually every standard data reduction technique. Since we are projecting the data into Cartesian coordinates, the two requirements are that the axes be at right angles to each other (orthogonal) and that, for the resulting map to be useful, the data can be reduced to just two dimensions. In his book, Gatrell (1983) used metric- and nonmetric-multidimensional scaling to find the best configuration for the distances in two dimensions. A simpler alternative that is sometimes appropriate is to use standard principal components analysis employing scores on the first two components as the locational coordinates. More recently, use has been made of a variety of other data reduction techniques such as "projection pursuit," "spring modelling," "pathfinder network scaling," "self-organizing maps," "tree maps," and so on (Skupin and Fabrikant, 2003). In practice, most workers in the field seem to use either the method with which they are most familiar or the one that gives the clearest arrangement on the output map.

3.10. CONCLUSION

This chapter has covered a lot of ground. Our aim has been to show how allegedly simple mapping can help us understand spatial data and to place this within a modern framework in which access to a GIS is assumed. The use of a GIS to create maps easily and quickly from digital data means that much of the craft involved in drawing maps has been consigned to history, and the temptation is to think that GIS analysis no longer needs to concern itself with ideas from traditional cartography. Our essential summary point is that replacing a craft skill with a computer does not mean that the art of map design and the science of map compilation are also dead. Rather, it is important that everybody who produces maps or performs analysis with GIS be aware of the complexities and difficulties involved. It may no longer be important to follow all the rules of cartographic design all of the time when using a map as a transient analysis tool rather than as a presentation medium, but understanding some of those fundamentals is still useful. Knowing when and why you are breaking the rules and how it can affect your understanding of the problem at hand is a valuable skill to develop. Equally, understanding how to make good maps improves your ability to use mapping as an essential tool for the exploration and understanding of your data.

CHAPTER REVIEW

- Maps are a very old form of display that are used in many ways.
- Exploring spatial data by visualization isn't new, but by providing more data, making maps easier to draw, and allowing new display methods, technological changes have made it a much more common analysis strategy.
- Traditionally, mapmakers were restricted to the use of a few basic graphic variables, such as the seven (location, value, hue, size, shape, spacing, pattern) suggested by Bertin.
- New technologies have enabled this list to be extended to include, for example, animation, linking, and brushing, but as yet, little is known about the ways in which map readers interpret and use such devices.
- For each of the spatial entity types (point, line, area, field) recognized in Chapter 1, there is a variety of possible mapping types, of which most GISs implement only a small subset.
- *Spatialization* is the name given to the process of making maps of nonspatial data.
- Finally, and despite all the technology, modern GIS users can learn a great deal about how to create effective displays from traditional cartography.

REFERENCES

Consistent with the view expressed in the last bullet point above, this chapter has a longer than usual reference list. We encourage those of you new to the art and science of cartography to explore some of this literature before you use any maps.

Anselin, L. (1996) The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: M. Fischer, H. J. Scholten, and D. Unwin, eds., *Spatial Analytical Perspectives on GIS*, (London: Taylor & Francis): pp. 111–125.

Bertin, J. (1983) *Semiology of Graphics* (Madison: University of Wisconsin Press).

Brody, H., RussellPip, M., Vinten-Johnasen, P., Paneth, N., and Rachman, S. (2000) Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *The Lancet*, 356(9223): 64–68.

Brunsdon, C. (1995) Estimating probability surfaces for geographical point data: an adaptive technique. *Computers and Geosciences*, 21: 877–894.

Choynowski, M. (1959) Maps based on probabilities. *Journal of the American Statistical Association*, 54: 385–388.

Cressie, N. A. (1993) *Statistics for Spatial Data*, rev. ed. (Hoboken, NJ: Wiley), pp. 385–393.

CRU, Census Research Unit. (1980) *People in Britain—A Census Atlas* (London: Her Majesty's Stationery Office).

Dent, B. D. (1990) *Cartography: Thematic Map Design* (DuBuque, IA: WCB Publishers).

DiBiase, D., MacEachren, A. M., Krygier, J., and Reeves, C. (1992) Animation and the role of map design in scientific visualization. *Cartography and Geographic Information Systems*, 19: 201–214.

Dobson, M. W. (1973) Choropleth maps without class intervals? A comment. *Geographical Analysis*, 5: 358–360.

Dodge, M., McDerby, M., and Turner, M., eds. (2008) *Geographic Visualization* (Chichester, England: Wiley).

Dodge, M., and Perkins, C. (2008) Reclaiming the map: British geography and ambivalent cartographic practice. *Environment and Planning A*, 40: 1271–1276.

Dolan, M. E., Holden, C. C., Beard, M. K., and Bult, C. J. (2006) Genomes as geography: using GIS technology to build interactive genome feature maps. *BMC Bioinformatics*, 7: 416 (available at www.biomedcentral.com/1471-2105/7/416).

Donnay, J. P. and Unwin, D. J. (2001) Modelling geographical distributions in urban areas. In: J. P. Donnay, M. J. Barnsley, and P. A. Longley, eds., *Remote Sensing and Urban Analysis, GISDATA 9* (London: Taylor & Francis), pp. 205–224.

Dorling, D. F. L. (1992) Visualizing people in time and space. *Environment and Planning B: Planning and Design*, 19: 613–637.

Dorling, D. (1994) Cartograms for visualizing human geography. In: H. Hearnshaw, and D. Unwin, eds., *Visualization in Geographical Information Systems* (Chichester, England: Wiley), pp. 85–102.

Dorling, D. F. L. (1995) *A New Social Atlas of Britain* (Chichester, England: Wiley).

Dorling, D. (1996) *Area cartograms: their use and creation. Concepts and Techniques in Modern Geography*, 59, 69 pages (Norwich, England: Geo Books). Available at http://www.qmrg.org.uk/catmog.

Dorling, D., Newman, M., and Barford, A. (2008) *The Atlas of the Real World* (London: Thames and Hudson) (also available at http://www.worldmapper.org).

Dykes, J. A., Fisher, P. F., Stynes, K., Unwin, D., and Wood, J. (1997) The use of the landscape metaphor in understanding population data. *Environment and Planning, B: Planning and Design*, 26: 281–295.

Dykes, J. A., MacEachren, A. M., and Kraak, M.-J., eds., (2005) *Exploring Geovisualization* (Amsterdam: Elsevier).

Dykes, J. and Unwin, D. J. (2001) *Maps of the Census: A Rough Guide* (available at http://www.agocg.ac.uk/reports/visual/casestud/dykes/dykes.pdf).

Earnshaw, R. A. and Wiseman, N. (1992) *An Introduction to Scientific Visualization* (Berlin: Springer-Verlag).

Evans, I. S. (1977) The selection of class intervals. *Transactions of the Institute of British Geographers*, Vol. 2, 98–124.

Fabrikant, S. I. (2003) Commentary on "A History of Twentieth-Century American Academic Cartography" by Robert McMaster and Susanna McMaster. *Cartography and Geographic Information Science*, 30: 81–84.

Fabrikant, S. I., Montello, D. R., Ruocco, M., and Middleton, R. S. (2004) The distance-similarity metaphor in network-display spatializations. *Cartography and Geographic Information Science*, 31: 237–252.

Ferreira, J., Jr. and Wiggins, L. L. (1990) The density dial: a visualization tool for thematic mapping. *Geo Info Systems*, 10: 69–71.

Fisher, P. F., Dykes, J., and Wood, J. (1993) Map design and visualisation. *The Cartographic Journal*, 30: 36–42.

Fisher, P. F. and Unwin, D. J., eds. (2002) *Virtual Reality in Geography* (London: Taylor & Francis).

Fisher, P., Wood, J., and Cheng, T. (2004) Where is Helvellyn? Fuzziness of multiscale landscape morphometry. *Transactions of the Institute of British Geographers*, 29: 106–128.

Foxell, S. (2008) *Mapping England* (London: Black Dog Publishing).

Gastner, M. T. and Newman, M. E. J. (2004) Diffusion-based method for produing density equalizing maps. *Proceedings of the National Academy of Science, USA*, 101: 7499–7504.

Gatrell, A. C. (1983) *Distance and Space: A Geographical Perspective* (Oxford: Clarendon Press).

Gatrell, A. C, and Gould, P. (1979) A micro-geography of team games: graphical explorations of structural relations. *Area* 11: 275–278.

Gusein-Zade, S. M. and Tikunov, V. (1993) A new technique for constructing continuous cartograms. *Cartography and Geographic Information Systems*, 20: 167–173.

Hall, S. S. (1992) *Mapping the Next Millennium: The Discovery of New Geographies* (New York: Random House).

Hearnshaw, H. and Unwin, D. J., eds. (1994) *Visualisation and GIS* (London: Wiley).

Hudson-Smith, A. (2008) *Digital Geography: Geographic Visualisation for Urban Environments* (London: UCL/CASA).

Hunter, J. M. and Young, J. C. (1968) A technique for the construction of quantitative cartograms by physical accretion models. *Professional Geographer*, 20: 402–407.

Johnson, S. (2006) *The Ghost Map* (New York: Riverhead; London: Penguin Books).

Keim, D., North, S., and Panse, C. (2004) CartoDraw: a fast algorithm for generating contiguous cartograms. *IEEE Transactions on Visualization and Computer Graphics*, 10: 95–110.

Kennelly, P. J. and Kimerling, A. J. (2001) Hillshading alternatives: new tools produce classic cartographic effects. *ArcUser* (available at http://www.esri.com/news/arcuser/0701/althillshade.html.

Koch, T. (2004) The map as intent: variations on the theme of John Snow *Cartographica*, 39: 1–13.

Koch, T. (2005) *Cartographies of Disease: Maps, Mapping and Medicine* (Redlands, CA: ESRI Press).

Koch, T. and Denke, K. (2004) Medical mapping: the revolution in teaching—and using—maps for the analysis of medical issues. *Journal of Geography*, 103: 76–85.

Kraak, M.-J. (2006) Why maps matter in GIScience. *The Cartographic Journal*, 43: 82–89.

Krygier, J. and Wood, D. (2005) *Making Maps: A Visual Guide to Map Design for GIS* (New York: Guildford Press).

Langford, I. (1994) Using empirical Bayes estimates in the geographical analysis of disease risk. *Area*, 26: 142–190.

Langford, M. and Unwin, D. J. (1994) Generating and mapping population density surfaces within a geographical information system. *The Cartographic Journal*, 31: 21–26.

MacEachren, A. M. and Fraser Taylor, D. R., eds. (1994) *Visualisation in Modern Cartography* (Oxford: Pergamon Press).

Mackay, J. R. (1949) Dotting the dot map. *Surveying and Mapping*, 9: 3–10.

Marshall, R. J. (1991) Mapping disease and mortality rates using empirical Bayes estimators. *Applied Statistics*, 40: 283–294.

Martin, D. (1989) Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers*, 14: 90–97.

Mennis, J. (2003) Generating surface models of population using dasymetric mapping. *Professional Geographer*, 55(1): 31–42.

Mennis, J. (2006) Mapping the results of geographically weighted regression. *Cartographic Journal*, 43(2): 171–179.

Monmonier, M. (1989) Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis*, 21: 81–84.

Monmonier, M. (1991) *How to Lie with Maps* (Chicago: University of Chicago Press).

Robinson, A. H., Morrison, J. L., Muehrcke, P. C., Kimerling, A. J., and Guptill, S. C. (1995) *Elements of Cartography*, 6th ed. (London: Wiley).

Schwartz, S. (2008) *The Mismapping of America* (Rochester, NY: University of Rochester Press).

Skoda, L. and Robertson, J. C. (1972) *Isodemographic Map of Canada* (Ottawa: Lands Directorate, Department of Environment, Geographical Papers, 50). See also http://www.csiss.org/classics/content/27.

Skupin, A. and Fabrikant, S. I. (2003) Spatialization methods: a cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science*, 30: 95–119.

Thurstain-Goodwin, M. and Unwin, D. J. (2000) Defining and delimiting the central areas of towns for statistical monitoring using continuous surface representations. *Transactions in GIS*, 4: 305–317.

Tobler, W. R. (1963) Geographic area and map projections. *Geographical Review*, 53: 59–78.

Tobler, W. R. (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46: 234–240.

Tobler, W. R. (1973) Choropleth maps without class intervals. *Geographical Analysis*, 5: 26–28.

Tobler, W. (2004) Thirty-five years of computer cartograms. *Annals of the Association of American Geographers*, 94: 1–58.

Unwin, D. J. (1994) Visualization, GIS and cartography. *Progress in Human Geography*, 18: 516–522.

Winchester, S. (2002) *The Map That Changed the World* (London: Penguin Books).

Wood, J. D., Fisher, P. F., Dykes, J. A., Unwin, D. J., and Stynes, K. (1999) The use of the landscape metaphor in understanding population data. *Environment and Planning B: Planning and Design*, 26: 281–295.

Wright, J. K. (1936) A method of mapping densities of population with Cape Cod as an example. *Geographical Review*, 26: 103–110.

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 4

Fundamentals—Maps as Outcomes of Processes

CHAPTER OBJECTIVES

In this chapter, we:

- Introduce the concept of *patterns* as *realizations* of *processes*
- Describe a simple process model for point patterns–the *independent random process* or *complete spatial randomness*
- Show how *expected values* for one measure of a point pattern can be derived from this process model
- Introduce the ideas of *stationarity* and of *first-* and *second-order* effects in spatial processes
- Differentiate between *isotropic* and *anisotropic* processes
- Briefly extend these ideas to the treatment of line and area objects and to spatially continuous fields

After reading this chapter, you should be able to:

- Justify the so-called *stochastic process* approach to spatial statistical analysis
- Describe and provide examples of *deterministic* and spatial *stochastic processes*
- List the two basic assumptions of the independent random process
- Outline the logic behind the derivation of long-run expected outcomes of this process using the quadrat counts as an example
- List and give examples of nonstationarity involving first- and second-order effects
- Differentiate between isotropic and anisotropic processes

- Outline how these ideas might also be applied to line, area, and field objects

4.1. INTRODUCTION: MAPS AND PROCESSES

In Chapter 1 we highlighted the importance of spatial *patterns* and spatial *processes* to the view of spatial analysis presented in this book. Patterns provide clues to a possible causal process. The continued usefulness of maps and other visualizations to analysts remains their ability to suggest patterns in the phenomena they represent. In this chapter, we look at this idea more closely and explain the view that maps can be understood as outcomes of processes.

At the moment, your picture of processes and patterns, as described in this book, may look something like that shown in Figure 4.1.

We would agree that this is not a very useful picture. In this chapter, we plan to develop your ideas about processes in spatial analysis so that the left-hand side of this picture becomes more complete. In the next chapter we develop ideas about patterns—filling in the right-hand side of the picture—and complete the picture by describing how processes and patterns may be related statistically. However, by the end of this chapter, you should already have a pretty good idea of where this discussion is going, because in practice, it is difficult to separate entirely these related concepts. We develop this discussion with particular reference to point pattern analysis, so by the time you have read these two chapters, you will be well on the way to an understanding of both general concepts in spatial analysis and more particular concepts relating to the analysis of point objects.

In Section 4.2 we define processes, starting with deterministic processes and moving on to stochastic processes. We focus on the idea that processes make patterns. In Section 4.3 we show how this idea can be made mathematically exact and that certain properties of the patterns produced by the independent random process can be predicted. This involves some mathematical derivation, but it is done in easy steps so that it is easy to follow. It is more important that you grasp the general principle that we can propose a mathematical model for a spatial process and then use that model to determine expected values for descriptive measures of the patterns that might result from that process. This provides a basis for the statistical

Processes	**Patterns**
?	**?**

Figure 4.1 Our current view of spatial statistical analysis. In this chapter and the next, we will be fleshing out this rather thin description.

assessment of the various point pattern measures discussed in Chapter 5. This chapter ends with a discussion of how this definition of a process can be extended to line, area, and field objects.

4.2. PROCESSES AND THE PATTERNS THEY MAKE

We have already seen that there are a number of technical problems in applying statistical analysis to spatial data—principally spatial autocorrelation, MAUP, and scale and edge effects. There is another, perhaps more troublesome problem, which seems to make the application of inferential statistics to geography at best questionable and at worst simply wrong: *geographic data are often not samples in the sense meant in standard statistics*. Frequently, geographic data represent the whole population. Often, we are only interested in understanding the study region, and not in making wider inferences about the whole world, so the data *are* the entire population of interest. For example, census data are usually available for a whole country. It would be perverse to study only the census data for the Eastern Seaboard if our interest extended to all of the lower 48 states of the United States, since data are available for all states. Therefore, we really don't need the whole apparatus of confidence intervals for the sample mean. If we want to determine the infant mortality rate for the lower 48 states, based on data for approximately 3000 counties, then we can simply calculate it, because we have all the data we need.

One response to this problem is not to try to say anything statistical about geographic data at all. Thus, we can describe and map geographic data without commenting on their likelihood, or on the confidence that we have a good estimate of their mean, or anything else. This is a perfectly reasonable approach. It certainly avoids the contradictions inherent in statements like "The mean Pennsylvania county population is 150,000 ± 15,000 with 95% confidence" when we have access to the full data set.

The other possibility is to think in terms of spatial *processes* and their possible *realizations*. In this view, an observed map pattern is *one of the possible patterns that might have been generated by a hypothesized process*. Statistical analysis, then, focuses on issues around the question "Could the pattern we observe have been generated by this particular process?"

Deterministic Processes

Process is one of those words that is tricky to pin down. Dictionary definitions tend to be unhelpful and a little banal: "something going on" is typical. Our definition will not be very helpful either, but bear with us, and it will all start

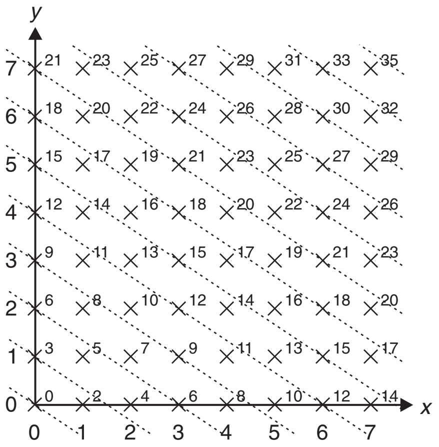


Figure 4.2 A realization of the deterministic spatial process $z = 2x + 3y$ for $0 \leq x \leq 7, 0 \leq y \leq 7$. Contours are shown as dashed lines. This is the only possible realization because the process is deterministic.

to make sense. *A spatial process is a description of how a spatial pattern might be generated*. Often, the process description is mathematical and it may also be *deterministic*. For example, if x and y are the two spatial coordinates, the equation

$$z = 2x + 3y \tag{4.1}$$

describes a *spatial* process that produces a numerical value for z at every location in the x–y plane. If we substitute any pair of location coordinates into this equation, then a value for z is returned. For example, location (3, 4) has $x = 3$ and $y = 4$, so that $z = (2 \times 3) + (3 \times 4) = 6 + 12 = 18$. The values of z at a number of other locations are shown in Figure 4.2. In the terms introduced in Chapters 1 and 2, the entity described by this equation is a spatially continuous field. The contours in the figure show that the field z is a simple inclined plane rising from southwest to northeast across the mapped area.

This spatial process is not very interesting because it always produces the same outcome at each location, which is what is meant by the term *deterministic*. The value of z at location (3, 4) will be 18 no matter how many times this process is realized or "made real."

A Stochastic Process and Its Realizations

Geographic data are rarely deterministic in this way. More often, they appear to be the result of a chance process, whose outcome is subject to variation that cannot be given precisely by a mathematical function. This apparently chance element seems inherent in processes involving the individual or collective results of human decisions. It also appears in applications such as meteorology, where, although the spatial patterns observed are the result of deterministic physical laws, they are often analyzed as if they were the results of chance processes. The physics of chaotic and complex systems has made it clear that even deterministic processes can produce seemingly random, unpredictable outcomes—see James Gleick's excellent nontechnical book *Chaos* for a thorough discussion (Gleick, 1987). Furthermore, the impossibility of exact measurement may introduce random errors into even uniquely determined spatial patterns. Whatever the reason for this chance variation, the result is that the same process may generate many different results.

If we introduce a random, or *stochastic*, element into a process description, then it becomes unpredictable. For example, a process similar to the previous one is $z = 2x + 3y + d$, where d is a randomly chosen value at each location (say) -1 or $+1$. Now different outcomes are possible each time the process is realized. Two realizations of

$$z = 2x + 3y \pm 1 \tag{4.2}$$

are shown in Figure 4.3. If you draw the same isolines, you will discover that, although there is still a general rise from southwest to northeast, the lines are no longer straight (try it). There is an effectively infinite number of possible realizations of this process. If only the 64 locations shown here are of

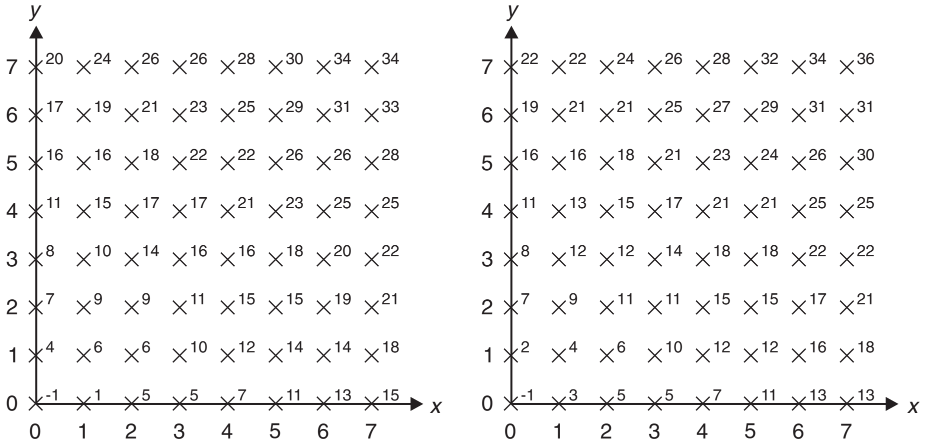


Figure 4.3 Two realizations of the stochastic spatial process $z = 2x + 3y \pm 1$ for $0 \le x \le 7, 0 \le y \le 7$.

interest, there are 2^{64} or 18,446,744,073,709,551,616 possible realizations that might be observed.

Thought Exercise to Fix Ideas

We concede that this is tedious, and if you understand things so far, then skip it. However, it is a useful exercise to fix ideas.

Use the basic equation above, but instead of adding or subtracting 1 from each value, randomly add or subtract an integer (whole number) in the range 0–9 and prepare an isoline map of the result you obtain. You can get random numbers from a spreadsheet or from tables in most statistics textbooks.Take two digits at a time. If the first digit is less than 5 (0–4), add the next digit to your result; if it is 5 or more, subtract the next digit.

Notice that this map pattern isn't random. The map still shows a general drift in that values increase from southwest to northeast, but it has a local chance component added to it. The word *random* refers to the way this second component was produced—in other words, it refers to the process, not to any resulting map.

What would be the outcome if there were absolutely no geography to a process—if it were completely random? If you think about it, the idea of no geography is the ultimate null hypothesis for any geographer to suggest, and we illustrate what it implies in the remainder of this section using as an example the creation of a *dot/pin map* created by a point process. Again, to fix ideas, we suggest that you undertake the following thought exercise.

All the Way: A Chance Map

The principles involved here can be demonstrated readily by the following experiment. If you have a spreadsheet on your computer with the ability to generate random numbers, it is easily done automatically. (Work out how for yourself!). By hand, proceed as follows:

1. Draw a square map frame, with eastings and northings coordinates from 0 to 99.
2. Use a spreadsheet program, random number tables, or the last two digits in a column of numbers in your telephone directory to get two random numbers each in the range 0–99.

3. Using these random numbers as the eastings and northings coordinates, mark a dot at the specified location.
4. Repeat steps 2 and 3 as many times as seems reasonable (50?) to get your first map.

To get another map, repeat steps 1–4.

The result is a dot/pin map generated by the *independent random process* (IRP), sometimes also called *complete spatial randomness* (CSR). Every time you locate a point, called an *event* in the language of statistics, you are randomly choosing a sample value from a fixed underlying probability distribution in which every whole number value in the range 0–99 has an equal chance of being selected. This is a uniform probability distribution. It should be evident that, although the process is the same each time, very different-looking maps can be produced. Each map is a *realization of a process* involving random selection from a fixed, uniform probability distribution. Strictly speaking, because in our exercise events can only occur at $100 \times 100 = 10,000$ locations and not absolutely everywhere in the study, the example isn't fully IRP/CSR. This issue can be easily addressed in a spreadsheet setting by generating real-valued random coordinates rather than integers.

It is important to be clear on three issues:

- The word *random* is used to describe the method by which the symbols are located, not the patterns that result. It is the process that is random, not the pattern. We can also generate maps of realizations randomly using other underlying probability distributions—not just uniform probabilities.

Different Distributions

If instead of selecting the two locational coordinates from a uniform probability distribution you had instead used a normal (Gaussian) distribution, how might the resulting realizations differ from the one you obtained?

Notice that the very clear tendency to create a pattern in this experiment is still a result of a random or stochastic process. It's just that, in this case, we chose different rules of the game.

- The maps produced by the stochastic processes we are discussing each display a spatial pattern. It often comes as a surprise to people doing these exercises for the first time that random selection from a uniform probability distribution can give marked clusters of events of the sort often seen, for example, in dot/pin maps of disease incidence.
- In no sense is it asserted that spatial patterns are ultimately chance affairs. In the real world, each point symbol on a map, whether it represents the incidence of a crime, illness, factory, or oak tree, has a good behavioral or environmental reason for its location. All we are saying is that, in aggregate, the many individual histories and circumstances might best be described by regarding the location process as a chance one—albeit a chance process with well-defined mechanisms.

4.3. PREDICTING THE PATTERN GENERATED BY A PROCESS

Warning: Mathematics Ahead!

So far, you may be thinking, ''This spatial statistical analysis is great—no statistics or mathematics!'' Well, all good things come to an end, and this section is where we start to look at the patterns in maps in a more formal or mathematical way. There is some possibly muddy mathematical ground ahead. As when dealing with really muddy ground, we will do better and not get stuck if we take our time and move slowly but surely ahead. The objective is not to bog down in mathematics (don't panic if you can't follow it completely), but rather to show that it is possible to suggest a process and then to use some mathematics to deduce its long-run average outcomes.

Now we will use the example of the dot/pin map produced by a point process to show how, with some basic assumptions and a little mathematics, we can deduce something about the patterns that result from a process. Of the infinitely many processes that could generate point symbol maps, the simplest is one where no spatial constraints operate, the IRP or CSR. You will already have a good idea of how this works if you completed the exercise in the previous section. Formally, the IRP postulates two conditions:

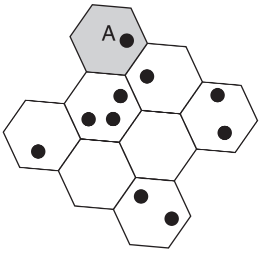


Figure 4.4 Quadrat counting for the example explained in the text.

1. The condition of *equal probability*. This states that any event has an equal probability of being in any position or, equivalently, that each small subarea of the map has an equal chance of receiving an event.
2. The condition of *independence*. This states that the positioning of any event is independent of the positioning of any other event.

Such a process might be appropriate in real-world situations where the locations of entities are not influenced either by the varying quality of the environment or by the distances between entities.

It turns out to be easy to derive the long-run expected results for this process, expressed in terms of the number of events we expect to find in a set of equal-sized and nonoverlapping areas, called *quadrats*. Figure 4.4 shows an area in which there are 10 events (points), distributed over eight hexagonal quadrats.

In the figure, a so-called *quadrat count* (see Chapter 5 for a more complete discussion) reveals that we have two quadrats with no events, three quadrats with one, two quadrats with two, and one quadrat with three events.

Our aim is to derive the *expected frequency distribution* of these numbers for the IRP outlined above. With our study region divided into these eight quadrats for quadrat counting, what is the probability that any one event will be found in a particular quadrat? Or two events? Or three? Obviously, this must depend on the number of events in the pattern. In our example there are 10 events in the pattern, and we are interested in determining the probabilities of 0, 1, 2 . . . up to 10 events being found in a particular quadrat. It is obvious that, under our assumptions, the chance that all 10 events will be in the same quadrat is very low, whereas the chance of getting just 1 event in a quadrat is relatively high.

To determine this expected frequency distribution, we need to build up the mathematics in a series of steps. First, we need to know the probability that *any* single event will occur in a *particular* quadrat. For each event in the pattern, the probability that it occurs in the particular quadrat we are looking at (say, the shaded one) is given by the fraction of the study area that the quadrat represents. This probability is given by

$$P(\text{event A in shaded quadrat}) = \frac{1}{8} \tag{4.3}$$

since all quadrats are equal in size and all eight together fill up the study region. This is a direct consequence of our assumption that an event has an equal probability of occurring anywhere in the study region and amounts to a declaration that there are no first-order effects in the imagined process.

Now, to the second step. For a particular event A to be the *only* event observed in the same *particular* quadrat, what must happen is that A is in that quadrat (with probability 1/8) and nine other events B, C, . . . J are not in the quadrat, which occurs with probability 7/8 for each of them. So, the probability that A is the only event in the quadrat is given by

$$P(\text{event A only}) = \frac{1}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \tag{4.4}$$

that is, 1/8, multiplied by 7/8 nine times—once for each of the events that we are not interested in seeing in the quadrat. The multiplication of the probabilities in the above equation is possible because of the second assumption—that each event location is independent of all other event locations—and is a declaration that there are no second-order effects in the imagined process.. Step three is as follows: that if we observe one event in a particular quadrat, it could be *any of the 10 events* in the pattern, not necessarily event A, so there are 10 ways of getting just one event in that quadrat. Thus, we have

$$P(\text{one event only}) = 10 \times \frac{1}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \times \frac{7}{8} \tag{4.5}$$

In fact, the general formula for the probability of observing k events in a particular quadrat is

$$P(k\text{ events}) = (\text{No. of possible combinations of } k \text{ events}) \times \left(\frac{1}{8}\right)^k \times \left(\frac{7}{8}\right)^{10-k} \tag{4.6}$$

The formula for "number of possible combinations of k events" from a set of n events is well known and is given by

$$C_k^n = \frac{n!}{k!(n-k)!} = \binom{n}{k} \tag{4.7}$$

where the exclamation symbol (!) represents the factorial operation and $n!$ is given by

$$n \times (n-1) \times (n-2) \ldots \times 1 \tag{4.8}$$

If we put this expression for the number of combinations of k events into equation (4.6), we have

$$\begin{aligned} P(k \text{ events}) &= C_k^{10} \times \left(\frac{1}{8}\right)^k \times \left(\frac{7}{8}\right)^{10-k} \\ &= \frac{10!}{k!(10-k)!} \times \left(\frac{1}{8}\right)^k \times \left(\frac{7}{8}\right)^{10-k} \end{aligned} \tag{4.9}$$

We can now substitute each possible value of k from 0 to 10 into this equation in turn and arrive at the probability distribution for the quadrat counts based on eight quadrats for a point pattern of 10 events. The probabilities that result are shown in Table 4.1.

This distribution is so commonplace in statistics that it has a name: the *binomial distribution*, given by

Table 4.1 Probability Distribution Calculations for the Worked Example in the Text, $n = 10$

No. of events in quadrant k	*No. of possible combinations of k events C_k^n*	$\left(\frac{1}{8}\right)^k$	$\left(\frac{7}{8}\right)^{10-k}$	*P(k events)*
0	1	1.00000000	0.26307558	0.26307558
1	10	0.12500000	0.30065780	0.37582225
2	45	0.01562500	0.34360892	0.24160002
3	120	0.00195313	0.39269590	0.09203810
4	210	0.00024412	0.44879532	0.02300953
5	252	0.00003052	0.51290894	0.00394449
6	210	0.00000381	0.58618164	0.00046958
7	120	0.00000048	0.66992188	0.00003833
8	45	0.00000006	0.76562500	0.00000205
9	10	0.00000001	0.87500000	0.00000007
10	1	0.00000000	1.00000000	0.00000000

$$P(n,k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.10}$$

A little thought will show that the probability p in the quadrat counting case is given by the size of each quadrat relative to the size of the study region. That is,

$$p = \frac{\text{quadrat area}}{\text{area of study region}} = \frac{a/x}{a} = \frac{1}{x} \tag{4.11}$$

where x is the number of quadrats into which the study area is divided. This gives us the final expression for the probability distribution of the quadrat counts for a point pattern generated by the IRP:

$$P(k,n,x) = \binom{n}{k} \left(\frac{1}{x}\right)^k \left(\frac{x-1}{x}\right)^{n-k} \tag{4.12}$$

which is simply a binomial distribution with $p = 1/x$, where n is the number of events in the pattern, x is the number of quadrats used, and k is the number of events in a quadrat.

The importance of these results cannot be overstated. In effect, we have specified a process—the IRP—and used some mathematics to predict the frequency distribution of quadrat counts that, in the long run, its realizations should yield. These probabilities may therefore be used as a standard by which any observed real-world distribution can be judged. For example, the small point pattern in Figure 4.4 has an observed quadrat count distribution shown in column 2 of Table 4.2.

We can compare this observed distribution of quadrat counts to that predicted by the binomial distribution calculations from Table 4.1. To make comparison easier, these proportions have been added as the last column in Table 4.2. The observed proportions appear very similar to those we would expect if the point pattern in Figure 4.4 had been produced by the IRP. This is confirmed by inspection of the two distributions plotted on the same axes, as in Figure 4.5.

Since we also know the theoretical mean and standard deviation of the binomial distribution, it is possible—as we shall see in the next chapter—to make this observation more precise using the usual statistical reasoning and tests.

In this section, we have seen that it is possible to describe a spatial process mathematically. We have also seen, by way of example, that we can predict

Table 4.2 Quadrat Counts for the Example in Figure 4.4 Compared to the Calculated Expected Frequency Distribution from the Binomial Distributions

k	*No. of quadrats*	*Observed proportions*	*Predicted proportions*
0	2	0.250	0.2630755
1	3	0.375	0.3758222
2	2	0.250	0.2416000
3	1	0.125	0.0920381
4	0	0.000	0.0230095
5	0	0.000	0.0039445
6	0	0.000	0.0004696
7	0	0.000	0.0000383
8	0	0.000	0.0000021
9	0	0.000	0.0000001
10	0	0.000	0.0000000

the outcome of a quadrat count description of a pattern generated by the IRP, and use this to judge whether or not a particular observed point pattern is unusual with respect to that process. In other words, we can form a null hypothesis that the IRP is responsible for an observed spatial pattern and judge whether or not the observed pattern is a likely realization of that process. In the next chapter, we discuss some statistical tests, based on this general approach, for various point pattern measures. This discussion should make the rather abstract ideas presented here more concrete.

We should note at this point that the binomial expression derived above is often not very practical. The calculation of the required factorials for even

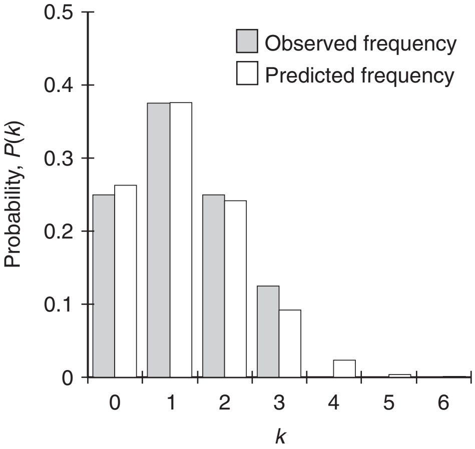


Figure 4.5 Comparison of the observed and predicted frequency distributions for the pattern in Figure 4.4.

Table 4.3 Comparison of the Binomial and Poisson Distributions for Small n

k	*Binomial*	*Poisson*
0	0.26307558	0.28650480
1	0.37582225	0.35813100
2	0.24160002	0.22383187
3	0.09203810	0.09326328
4	0.02300953	0.02914478
5	0.00394449	0.00728619
6	0.00046958	0.00151796
7	0.00003833	0.00027106
8	0.00000205	0.00004235
9	0.00000007	0.00000588
10	0.00000000	0.00000074

medium-sized values of n and k is difficult. For example, $50! \approx 3.0414 \times 10^{64}$ and $n = 50$ would represent a small point pattern—values of n of 1000 or more are not uncommon. Fortunately, it turns out that even for modest values of n the *Poisson distribution* is a very good approximation to the binomial distribution. The Poisson distribution is given by

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{4.13}$$

where λ is the total *intensity* of the pattern per quadrat and $e \approx 2.7182818$ is the base of the natural logarithm system. To confirm that this is a good approximation, for the example considered in Figure 4.4, if each hexagonal quadrat has unit area (i.e., 1), then $\lambda = 10/8 = 1.25$, and we obtain the proportions given in Table 4.3.

For larger n the Poisson approximation is closer than this, so it is almost always adequate—and it is always considerably easier to calculate.

4.4. MORE DEFINITIONS

The IRP is mathematically elegant and forms a useful starting point for spatial analysis, but its use is often exceedingly naive and unrealistic. Many applications of the model are made in the expectation of being forced to reject the null hypothesis of independence and randomness in favor of some alternative hypothesis that postulates a spatially dependent process. If real-world spatial patterns were indeed generated by unconstrained randomness, then geography as we understand it would have little meaning or interest and most GIS operations would be pointless.

An examination of most point patterns suggests that some other process is operating. In the real world, events at one place and time are seldom independent of events at another, so as a general rule, we expect point patterns to display spatial dependence, and hence to not match a hypothesis of spatial randomness. There are two basic ways in which we expect real processes to differ from IRP/CSR. First, variations in the receptiveness of the study area mean that the assumption of an equal probability of each area receiving an event cannot be sustained. For example, if events happen to be plants of a certain species, then almost certainly they will have a preference for patches of particular soil types, with the result that these plants would probably cluster on the favored soils at the expense of those less favored. Similarly, in a study of the geography of a disease, if our point objects represent locations of cases of that disease, these will naturally tend to cluster in more densely populated areas. Statisticians refer to this type of influence on a spatial process as a *first-order effect*.

Second, the assumption that event placements are independent of each other often cannot be sustained. Two deviations from independence are possible. Consider, for example, the settlement of the Canadian prairies in the latter half of the nineteenth century. As settlers spread, market towns grew up in competition with one another. For various reasons, notably the competitive advantage conferred by being near a railway lines, some towns prospered while others declined, with a strong tendency for successful towns to be located far from other successful towns as the market areas of each expanded. The result was distinct spatial separation in the distribution of towns, with a tendency toward uniform spacing of the sort predicted by central place theory (see King, 1984). In this case, point objects tend to suppress nearby events, reducing the probability of another point close by. Other real-world processes involve *aggregation* or *clustering* mechanisms where by the occurrence of one event at a particular location increases the probability of other events nearby. Examples include the spread of contagious diseases, such as foot and mouth disease in cattle and tuberculosis in humans, or the diffusion of an innovation through an agricultural community—where farmers are more likely to adopt new techniques that their neighbors have already used with success. Statisticians refer to this second type of influence as *a second-order effect*.

Both first- and second-order effects mean that the chances of an event occurring change over space, and we say that the process is no longer *stationary*. The concept of stationarity is not a simple one, but is essentially the idea that the rules that govern a process and control the placement of entities, although probabilistic, do not change, or *drift* over space. In a point process, the basic properties of the process are set by a single parameter—the probability that any small area will receive a point—called, for obvious reasons, the *intensity* of the process. Stationarity implies that the intensity does not change over space. To complicate matters further, we can also think in terms of first- and second-order stationarity. A spatial process is *first-order stationary* if there is no variation in its intensity over space, and it is *second-order stationary* if there is no interaction between events. The IRP is *both first- and second-order stationary*. Another possible class of intensity variation is where a process varies with spatial direction. Such a process is called *anisotropic* and may be contrasted with an *isotropic* process, where directional effects do not occur.

So, we have the possibility of both first- and second-order effects in any spatial process, and both can lead to either uniformity or clustering in the distribution of the point objects. Herein lies one important weakness of spatial statistical analysis: observation of just a single realization of a process—for example, a simple dot/pin map—is almost never sufficient to enable us to decide which of these two effects is operating. In other words, departures from an independent random model may be detected using the tests we outline in Chapter 5, but it will almost always be impossible to say whether they are due to variations in the environment or to interactions between events.

4.5. STOCHASTIC PROCESSES IN LINES, AREAS, AND FIELDS

So far, we have concentrated on IRP/CSR applied to spatial point processes. At this stage, if you are primarily interested in analyzing point patterns, you may want to read the next chapter. However, it is important to note that the same idea of mathematically defining spatial processes has also been applied to the generation of patterns of lines, areas, and the values in continuous fields. In this section, we briefly survey these cases. Many of these ideas will be taken further in later chapters.

Line Objects

Just as point objects have spatial pattern, line objects have length, direction, and, if they form part of a network, connection. It is theoretically possible to apply similar ideas to those we have used above to determine expected path lengths, directions, and connectivity for mathematically defined processes that generate sets of lines. However, this approach has not found much favor.

Random Lines

Consider a blank area such as an open park or plaza to be crossed by pedestrians or shoppers and across which no fixed paths exist. An analogous process to the independent random location of a point is to randomly select a location on the perimeter of the area, allowing each point an equal and independent chance of being selected, and then to draw a line in a random direction from the selected point until it reaches the perimeter. As an alternative, and generating a different distribution, we could randomly select a second point, also on the perimeter, and join the two points. Draw such an area and one such line on a sheet of paper. Next, produce a series of random lines, so that the pattern they make is one realization of this random process. What do you think the frequency distribution of these line lengths would look like?

What values would we expect, in the long run, from this IRP? Although the general principles are the same, deducing the expected frequencies of path lengths given an IRP is more difficult than it was for point patterns. There are three reasons for this. First, recall that the frequency distribution of quadrat counts is *discrete*; they need only be calculated for whole numbers corresponding to cell counts with $k = 0, 1, 2, \ldots,\ n$ points in them. Path lengths can take on any value, so the distribution involved is a *continuous probability density function*. This makes the mathematics a little more difficult. Second, a moment's doodling quickly shows that, because they are constrained by the perimeter of the area, path lengths strongly depend on the shape of the area they cross. Third, mathematical statisticians have paid less attention to line-generating processes than they have to point-generating ones. One exception is the work of Horowitz (1965), described by Getis and Boots (1978).

Starting from the independent random assumptions already outlined, Horowitz derives the probabilities of lines of a given length for five basic shapes: squares, rectangles, circles, cubes, and spheres. His results for a rectangle are shown in Figure 4.6. The histogram in the plot is based on a spreadsheet simulation of this situation, while the line shows the theoretical probability density function derived by Horowitz.

There are several points to note: The probability associated with any exact path length in a continuous probability distribution is very small. Thus, what is plotted is the probability density, that is, the probability per unit change in length. This probability density function is strongly influenced by the area's shape. There are a number of very practical situations in which the statistical properties of straight-line paths across specific geometric shapes are required,

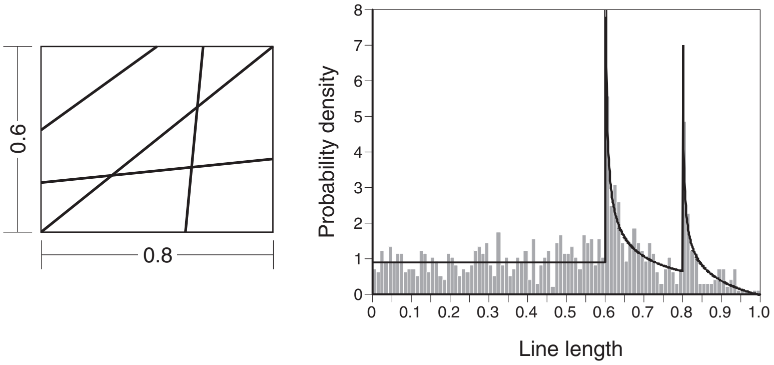


Figure 4.6 Theoretical probability density function (the line) and a single realization of the distribution of line lengths across a rectangular area (the histogram).

but these occur mostly in physics (gamma rays across a reactor, sound waves in a room, and so on) rather than in geography. A possible application, to pedestrian paths across a circular shopping plaza, is given in Getis and Boots (1978), but it is not very convincing. A major difficulty is that few geographic problems of interest have the simple regular shapes that allow precise mathematical derivation of the probabilities. Instead, it is likely to be necessary to use computer simulation to establish the expected independent random probabilities appropriate to more complex real-world shapes.

A related but more complex problem, with more applicability in geography, is that of establishing the probabilities of all possible distances *within* irregular shapes, rather than simply across the shape, as in the Horowitz model. Practical applications might involve the lengths of journeys in cities of various shapes, the distances between the original homes of marriage partners, and so on. Given such data, the temptation is to test the observed distribution of path lengths against some uniform or random standard without taking into account the constraints imposed by the shape of the study area. In fact, a pioneering paper by Taylor (1971) shows that the shape strongly influences the frequency distribution of path lengths obtained, and it is the constrained distribution that should be used to assess the observed results. As suggested above, Taylor found it necessary to use computer simulation rather than mathematical analysis.

Sitting Comfortably?

An illustration of the importance of considering the possibilities created by the shapes of things is provided by the following example. Imagine a coffee shop where all the tables are square, with one chair on each of the four sides. An observational study finds that when pairs of customers sit at a table, those who choose to sit across the corner of a table outnumber those who prefer to sit opposite one another by a ratio of 2 to 1. Can we conclude that there is a psychological preference for corner sitting? Think about this before reading on.

In fact, we can draw no such conclusion. As Figure 4.7 clearly shows, there are only two possible ways that two customers can sit opposite one another across a table, but there are four ways—twice as many—that they may sit across a table corner. It is perfectly possible that the observation described tells us nothing at all about the seating preferences of customers, because it is exactly what we would expect to find if people were making random choices about where to sit.

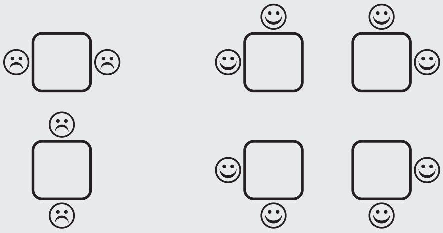


Figure 4.7 Possible ways of sitting at a coffee shop table.

The shape of the tables affects the number of possible arrangements, or the configurational possibilities. In much the same way, the shape of an urban area, and the structure of its transport networks, affect the possible journeys and journey lengths that we might observe. Of course, the coffee shop seating arrangement is a much easier example to do the calculations for than is typical in a geographic application.

The idea of an IRP has been used more successfully to study the property of line direction. Geologists interested in sediments such as glacial tills, where the orientations of the particles have process implications, have done most of this work. In this case, we imagine lines to have a common origin at the center of a circle and randomly select points on the perimeter, measuring the line direction as the angle from north, as shown in Figure 4.8.

A comprehensive review of this field, which is required reading for anyone with more than a passing interest, is the book by Mardia (1972) or its more recent, substantially revised edition (Mardia and Jupp, 1999). In till fabric analysis, any directional bias is indicative of the direction of a glacier flow. In

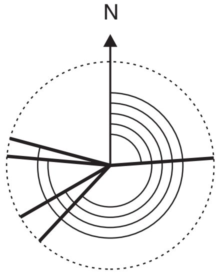


Figure 4.8 Randomly generated line segments are produced and their angle measured relative to north.

transport geography it could indicate a directional bias imparted by the pattern of valleys along which the easiest routes were found, and so on.

Line data are often organized in networks. There is a large body of recent work in numerous disciplines examining the statistical properties of how networks that grow in a variety of ways are structured (see Watts, 2003, and Barabàsi, 2002, for accessible introductions to this vast literature). The properties of these networks are relevant to the structure of the Internet, the brain, social networks, and epidemic spread (among many other things). However, because the nodes in such networks are not necessarily spatially embedded, such work is less relevant to situations where the nodes linked into a network have well-defined geographic locations.

In the past, geographers usually took the structure of a network expressed by its pattern of connections as a given, attempting to relate that structure to flows along the various paths. However, there is also a literature exploring the idea of network generation by random joining of segments. This is in the field of geomorphology, where attempts have been made, notably by Shreve (1966), to relate the observed "tree" networks of rivers in a drainage basin to predictions of possible models of their evolution. It turns out that natural tree networks have patterns of connection that could be fairly probable realizations of a random model. The geomorphologic consequences of this discovery, together with further statistical tests and an in-depth review, are to be found in Werritty (1972). For dissenting views, see Milton (1966) and Jones (1978).

In contrast, less attention has been paid to the statistical analysis of spatially embedded networks that are not tree-like (in practice, most networks). Exceptions are the work of the statistician Ling (1973), summarized in Getis and Boots (1978, p. 104) and Tinkler (1977). As before, we can propose a random model as a starting point and compare its predictions with those for any observed network with the same number of nodes and links. This problem turns out to be mathematically very similar to the basic binomial model used in describing the random allocation of point events to quadrats. Here we assign links to nodes, but with an important difference. Although the assignment of each link between nodes may be done randomly, so that at each step all nodes have an equal probability of being linked, each placement reduces the number of available possible links; hence, the probability of a specific link will change. The process is still random, but it now involves *dependence* between placements. If there are n nodes with q paths distributed among them, it can be shown (see Tinkler, 1977, p. 32) that the appropriate probability distribution taking this dependence into account is the *hypergeometric distribution*.

Another area of related work is the statistics of random walks. A random walk is a process that produces a sequence of point locations either in continuous space or on a lattice or grid. Random walk theory has important application in physics, where it is closely related to real-world physical processes such as Brownian motion and the diffusion of gases. A fairly accessible introduction to the theory of random walks can be found in Berg (1993), where the examples are from biology. In recent years, this work has become more relevant to the study of topics such as the movement patterns of animals, and those of people in crowded buildings and streets, as our ability to record movement tracks has increased through the miniaturization of GPS devices. As GPS becomes more commonplace in everyday life, most obviously in cellular phones, so that such tracking data are more readily available for analysis, it is likely that the basic ideas of random walk theory will become relevant in geographic applications. As with the other work mentioned here, a critical challenge will be applying highly abstract models of pure random walks to more constrained situations such as journeys on road networks.

Area Objects

Maps based on area data are probably the most common in the geographic literature. However, in many ways, they are the most complex cases to map and analyze. Just as with points and lines, we can postulate a process and then examine how likely a particular observed pattern of area objects and the values assigned to them is as a realization of that process. Imagine a pattern of areas. The equivalent of the IRP/CSR process would be either to “color” areas randomly to create a chorochromatic map or to assign values to areas, as in a choropleth map. In both cases, it is possible to think of this process as independent random (IRP/CSR) and to treat observed maps as potential realizations of that process.

Well, Do It!

On squared paper, set out an 8 by 8 ‘‘chessboard’’ of area objects. Now, visit the squares one after another and flip a coin for each one. If it lands heads, color the square black; if it lands tails, color it white. The resulting board is a realization of IRP/CSR in the same way as point placement. You can see that a perfect alternation of black and white squares, as on a real chessboard, is unlikely to result from this process. We consider the analysis of this type of setting in more detail in Chapter 7.

In fact, as we will find in Chapter 7, in the real world, randomly shaded maps are rare as a direct consequence of the “first law of geography.” This is occasionally also called Tobler’s Law and states that “[e]verything is related to everything else, but near things are more related than distant things” (Tobler, 1970, p. 234). Properly speaking, the first law of geography is an observational law, derived from the fact that much of what we see around us is spatially autocorrelated. To say that observed data are spatially autocorrelated is equivalent to saying that we do not think that they were generated by IRP/CSR.

A further complication that arises in trying to apply IRP/CSR to areal data is that the pattern of adjacency between areas is involved in the calculation of descriptive measures of the pattern. This means that information about the overall frequency distribution of values or colors on a map is insufficient to allow calculation of the expected range of map outcomes. In fact, any particular spatial arrangement of areal units must be considered separately in predicting the likely arrangements of values. This introduces formidable extra complexity—even for mathematicians—so it is common to use computer simulation rather than mathematical analysis to predict likely patterns.

Fields

An IRP may also be used as a starting point for the analysis of continuous spatial fields. First, consider the following thought exercise.

Random Spatial Fields

There are two clear differences between a point process and the same basic idea applied to spatial fields. First, a field is by definition continuous, so that every place has a value assigned to it and there are no abrupt jumps in value as one moves across the study region, whereas a point process produces a discontinuous pattern of dots. Second, the values of a scalar field aren't simply 0/1, present/absent; instead, they are ratio or interval-scaled numbers. So, a ''random'' field model will consist of random sampling at every point in the plane from a continuous probability distribution.

It is possible to construct such a random field using randomly chosen values sampled from the standard normal distribution, and you are invited to try to do this. Set out a grid of size (say) 20 by 20, and at each grid intersection, write in a value taken from the standard normal distribution. Now produce an isoline map of the resulting field.

Even without actually doing this exercise, you should realize that it won't be easy, since the random selection and independence assumption means that any value from $-\infty$ to $+\infty$ can occur anywhere across the surface, including right next to each other! In fact, with this type of model, from time to time you will get distributions that can be isolined and look vaguely real. As a bad joke, one of us used to introduce a laboratory class on isolining by asking students to isoline random data without revealing that the data were random. Often students produced plausible-looking spatial patterns.

As with area objects, it should be clear that, although it is used in many other sciences, this simple random field model is just the beginning as far as geography is concerned. The branch of statistical theory that deals with continuous field variables is called *geostatistics* (see Isaaks and Srivastava, 1989; Cressie, 1991) and develops from IRP/CSR to models of field variables that have three elements:

- A deterministic, large-scale spatial trend or "drift."
- Superimposed on this is a "regionalized variable" whose values depend on the autocorrelation and that is partially predictable from knowledge of the spatial autocorrelation.
- A truly random error component or "noise" that cannot be predicted.

For example, if our field variable consisted of the rainfall over a maritime region such as Great Britain, then we might identify a broad regional decline in average values (the drift) as we go inland, superimposed on which are local values dependent on the height of the immediate area (the values for the regionalized variable), on top of which is a truly random component that represents very local effects and inherent uncertainty in measurement (see, for example, Bastin et al., 1984). In Chapter 10, we discuss how the geostatistical approach can be used to create optimum isoline maps.

4.6. CONCLUSIONS

In this chapter, we have taken an important step down the road to spatial statistical analysis by giving you a clearer picture of the meaning of a spatial process. Our developing picture of spatial statistical analysis is shown in Figure 4.9. We have seen that we can think of a spatial process as a description of a method for generating a set of spatial objects. We have concentrated on the idea of a mathematical description of a process, partly because it is the easiest type of process to analyze and partly because mathematical descriptions or models of processes are common in spatial analysis.

Another possibility, which we have not examined at in any detail, is mentioned in Figure 4.9 and is of increasing importance in spatial analysis, as we shall see in the coming chapters. A *computer simulation* or *model* may also represent a spatial process. It is easy to imagine automating the rather arduous process of obtaining random numbers from a phone book in order to

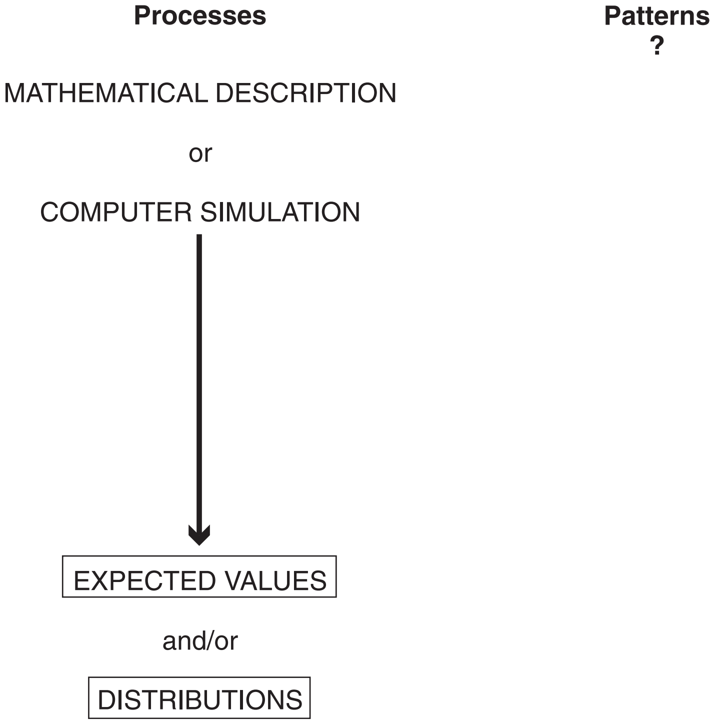


Figure 4.9 The developing framework for spatial statistical analysis. We now have a clearer picture of the meaning of a spatial process. Patterns will be tackled in the next chapter.

generate a set of points according to the IRP/CSR. A few minutes with the random number generation functions and scatterplot facilities of any spreadsheet program should convince you of this. In fact, it is also possible to represent much more complex processes using computer programs. The simulations used in weather prediction are the classic example of a complex spatial process represented in a computer simulation.

Whatever way we describe a spatial process, the important thing is that we can use the description to determine the expected spatial patterns that might be produced by that process. In this chapter, we have done this mathematically for IRP/CSR. As we shall see, this is important because it allows us to make comparisons between the *predicted outcomes* of a process and the *observed patterns* of distribution of phenomena we are interested in. This is essential to the task of making statistical statements about spatial phenomena. In the next chapter, we will take a much closer look at the concept of pattern so that we can fill in the blank on the right-hand side of our diagram.

This chapter has covered a lot of ground and introduced some possibly unfamiliar concepts. Many of these are taken up in succeeding chapters as we look in detail at how spatial analysis is applied to point objects, area objects, and fields. For the moment, there are four key ideas that you should remember. First is the idea that any map, or its equivalent in spatial data, can be regarded as the outcome of a spatial process. Second, although spatial processes can be deterministic in the sense that they permit only one outcome, most of the time we think in terms of stochastic processes where random elements are included in the process description. Stochastic processes may yield many different patterns, and we think of a particular observed map as an individual outcome, or realization, of that process. Third, we can apply the basic idea of the IRP in various ways to all of the entity types (point, line, area, and field) discussed in Chapter 1. Finally, as illustrated using the case of point patterns and IRP/CSR, this approach enables us to use mathematics to make precise statements about the expected long-run average outcomes of spatial processes.

CHAPTER REVIEW

- In spatial analysis, we regard maps as outcomes of processes that can be *deterministic* or *stochastic*.
- Typically, we view spatial patterns as potential *realizations* of stochastic processes.
- The classic stochastic process is *complete spatial randomness* (CSR), also called the *independent random process* (IRP).

- When dealing with a pattern of point objects, under CSR the points are randomly placed, so that every location has an equal probability of receiving a point, and points have no effects on each other—so that there are no *first-* or *second-order effects*.
- The expected quadrat count distribution for CSR conforms to the *binomial distribution,* with p given by the area of the quadrats relative to the area of the study region and n by the number of events in the pattern. This can be approximated by the *Poisson distribution*, with the intensity given by the average number of events per quadrat.
- These ideas can also be applied, with modification as appropriate, to properties of other types of spatial objects—for example, to line object length and direction, to networks, to autocorrelation in area objects, and, finally, to spatially continuous fields.
- Tobler's first law of geography tells us that real-world geography almost never conforms to IRP/CSR since "Everything is related to everything else, but near things are more related than distant things."
- Sometimes this is a result of variation in the underlying geography that makes the assumption of equal probability (first-order stationarity) untenable. At other times, what has gone before affects what happens next, and so makes the assumption of independence between events (second-order stationarity) untenable. In practice, it is very hard to disentangle these effects merely by the analysis of spatial data.

REFERENCES

Barabàsi, A.-L. (2002) *Linked: The New Science of Networks* (Cambridge, MA: Perseus).

Bastin, G., Lorent, B., Duque, C., and Gevers, M. (1984) Optimal estimation of the average rainfall and optimal selection of rain gauge locations. *Water Resources Research*, 20: 463–470.

Berg, H. C. (1993) *Random Walks in Biology* (Princeton, NJ: Princeton University Press).

Cressie, N. A. C. (1991) *Statistics for Spatial Data* (Chichester, England: Wiley).

Getis, A. and Boots, B. (1978) *Models of Spatial Processes* (Cambridge: Cambridge University Press).

Gleick, J. (1987) *Chaos: Making a New Science* (New York: Viking Penguin).

Horowitz, M. (1965) Probability of random paths across elementary geometrical Shapes. *Journal of Applied Probability*, 2(1): 169–177.

Isaaks, E. H. and Srivastava, R. M. (1989) *An Introduction to Applied Geostatistics* (New York: Oxford University Press).

Jones, J. A. A. (1978) The spacing of streams in a random walk model. *Area*, 10: 190–197.

King, L. J. (1984) *Central Place Theory* (Beverly Hills, CA: Sage).

Ling, R. F. (1973) The expected number of components in random linear graphs. *Annals of Probability*, 1: 876–881.

Mardia, K. V. (1972) *Statistics of Directional Data* (London: Academic Press).

Mardia, K. V. and Jupp, P. E. (1999) *Directional Statistics* (Chichester, England: Wiley).

Milton, L. E. (1966) The geomorphic irrelevance of some drainage net laws. *Australian Geographical Studies*, 4: 89–95.

Shreve, R. L. (1966) Statistical law of stream numbers. *Journal of Geology*, 74: 17–37.

Taylor, P.J. (1971) Distances within shapes: an introduction to a family of finite frequency distributions. *Geografiska Annaler*, B53: 40–54.

Tinkler, K. J. (1977) *An Introduction to Graph Theoretical Methods in Geography*. Concepts and Techniques in Modern Geography; 14,56 pages (Norwich, England: Geo Books). Available at http://www.qmrg.org.uk/catmog.

Tobler, W. (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46: 23–40.

Watts, D. J. (2003) *Six Degrees: The Science of a Connected Age* (New York: Norton).

Werritty, A. (1972) The topology of stream networks. In: R.J. Chorley, ed., *Spatial Analysis in Geomorphology* (London: Methuen), pp. 167–196.

Chapter 5

Point Pattern Analysis

CHAPTER OBJECTIVES

In this chapter, we attempt to:

- Define the meaning of a *pattern* in spatial analysis
- Come to a better understanding of the concept of pattern generally
- Introduce and define a number of descriptive measures for point patterns
- Show how we can use the idea of the IRP/CSR as a standard against which to judge observed real-world patterns for a variety of possible measures

After reading this chapter, you should be able to:

- Define what is meant by *point pattern analysis* and list conditions that are necessary for it to make sense to undertake it
- Suggest measures of pattern based on first- and second-order properties such as the *mean center* and *standard distance, quadrat counts, nearest-neighbour distance,* and the more modern G, F, and K functions
- Describe how IRP/CSR may be used to evaluate various point pattern measures, and hence to make statistical statements about point patterns and outline the basic process involved

5.1. INTRODUCTION

Point patterns, where the only data are the locations of a set of point objects, represent the simplest possible spatial data. Nevertheless, this does not mean that they are especially simple to analyze. In applied geography using GIS, pure point patterns occur fairly frequently. We might, for example, be interested in so-called hot spot analysis, where the point events studied are the locations of crimes or of deaths from some disease. The locations of plants of various species, or of archaeological finds, are other commonly investigated point patterns. In these applications, it is vital to be able to describe the patterns made by the point events and to test whether or not there is either some concentration of events, or *clustering,* in particular areas or, alternatively, some evidence that the objects are evenly spread in space. In this chapter, we outline what is meant by a point pattern and what we mean when we talk about describing examples of such patterns. We then show how we can relate observed real-world patterns to the IRP described in Section 4.3.

We need at the outset to present some terminology. A *point pattern* consists of a set of events in a study region. Each *event* represents the existence of a point object of the type we are interested in at a particular location in the study region, but there may be more than one such event at any given location. In the sections that follow, a point pattern of n events is a set of locations $S = \{\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_i, \ldots, \mathbf{s}_n\}$ in which each event (or point) $\mathbf{s}_i$ has locational coordinates (x_i, y_i). The pattern occurs in a study region A that has area a. Note that we use the term *event* to mean the occurrence of an object of interest at a particular location. This is useful because we can distinguish between events in the point pattern and any other arbitrary locations in the study region. In the simple case, each event is simply the occurrence of the object, but it can also have additional information attached to it, in which case the set of events is called a *marked point pattern*. In ecology the "marks" might consist of other information about a plant, such as its age or health; in spatial epidemiology it might be the date of onset of a disease. Note also that the location of each event is represented mathematically by a vector, written in boldface roman type: $\mathbf{s}$.

There are a number of requirements for such a set of events to constitute a point pattern:

- The pattern should be *mapped on the plane*. Latitude/longitude data should therefore be projected appropriately, preferably to preserve distances between points. Generally, it is inappropriate to perform point pattern analysis on events scattered over a very wide geographic area unless the methods used take account of the distortions introduced by the map projection used.
- The study area should be *objectively determined*. Ideally, this should be independent of the pattern of events and is a consequence of the MAUP discussed in Section 2.2. It is important because a different study area might give us different analytical results leading to different conclusions. In practice, such independence is hard, often impossible, to achieve. It might be given by the borders of a country,

shoreline of an island, or edge of a forest, but often no such natural boundary to the study region exists. In this circumstance, we need to consider carefully the rationale behind the study area's definition and think about applying *edge corrections* that attempt to correct for at least some of the consequences.

- The pattern should be an enumeration, or *census*, of the entities of interest, not a sample; that is, all relevant entities in the study region should be included. The relevance of this issue depends greatly on the choice of statistic and the availability of data. It can be ignored in studies that use a formal sampling procedure, such as quadrat sampling in ecology, discussed in the next section. Similarly, in ecology, the measure known as the *mean distance to the nearest neighbor* can be estimated by sampling the events that make up a pattern. However, in most studies in geography, the point event locations will be given in advance of the study, with no possibility of being able to sample the pattern.
- There should be a *one-to-one correspondence* between objects in the study area and events in the pattern.
- Event locations must be *proper*. They should not be, for example, the centroids of areal units chosen as "representative", nor should they be arbitrary points on line objects. They should represent the point locations of entities that can sensibly be considered points at the scale of the study.

This is a restrictive set of requirements that is only rarely met. In this chapter, we will present what is therefore an idealized picture of spatial point pattern analysis as it might be applied to a very "clean" problem. In the real world such problems are rare, so why do we spend time on them? Our view is that it is only by working through the approaches and assumptions made in the analysis of a clean problem that one can understand the complications that arise in analysis using real-world data to address real-world problems. In Chapter 6, we extend the discussion to examine some of the better-understood complexities.

5.2. DESCRIBING A POINT PATTERN

Revision

One of the most important ways of describing a point pattern is to visualize it as a map. We suggest that at this point you revisit Section 3.6 to review the

(*continues*)

(*box continued*)

ways in which a point pattern can be mapped. A pin or dot map is the basic method employed, but increasingly, analysts prefer to use the output from a kernel density estimation to transform the pattern into an isoline (contour-type) map of estimated spatial densities. If the point pattern is marked in some way by an interval- or ratio-scaled variable, the most appropriate display is a located proportional symbol map.

With point objects, the pattern that we see on a map of the events is really all that there is. So, how do we describe a pattern quantitatively? This is surprisingly difficult. In general, there are two interrelated approaches based on *point density* and *point separation*. These are related, in turn, to the two distinct aspects of spatial patterns that we have already mentioned: first- and second-order effects. Recall that first-order effects are manifest as variations in the *intensity* of the process across space, which we estimate as the observed spatial density of events. When first-order effects are marked, the *absolute location* is an important determinant of observations. In a point pattern, clear variations across space in the number of events per unit area are observed that arise because of variations in some factor that makes locations more or less "attractive" for events to occur at them. When second-order effects are strong, there is *interaction* between locations, depending on the distance between them, and *relative location* is important. In point patterns, such effects are manifest as reduced or increased distances between neighboring or nearby events.

This first-order/second-order distinction is important but, again, it is necessary to emphasize that it is usually impossible to distinguish the effects in practice simply by observing spatial variations in the density of events.

This difficulty is illustrated in Figure 5.1. In the first panel, we would generally say that there is a first-order variation in the point pattern whose observed density increases from northeast to the southwest corner, where

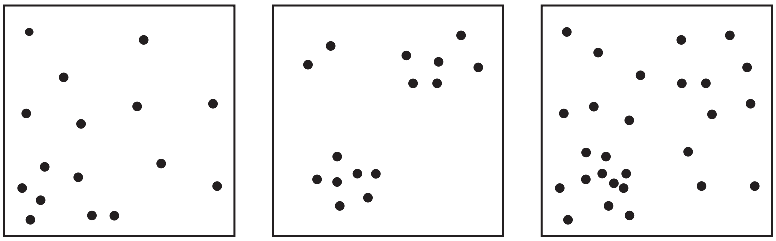


Figure 5.1 The difficulty of distinguishing first- and second-order effects.

it is highest. In the second panel, second-order effects are strong, with events grouped in distinct clusters. Obviously, this distribution could just as well be described in terms of first-order intensity variations, but it makes more sense to think of it in terms of grouping of events near one another. The third panel shows the difficulty of distinguishing the two effects in a more complex case. There is still a northeast-southwest trend, as in the first panel, but there is also a suggestion of clusters, as in the second panel. No simple description of this pattern in terms of separate first- and second-order effects is possible.

Centrography

Before considering more complex approaches, note that we can apply simple descriptive statistics to provide summary descriptions of point patterns. For example, the *mean center* of a point pattern S is given by

$$\bar{\mathbf{s}} = \left(\mu_x, \mu_y\right) = \left(\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n}\right) \tag{5.1}$$

That is, $\bar{\mathbf{s}}$ is the *point* whose coordinates are the average (or mean) of the corresponding coordinates of all the *events* in the pattern. We can also calculate a *standard distance* for the pattern:

$$d = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \mu_x\right)^2 + \left(y_i - \mu_y\right)^2}{n}} \tag{5.2}$$

Recalling basic statistics, this quantity is obviously closely related to the usual definition of the *standard deviation* of a data set, and it provides a measure of how dispersed the events are around their mean center. Taken together, these measurements can be used to plot a *summary circle* for the point pattern, centered at $\left(\mu_x, \mu_y\right)$ with radius d, as shown in the first panel of Figure 5.2.

More complex manipulation of the event location coordinates, in which the standard distance is computed separately for each axis, produces standard distance ellipses, as shown in the second panel of Figure 5.2. Summary ellipses gives some indication of the overall shape of the point pattern as well as its location. The calculation in this case is done for two orthogonal directions separately, and the results are resolved by trigonometry to get the correct orientation for the long and short axes of the ellipse.

These approaches, for obvious reasons called *centrography* in the literature, are sometimes useful for comparing point patterns or for tracking

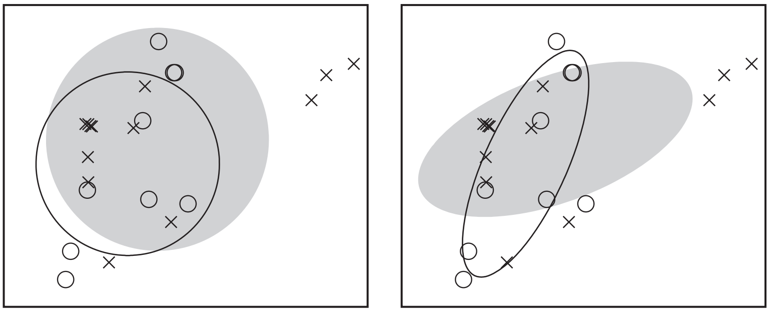


Figure 5.2 Summary circles and mean ellipses for two point patterns (open circles and crosses). The black outlined circle and ellipse summarize the open circle events, while the gray-shaded circle and ellipse summarize the crosses.

change in a pattern over time, but they don't provide much information about the pattern itself and they are extremely sensitive to the borders of the study region chosen. Description of the pattern itself has more to do with variations from place to place within the pattern and with the relationships between events in the pattern study region. More complex measures are therefore required to fully characterize a pattern, as discussed in the next sections.

Density-Based Point Pattern Measures

Density-based approaches to the description of a point pattern characterize the pattern in terms of its *first-order properties*. In doing this, we must be careful to maintain a distinction between the *intensity* of the spatial process itself, λ, and the observed density of events in the study region, which is frequently taken as an estimate of this real intensity. We can readily determine the crude density, or estimate of the overall *intensity*, of a point pattern. This is given by

$$\hat{\lambda} = \frac{n}{a} = \frac{\#(S \in A)}{a} \tag{5.3}$$

where $\hat{\lambda}$ is the estimated intensity and $\#(S \in A)$ is the number of events in pattern S found in study region A of area a in appropriate squared distance units such m^2 or km^2. One serious difficulty with density as a measure is its sensitivity to the definition of the study area. This is a difficulty with all density measures and is especially problematic when we attempt to calculate a "local" density. In Figure 5.3 the total number of events in successively

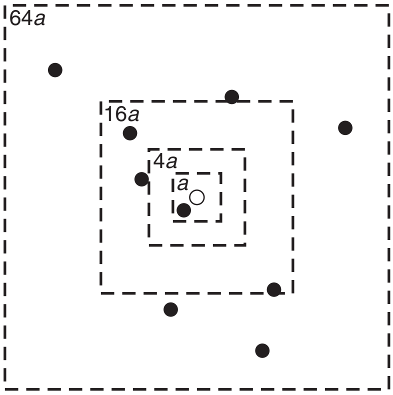


Figure 5.3 The difficulty with density. Calculating a local density measure for study areas defined in different ways.

larger regions with areas a, $4a$, $16a$, and $64a$ is 2, 2, 5, and 10, respectively. If a is a unit area (say, 1 km^2), then this gives us densities of 2.0, 0.5, 0.31, and 0.15, and the density around the central point changes depending on the study area. Without resorting to the calculus, there is no easy way to deal with this problem, and such methods are beyond the scope of this book. Kernel density methods (see Section 3.6) are one possible approach to this issue. In the next section, we discuss another.

Quadrat Count Methods

We lose a lot of information when we calculate a single summary statistic like overall density, and we have just seen that there is strong dependence on the definition of the study area. One way of getting around this problem is to record the number of events in the pattern that occur in a set of cells, or *quadrats*, of some fixed size. You will recall that this approach was discussed in Section 4.3 (see Figure 4.4). This can be done either by taking an exhaustive *census* of quadrats that completely fill the study region, with no overlaps, or by *randomly placing* quadrats across the study region and counting the number of events that occur in them (see Rogers, 1974; Thomas, 1977). The two approaches are illustrated in Figure 5.4.

The random sampling approach is more frequently applied in field work—for example, in surveying vegetation in plant ecology (see Greig-Smith, 1964). Much of the statistical theory of quadrat measures relates to the sampling approach, which also has the merit of allowing shapes that do not tessellate the plane (such as circles) to be used. With random sampling, it is

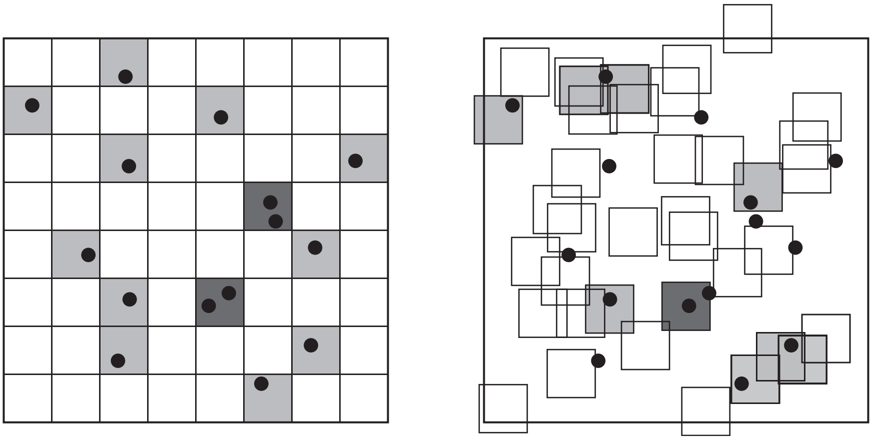


Figure 5.4 The two quadrat count methods: an exhaustive census (left) and random sampling (right). Quadrats containing events are shaded.

also possible to increase the sample size simply by adding more quadrats. This may be advantageous for relatively sparse patterns where a larger quadrat is required to “catch” any events, but it would rapidly exhaust a study region with only a small number of quadrats. The sampling approach also makes it possible to describe a point pattern without having complete data on the whole pattern. This is a distinct advantage for work in the field, provided that care is taken to remove any biases concerning where quadrats are placed; otherwise, a very misleading impression might be obtained. It is worth noting that the sampling approach can miss events in the pattern. Several events in the pattern in Figure 5.4 are not counted by the quadrats indicated, and some are double counted. The important thing is that all the events in any quadrat are counted. The sampling approach is really an attempt to *estimate* the likely number of events in a quadrat-shaped region by random sampling.

The exhaustive census-based method is used more commonly in geographic applications such as spatial epidemiology or criminology, where the measured event data are all that we have and there is no opportunity to sample the pattern. The choice of origin and quadrat orientation affects the observed frequency distribution, and the chosen size of quadrats also has an effect. Large quadrats produce a very coarse description of the pattern, but as quadrat size is reduced, many will contain no events and only a few will contain more than one, so the set of counts is not useful as a description of pattern variability. Note that, although rare in practice, exhaustive quadrats could also be hexagonal or triangular, as shown in Figure 5.5.

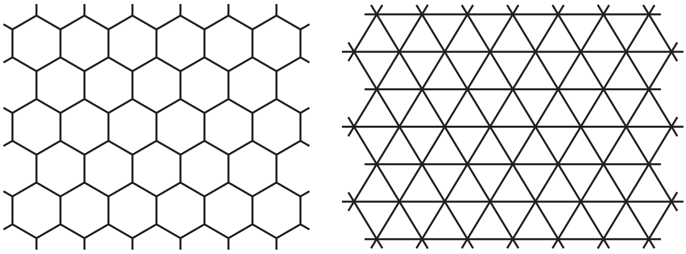


Figure 5.5 Alternative quadrat shapes used in a quadrat census.

Thought Exercise

Why do you think it is usual to use a regular grid of quadrats? Of the three shapes we have illustrated (squares, hexagons, and triangles), which would allow you to create the same shape at both larger and smaller scales by combining quadrats or subdividing them? What other shapes are possible, and what properties might they have? Do circular quadrats tessellate the plane? Would it surprise you to learn that whole books have been written on the topic of ''tiling''? (see Grünbaum and Shephard, 1987).

Whichever approach we adopt, the outcome is a list of *quadrat counts* recording the number of events that occur in each quadrat. These are compiled into a frequency distribution listing how many quadrats contain zero events, how many contain one event, how many contain two, and so on.

As an example, we look at the distribution of the coffee shops of a particular company in central London (in late 2000). This distribution is mapped in Figure 5.6 and has $n = 47$ coffee shops. Using $x = 40$ quadrats to compile the

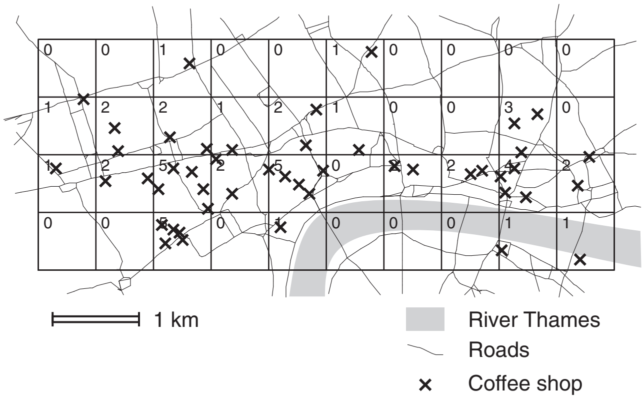


Figure 5.6 Coffee shops in central London.

Table 5.1 Quadrat Counts and Calculation of the Variance for the Coffee Shop Pattern

No. of events, K	*No. of quadrats, X*	$K - \mu$	$(K - \mu)^2$	$X(K - \mu)^2$
0	18	−1.175	1.380625	24.851250
1	9	−0.175	0.030625	0.275625
2	8	0.825	0.680625	5.445000
3	1	1.825	3.330625	3.330625
4	1	2.825	7.980625	7.980625
5	3	3.825	14.630625	43.891875
Totals	**40**			**85.775000**

count, we have a mean quadrat count $\mu = 47/40 = 1.175$. Counts for each quadrat are indicated in the diagram.

These quadrat counts are compiled in Table 5.1, from which we calculate the observed variance s^2 to be $85.775/(40 - 1) = 2.19936$. A useful summary measure of these counts is the ratio of their variance to their mean, the variance-mean ratio (VMR), which, as the table shows, is observed to be $2.19936/1.175 = 1.87180$. A property of the Poisson distribution that we introduced in Section 4.3 as resulting from the IPR/CSR process is that its mean and variance are equal, so that if the quadrat counts describe such a distribution, their VMR should be 1.0. Clearly it isn't: the observed VMR is almost double the Poisson distribution value. Because it indicates high variability among the quadrat counts, implying that more quadrats contain very few or very many events than would be expected by chance, a moment's thought will suggest that this is indicative of clustering. In this case, three quadrats contain five coffee shops, and this finding contributes heavily to the result. In general, a VMR greater than 1.0 indicates a tendency toward clustering in the pattern, and a VMR less than 1.0 is indicates an evenly spaced arrangement.

Distance-Based Point Pattern Measures

The alternative to using density-based methods is to look at the distances between events in a point pattern. This provides a more direct description of the second-order properties of the pattern. In this section, we describe the more frequently used distance-based methods.

The *nearest-neighbor distance* for an event in a point pattern is the distance from that event to the nearest event also in the point pattern. The distance $d(\mathbf{s}_i, \mathbf{s}_j)$ between events at locations $\mathbf{s}_i$ and $\mathbf{s}_j$ may be calculated

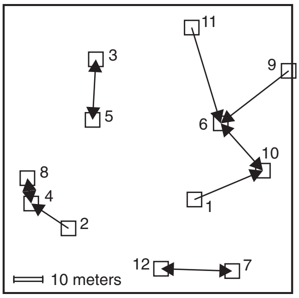


Figure 5.7 Distances to the nearest neighbor for a small point pattern. The nearest neighbor to each event lies in the direction of the arrow pointing away from it.

using Pythagoras's theorem:

$$d(\mathbf{s}_i, \mathbf{s}_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{5.4}$$

The nearest event in the pattern to each event can therefore be easily found. If we denote this value for event $\mathbf{s}_i$ by $d_{\min}(\mathbf{s}_\mathrm{i})$, then a frequently used measure is the *mean nearest-neighbor distance* originally proposed by Clark and Evans (1954):

$$\bar{d}_{\min} = \frac{\sum_{i=1}^{n} d_{\min}(\mathbf{s}_i)}{n} \tag{5.5}$$

Calculations for the point pattern in Figure 5.7 are shown in Table 5.2.

Table 5.2 Calculations for the Nearest-Neighbor Distances for the Point Pattern Shown in Figure 5.7

Point	*X*	*Y*	*Nearest neighbor*	$D_{\min}$
1	66.22	32.54	10	25.59
2	22.52	22.39	4	15.64
3	31.01	81.21	5	21.11
4	9.47	31.02	8	9.00
5	30.78	60.10	3	21.14
6	75.21	58.93	10	21.94
7	79.26	7.68	12	24.81
8	8.23	39.93	4	9.00
9	98.73	77.17	6	29.76
10	89.78	42.53	6	21.94
11	65.19	92.08	6	34.63
12	54.46	8.48	7	24.81

Note that it is not unusual for points to have the same nearest neighbor (9:10:11, 2:8, and 1:6) or to be nearest neighbors of each other (3:5, 7:12, and 4:8). In this case, $\Sigma d_{\min} = 259.40$, so the mean nearest-neighbor distance is 21.62.

In some problems—for example, when dealing with trees in a forest area—it might be possible to obtain the mean nearest-neighbor distance by sampling a population of events, and for each event in the sample, finding its nearest neighbor and the relevant distance. No matter how we find it, a drawback of the mean nearest-neighbor distance is that it throws away a lot of information about the pattern. Summarizing all the nearest-neighbor distances in Table 5.2 by a single mean value is convenient, but it seems almost too concise to be really useful. This drawback of the method is addressed in more recently developed approaches.

A number of extensions to the nearest-neighbor approach have been developed. These go by the unexciting names of the G and F functions. Of these, the G function, sometimes called the *refined nearest neighbor*, is the simplest. It uses exactly the same information contained in Table 5.2, but instead of summarizing it using the mean, we examine the *cumulative frequency distribution of the nearest-neighbor distances*. Formally, this is defined as

$$G(d) = \frac{\#(d_{\min}(\mathbf{s}_i) < d)}{n} \tag{5.6}$$

so the value of G for any particular distance, d, tells us what fraction of all the nearest-neighbor distances in the pattern is less than d. Figure 5.8 shows the G function for the example in Figure 5.7.

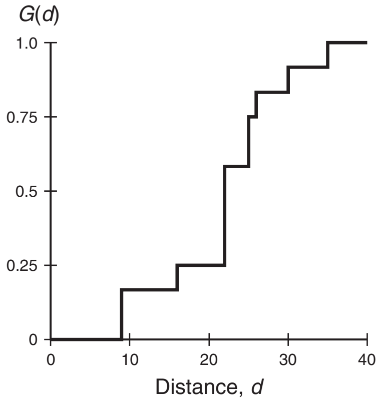


Figure 5.8 The G function for the point pattern of Figure 5.7 and Table 5.2.

Refer back to Table 5.2. The shortest nearest-neighbor distance is 9.00 between events 4 and 8. Thus, 9.00 is the nearest-neighbor distance for two events in the pattern. Since 2 out of 12 is a proportion of 2/12 = 0.167, $G(d)$ at distance $d = 9.00$ has the value 0.167. The next nearest-neighbor distance is 15.64, for event 2, and three events have nearest neighbors at this distance or less. Since 3 out of 12 is a proportion of 0.25, the next point plotted in $G(d)$ is 0.25 at $d = 15.64$. As d increases, the fraction of all nearest-neighbor distances that are less than d increases. This process continues until we have accounted for all 12 events and their nearest-neighbor distances.

The shape of this function can tell us a lot about the way events are spaced in a point pattern. If events are closely clustered together, then G increases rapidly at short distances. If events tend to be evenly spaced, then G increases slowly up to the range of distances at which most events are spaced, and only then increases rapidly. In our example, G increases most quickly in the $20 < d < 25$ range, reflecting the fact that many of the nearest-neighbor distances in this pattern are in that distance range. This example has a very "bumpy" plot because it is based on only a small number of nearest-neighbor distances ($n = 12$). Usually, n will be greater than this and smoother changes in G are observed.

The F function is closely related to G but may reveal other aspects of the pattern. Instead of accumulating the fraction of nearest-neighbor distances *between events* in the pattern, point locations anywhere in the study region are selected at random, and the minimum distance from these locations to any event in the pattern is determined. The F function is the cumulative frequency distribution for this new set of distances. If $\{\mathbf{p}_1 \ \ldots \ \mathbf{p}_i \ \ldots \ \mathbf{p}_m\}$ is a set of m randomly selected locations used to determine the F function, then formally

$$F(d) = \frac{\#[d_{\min}(\mathbf{p}_i, S) < d]}{m} \tag{5.7}$$

where $d_{\min}(\mathbf{p}_i, S)$ is the minimum distance from location $\mathbf{p}_i$ in the randomly selected set to any event in the point pattern S. Figure 5.9 shows a set of randomly selected locations in the study region for the same point pattern as before, together with the resulting F function. This has the advantage over G that we can increase the sample size m to get a smoother cumulative frequency curve that should give a better impression of the point pattern's properties. In practice, in software $F(d)$ is computed using an appropriate regular grid of locations rather than the random ones shown here.

It is important to note the difference between the F and G functions, as it is easy to get them mixed up and also because they behave differently for clustered and evenly spread patterns. This happens because while G shows

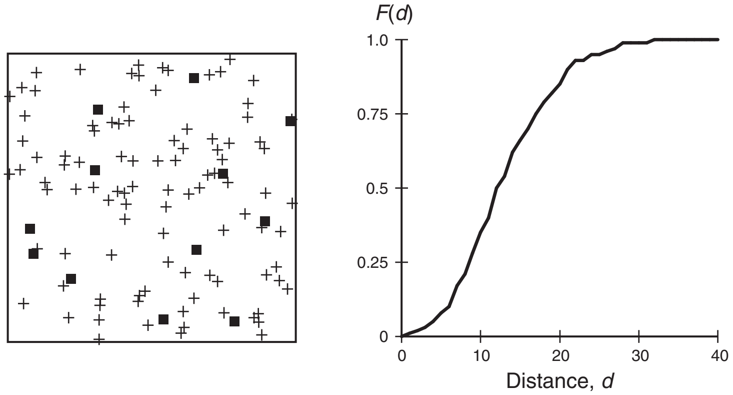


Figure 5.9 Random points (shown as crosses) for the same point pattern as before and the resulting *F* function.

how close together events in the pattern are, *F* relates to how far events are from arbitrary locations in the study area. So, if events are clustered in a corner of the study region, *G* rises sharply at short distances because many events have a very close nearest neighbor. The *F* function, on the other hand, is likely to rise slowly at first, but more rapidly at longer distances, because a good proportion of the study area is fairly empty, so that many locations are at quite long distances from the nearest event in the pattern. For evenly spaced patterns, the opposite is true. Most locations in *P* are relatively close to an event, so that *F* rises quickly at low *d*. However, events are relatively far from each other, so that *G* initially increases slowly and rises more quickly at longer distances.

It is possible to examine the relationship between *G* and *F* to take advantage of this different information. The likely relationships are demonstrated by the examples in Figure 5.10. The upper example is clearly clustered. As a result, most events (around 80% of them) have close near neighbors, so that the *G* function rises rapidly at short distances up to about 0.05. In contrast, the *F* function rises steadily across a range of distances. The lower example is evenly spaced, so that *G* does not rise at all until the critical spacing of about 0.05, after which it rises quickly, reaching almost 100% by a distance of 0.1. The *F* function again rises smoothly in this case. Note that the horizontal scale has been kept the same in these graphs. The important difference between the two cases is the relationship between the functions, which is reversed.

One failing of all the distance-based measures discussed so far, the nearest-neighbor distance and the *G* and *F* functions, is that they only make use of the *nearest* neighbor for each event or location in a pattern. This can be a major drawback, especially with clustered patterns where

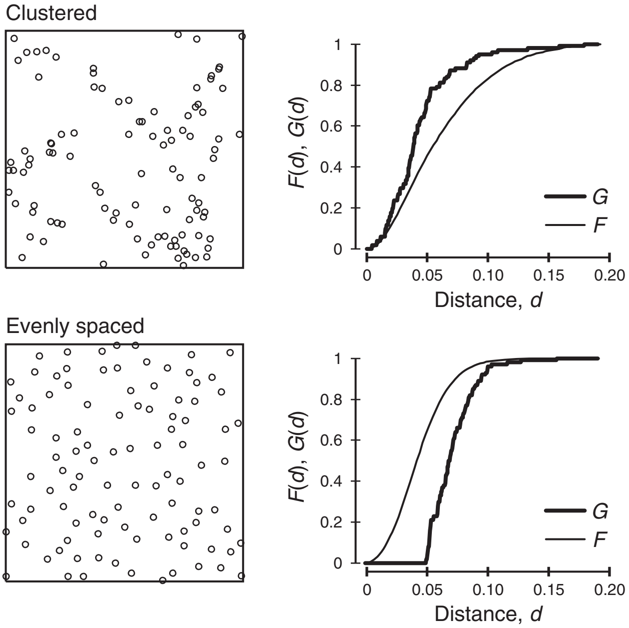


Figure 5.10 Comparing *F* and *G* functions for clustered and evenly distributed data.

nearest-neighbor distances are very short relative to other distances in the pattern, and can “mask” other structures in the pattern. A relatively simple way around this problem, which was suggested many years ago (Thompson, 1956; see also Davis et al., 2000), is to find the mean distances to the first, second, third, and so on nearest neighbors; however, in practice, a more common approach is to use K functions (Ripley, 1976) based on *all the distances* between events in S.

The easiest way to understand the calculation of a K function at a series of distances d is to imagine placing circles, of each radius d, centered on each of the events in turn as shown in Figure 5.11. The numbers of other events inside each circle of radius d is counted, and the mean count for all events is calculated. This mean count is divided by the overall study area event density to give $K(d)$. This process is repeated for a range of values of d. So, we have

$$\begin{aligned} K(d) &= \frac{\sum_{i=1}^{n} \#[S \in C(\mathbf{s}_i, d)]}{n\lambda} \\ &= \frac{a}{n} \cdot \frac{1}{n} \sum_{i=1}^{n} \#[S \in C(\mathbf{s}_i, d)] \end{aligned} \tag{5.8}$$

Remember that $C(\mathbf{s}_i, d)$ is a circle of radius d centered at $\mathbf{s}_\mathrm{i}$. The K function for clustered and evenly spaced patterns is shown in Figure 5.12.

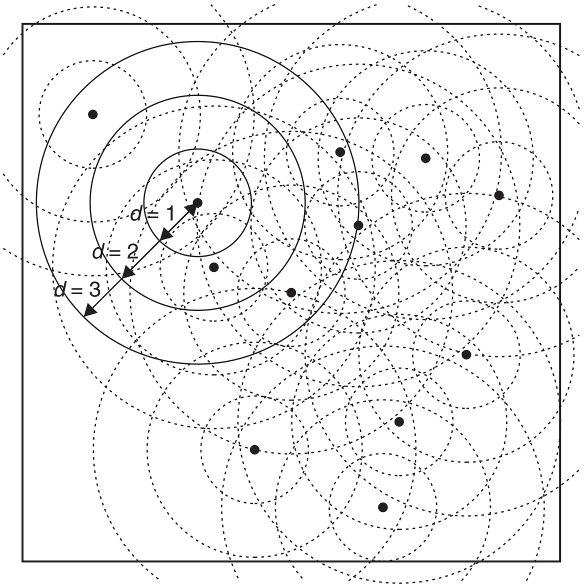


Figure 5.11 Determining the *K* function for a pattern. The measure is based on counting events within a series of distances of each event. Note how higher values of *d* result in more of the circular region around many events lying outside the study region.

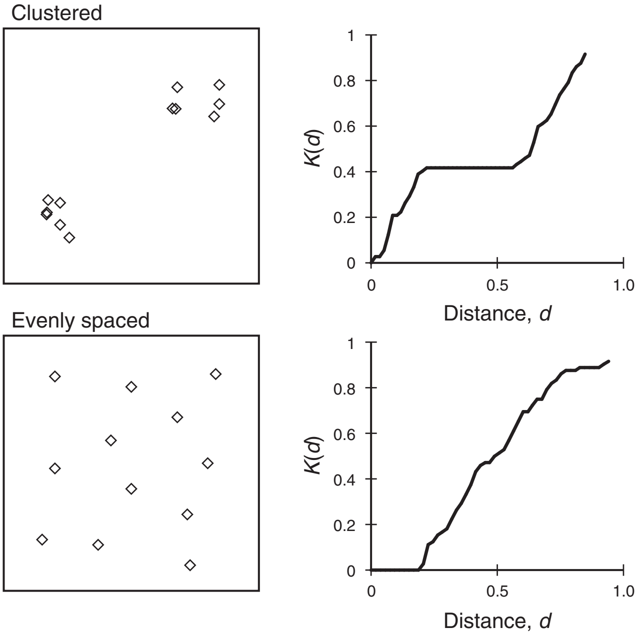


Figure 5.12 The *K* function for clustered and evenly spaced events.

Because *all* distances between events are used, this function provides more information about a pattern than either G or F. For the small patterns shown, it is easily interpreted. For example, the level portion of the curve for the clustered pattern in Figure 5.12 extends over a range of distances that does not match the separation between any pair of events. The lower end of this range (≈ 0.2) corresponds to the size of the clusters in the pattern, and the top end of this range (≈ 0.6) corresponds to the cluster separation. In practice, because there will usually be event separations across the whole range of distances, interpretation of K is usually less obvious than this. We will consider interpretation of the K function in more detail when we discuss how it is compared to expected functions for IRP/CSR.

Recent years have seen a variation on Ripley's $K(d)$ function being used, which has been variously called the *O-ring* statistic (Wiegand and Moloney, 2004) and the *pair correlation function* (or *neighborhood density function,* NDF) (see Perry et al., 2006), and for some patterns it may be more informative. As Figure 5.12 shows, the original $K(d)$ function is cumulative, with the proportion of events from each circle plotted as a function of the radius, d. All that the pair correlation function does is to plot the actual proportion in a series of annuli centered on each event. This somewhat reduces the ability to choose to measure pairs at any arbitrary separation distance, a problem avoided in the standard $K(d)$ by use of the cumulative distribution. Probability density estimation methods are used to convert the counts of pairs in various separation distance bins into a continuous function. The pair correlation function approach enables an analyst to get a clearer picture of any particular separation distances at which there are many or few pairs of events.

Edge Effects

A problem with all the distance functions we have discussed is that *edge effects* may be pronounced, especially if the number of events in the pattern is small. Edge effects arise from the fact that events (or point locations) near the edge of the study area tend to have higher nearest-neighbor distances even though they might have neighbors outside the study area that are closer than any inside it. Inspection of Figure 5.11 highlights how the problem becomes more severe at higher values of d when the circular region around many of the events extends outside the study area.

The easiest way to counter edge effects is to incorporate a *guard zone* around the edge of the study area. This is shown in Figure 5.13. Filled black dots in the study region are considered part of the point pattern for all purposes. Unfilled circles in the guard zone are considered in the determination of interevent distances for the G and K functions, or of point–event

Using Simulation to Show Why Edge Effects Matter

We can show the likely magnitude of this edge effect by a simple simulation exercise. Table 5.3 shows the results for the mean value of the Clark and Evans (1954) *R*-index from 100 simulated realizations of IRP/CSR using different numbers of events in the pattern.

Table 5.3 Simulation Results for the Clark and Evans *R* Statistic

No. of events, n	*Mean R value*
10	1.1628
12	1.1651
25	1.1055
50	1.0717
100	1.0440

Mathematical theory that we discuss in Section 5.3 tells us that the mean value for *R* in IRP/CSR should be 1.0 precisely. So, why do the simulation results not give this value?

Well, first, these are results from a simulation, so we would expect some differences between the values obtained from a relatively small number of realizations and the very-long-run results that theory predicts.

However, looking at the change in *R* more closely, as we increase the number of events in the pattern, what the table really shows is the bias due to edge effects. If you look back to Figure 5.7, you will see that events close to the study region border, such as 9, 10, and 11, are ''forced'' to find nearest neighbors inside the region when in all probability their true nearest neighbors would be outside it. In turn, this introduces into estimation of the mean some longer distances than an unbounded region would have given. Thus, instead of being 1.000, the random expectation produced by simulation is slightly higher. This effect tends to be greatest when the number of events is low, so that a large fraction of the events are on the edge of the pattern. As you can see, with only 10 or 12 events the effect is quite marked—a result 16% higher than theory predicts. As we increase the number of events, the relative effect of events near the border is reduced, so that at $n = 100$ we are pretty close to theoretical values. Even so, the result is still on the high side.

distances for the F function, but are not considered part of the pattern. Three examples are shown where an event's nearest neighbor is in the guard zone.

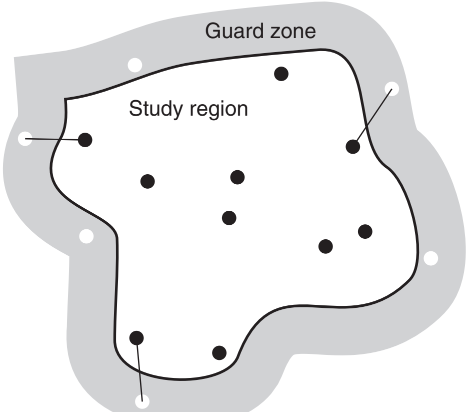


Figure 5.13 The use of a guard zone in a distance-based point pattern measure.

Using a guard zone has the disadvantage of involving collection of data that are not used in subsequent analysis. To use all the available data, Ripley (1977) suggested a weighted edge correction in which the distance between a pair of events is given a weight based on properties of the circle centered on the first point and passing through the second. If the circle is wholly within the study region, the weight is simply 1, but it is scaled by either the length of its circumference or the proportion of its area contained in the study region. A third approach uses the toroidal "wrap," which joins the top and left parts of the study region to the bottom and right, respectively, and then proceeds in the usual way to compute distances. Yamada and Rogerson (2003) provide an empirical study of the effect on the K function of these various corrections. They conclude that if the analysis is largely descriptive, to detect and characterize an observed pattern rather than to estimate parameters of a specific hypothesized point process, there is little point in using any of these corrections.

5.3. ASSESSING POINT PATTERNS STATISTICALLY

So far, we have presented a number of measures or *descriptive statistics* for point patterns. In principle, these are calculated and each may shed some light on the structure of the pattern. In practice, it is likely that different measures will reveal similar, but not necessarily the same, things about the pattern (see Perry et al., 2006), and especially whether it tends to be *clustered* or *evenly spaced*. When mapped, a clustered pattern is likely to have a "peaky" density pattern, which will be evident either in the quadrat counts or in strong peaks on a kernel-density estimated surface. It will also have short nearest-neighborhood distances, which will show up in the distance functions we have considered. An evenly distributed pattern exhibits the opposite: an even distribution of quadrat counts or a "flat" kernel-density estimated surface and relatively long nearest-neighbor distances. Such description may be quantitative, but it remains informal. In spatial statistical analysis the key questions are: *How clustered*? *How evenly spaced*? Against what benchmark are we to assess these measures of pattern?

The framework for spatial analysis that we have been working toward is now almost complete, and it enables us to ask such questions about spatial data and answer them statistically. Within this framework, we ask whether or not a particular set of observations could be a realization of some hypothesized process.

In statistical terms, our null hypothesis is that the pattern we are observing has been produced by a particular spatial process. A set of spatial data, a pattern or a map, is then regarded as a sample from the set of all possible realizations of the hypothesized process, and we use statistics to ask how *unusua*l the observed pattern would be if the hypothesized spatial process were operating. The complete framework is illustrated schematically in Figure 5.14.

Thus far, we have progressed separately down both branches of this framework. Chapter 4 focused on the left-hand branch of the framework. In it we saw how we could suggest a process, such as the IRP/CSR, and then use some relatively straightforward mathematics to say something about its outcomes. We will see later that computer simulation is now often used to do the same thing, but the outcome is the same: a description of the likely outcomes of the process in terms of the expected values of one of our measures of pattern.

In the first half of this chapter, we showed how we can follow the right-hand branch, taking a point pattern of events and deriving some measure such as quadrat counts, nearest-neighbor distances, or any of the F, G, K, and pair correlation functions. In the remainder of the chapter we consider the final step, which, as indicated in Figure 5.14, is to bring these two views together and compare them statistically to infer what we can about the underlying spatial process based on our observations of the pattern.

As we discuss in Section 6.2, there are actually a variety of approaches to making a statistical assessment. For now, because it relates directly to our discussion of the null hypothesis of IRP/CSR, we focus in this section on a hypothesis-testing approach from classical statistics. This approach asks the fundamental question: If IRP/CSR were the process operating, how probable would the observed pattern be? The probability value we arrive at, known as the *p-value*, is the probability that the observed pattern would have occurred

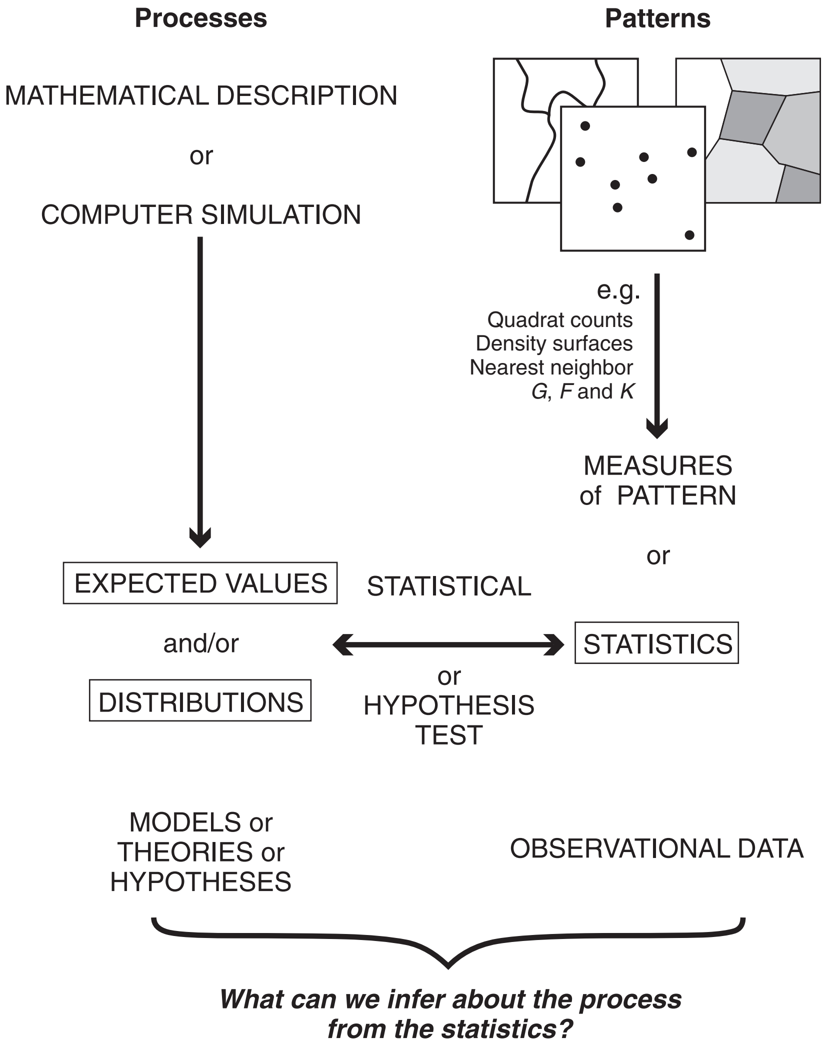


Figure 5.14 The conceptual framework for the statistical approach to spatial analysis.

as a realization of IRP/CSR. If the p-value is low ($p = 0.05$ is a commonly used threshold value), then we reject the null hypothesis and conclude that the observed pattern is unlikely to have been produced by IRP/CSR. A higher p-value leaves us unable to reject the null hypothesis, and we must acknowledge that the observed pattern might well be a realization of IRP/CSR.

This approach relies on having a good knowledge of the *sampling distribution* of expected values of our pattern measures for IRP/CSR. We have already seen that, in some cases, statisticians have developed theory that enables the sampling distribution to be predicted exactly for simple processes like IRP/CSR. In other cases, where analytic results are not known, computer simulation is often used to generate synthetic sampling distributions. This approach is becoming increasingly common, and we examine it in connection with the K function in Section 5.4.

Quadrat Counts

We saw in Section 4.3 that the expected probability distribution for a quadrat count description of a point pattern under the assumptions in IRP/CSR is given by the binomial distribution or, more practically, by a Poisson distribution approximation:

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{5.9}$$

where λ is the average *intensity* of the pattern per quadrat and e is the base of the natural logarithm system. Therefore, to assess how well a null hypothesis of complete spatial randomness explains an observed point pattern, we may compile a quadrat count distribution and compare it to the Poisson distribution with λ estimated from the point pattern. We have already seen that a simple measure for how well an observed distribution of quadrat counts fits a Poisson prediction is based on the property that its mean and variance are equal ($\mu = \sigma$), so the *variance-mean ratio* (VMR) is expected to be 1.0 if the distribution is Poisson. It is one thing to create an index such as this, but it is quite another to generate a significance test that answers the basic question posed at the bottom of Figure 5.14.

The most commonly suggested approach treats the problem as a goodness-of-fit test using the chi-square distribution as its standard. Table 5.4 summarizes this approach using the quadrat counts from Figure 5.6 and Table 5.1 for the London coffee shops example.

With six nonzero bins, we have $6 - 1 = 5$ degrees of freedom. The resulting chi-square value, at 32.261 ($p < 0.00001$!), is well above that required for significance at the 95% level, and we might be fairly confident in rejecting the null hypothesis that the underlying process is IRP/CSR. There are serious difficulties with this approach, however. The approximation of the chi-square statistic by the theoretical distribution is not good for a table such as this. The major part of the total obtained comes from the three quadrats that contain five or more coffee shops, which is consistent with our conclusion that the pattern is clustered. In addition, three of the bins contain expected frequencies that are less than 5, which is generally recommended for this

Table 5.4 Chi-Square Analysis for the London Coffee Shops Data from Figure 5.6 and Table 5.1

K, number of events in quadrat	*Observed number of quadrats, O*	*Poisson probability*	*Expected number, E*	*Chi-square $(O - E)^2/E$*
0	18	0.308819	12.35276	2.5817
1	9	0.362862	14.51448	2.0951
2	8	0.213182	8.52728	0.0326
3	1	0.083496	3.33984	1.6393
4	1	0.024527	0.98108	0.0004
5 or more	3	0.007114	0.28456	25.9123
Totals	40	1.000000	40.00000	32.2614

test; this issue will always arise when we look at the higher-frequency quadrats that actually contain the clusters. To follow this recommendation, we can group these three bins into one, representing quadrats with three or more coffee shops; the result is to make the pattern indistinguishable from a random one.

An equivalent test is to assess the calculated VMR of the quadrat counts statistically. The expected VMR value for a Poisson distribution is 1.0, and the product $(n - 1)$VMR where n is the number of quadrats is chi-square with $(n - 1)$ degrees of freedom. In this case, we get a chi-square test statistic of $1.8718 \times 39 = 73.0$. This value has an associated p-value of 0.0007, meaning that we would expect to observe such an extreme result in less than 1 case in 1,000, if the pattern were to have been generated by IRP/CSR. Again, this would lead us to reject the null hypothesis of IRP/CSR in this case. However, like the chi-square goodness-of-fit test, this approach is generally considered unreliable unless the mean number of events per quadrat is 10 or more. This requires us to use very large quadrats in most cases.

We can only conclude that while hypothesis testing is possible for quadrat count data, it is not reliable unless we are dealing with very large point data sets with a high mean intensity of events per quadrat.

Nearest-Neighbor Distances

If, instead of quadrat counts, we use mean nearest-neighbor distance to describe a point pattern, then we can use Clark and Evans's R statistic to test for conformance with IRP/CSR. Clark and Evans (1954) show that the expected value for the mean nearest-neighbor distance is

$$E(d) = \frac{1}{2\sqrt{\lambda}} \tag{5.10}$$

and they suggest that the ratio R of the observed mean nearest-neighbor distance to this value be used in assessing a pattern relative to IRP/CSR. Thus:

$$R = \bar{d}_{\min} \bigg/ \frac{1}{(2\sqrt{\lambda})} \tag{5.11}$$

An R value of *less than* 1 indicates of a tendency toward clustering, since it shows that observed nearest neighbor distances are shorter than expected. An R value of more than 1 indicatives of a tendency toward evenly spaced events. It is also possible to make this comparison more precise and to offer a significance test, this time based on the familiar normal distribution (see Bailey and Gatrell, 1995, pp. 98–101).

A Cautionary Tale: Are Drumlins Randomly Distributed?

The word *drumlin* is Irish and describes a long, low, streamlined hill. Drumlins occur in swarms, or drumlin fields, giving what geologists call *basket of eggs* landscapes. Although it is generally agreed that drumlins are a result of glaciation by an ice sheet, no one theory of their formation has been accepted. A theory proposed by Smalley and Unwin (1968) suggested that within drumlin fields the spatial distribution would conform to IRP/CSR. With data derived from map analysis, they used the nearest-neighbor statistic to show that this seemed to be true and provided strong evidence supporting their theory. The theory was tested more carefully in later papers by Trenhaile (1971, 1975) and Crozier (1976).

With the benefit of many years of hindsight, it is clear that the point pattern analysis methods used were incapable of providing a satisfactory test of the theory. First, there are obvious difficulties in considering drumlins as point objects and in using topographic maps to locate them (see Rose and Letzer, 1976). Second, use of just the mean nearest-neighbor distance means that any patterns examined are only those at short ranges. It may well be that at a larger scale, nonrandom patterns would have been detected. Finally, it is clear that the nearest-neighbor tests used by all these early workers show major dependence on the boundaries of the study region chosen for analysis. If you examine how *R* is calculated, you will see that by varying the area *A* used to estimate the intensity, it is possible to get almost any value for the index! Drumlins may well be distributed randomly, but the original papers neither proved nor disproved this. Better evidence about their distribution might be obtained by, for example, use of plots of the *G*, *F*, and *K* functions.

There are a number of complications with this approach, of which, as the cautionary tale above shows, the definition of area *A* used in all the computations is one. Another is that the expected values are strictly correct only for an unbounded study area with no edge effects.

The *G* and *F* Functions

Expected values of the *G* and *F* functions under IRP/CSR have also been determined. These both have the same well-defined functional form given by

$$\begin{aligned} E(G(d)) &= 1 - e^{-\lambda\pi d^2} \\ E(F(d)) &= 1 - e^{-\lambda\pi d^2} \end{aligned} \tag{5.12}$$

It is instructive to note why the two functions have the same expected form for a random point pattern. This is because, for a pattern generated by IRP/CSR, the events used in the *G* function, and the random point set used in the *F* function, are effectively equivalent—since they are both random. In either case, the predicted function may be plotted on the same axes as the observed *G* and *F* functions. Comparison of the expected and observed functions provides information on how unusual the observed pattern is. For the examples of clustered and evenly spaced arrangements considered previously (see Figure 5.10), this is plotted as Figure 5.15. In each plot, the expected function is the smooth curve between the two observed empirical curves.

In each case, the *G* and *F* functions lie on *opposite* sides of the expected curve. For a clustered pattern, the *G* function reveals that events in the pattern are closer together than expected under IRP/CSR, whereas the *F* function shows that typical locations in the study region are farther from any event in the pattern than would be expected (because they are empty). For the evenly spaced pattern, the opposite is the case. The *G* function clearly shows that an evenly spaced pattern has much greater nearest-neighbor distances than would be expected from a realization of IRP/CSR, while for the *F* function, because of the even spacing, typical locations in the empty

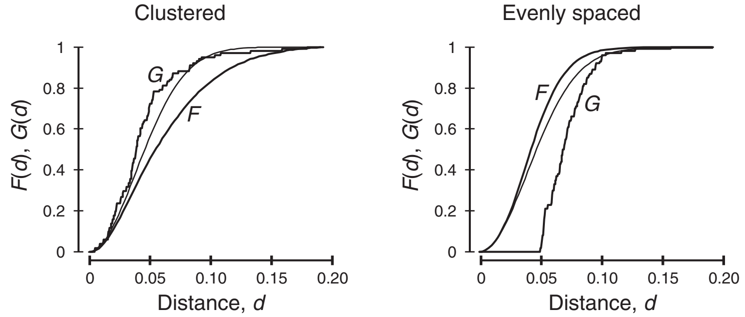


Figure 5.15 Comparison of the *G* and *F* functions for the patterns in Figure 5.10 against IRP/CSR. The middle curve in each plot is the expected value for both functions.

space are nearer to an event in the pattern than would be expected under IRP/CSR.

Again, these results can be made more precise by being given statements of probability or significance but it should be noted that all distance-based methods are subject to the same considerable problem: they are sensitive to changes in the study region. This affects estimation of λ, which must be used to determine the expected functions. Although the mathematics required is rather involved, it is possible to correct for edge effects. In practice, it is often more fruitful to use computer simulation to develop a "synthetic" prediction for the expected value of the descriptive measure of interest. This is discussed in more detail in connection with the *K* function.

The *K* Function

The expected value of the *K* function under IRP/CSR is easily determined. Since $K(d)$ describes the average number of events inside a circle of radius d centered on an event, for an IRP/CSR pattern we expect this to be directly dependent on d. Since πd^2 is the area of each circle and λ is the mean density of events per unit area, the expected value of $K(d)$ is

$$\begin{aligned} E(K(d)) &= \frac{\lambda \pi d^2}{\lambda} \\ &= \pi d^2 \end{aligned} \tag{5.13}$$

We can plot this curve on the same axes as an observed *K* function in much the same way as for the *G* and *F* functions. However, because the expected function depends on distance *squared*, both the expected and observed $K(d)$ functions become large as d increases. As a result, it is difficult to see small differences between expected and observed values when they are plotted on appropriately scaled axes.

One way around this problem is to calculate functions derived from K that have zero value if K is well matched to the expected value. For example, to convert the expected value of $K(d)$ to zero, we can divide by π, take the square root, and subtract d. If the pattern conforms to IRP/CSR, and if we perform the same operation on the observed values of $K(d)$, we should get values near zero. Perhaps unsurprisingly, this function is excitingly termed the *L function*

$$L(d) = \sqrt{\frac{K(d)}{\pi}} - d \tag{5.14}$$

and is plotted in Figure 5.16 for two well-known spatial data sets: Numata's Japanese pines data (Numata, 1961; Diggle, 2003) and Ripley's subset (1977) of Strauss's redwood seedlings data (Strauss, 1975). The former data set is indistinguishable from a random pattern, while the latter is clustered.

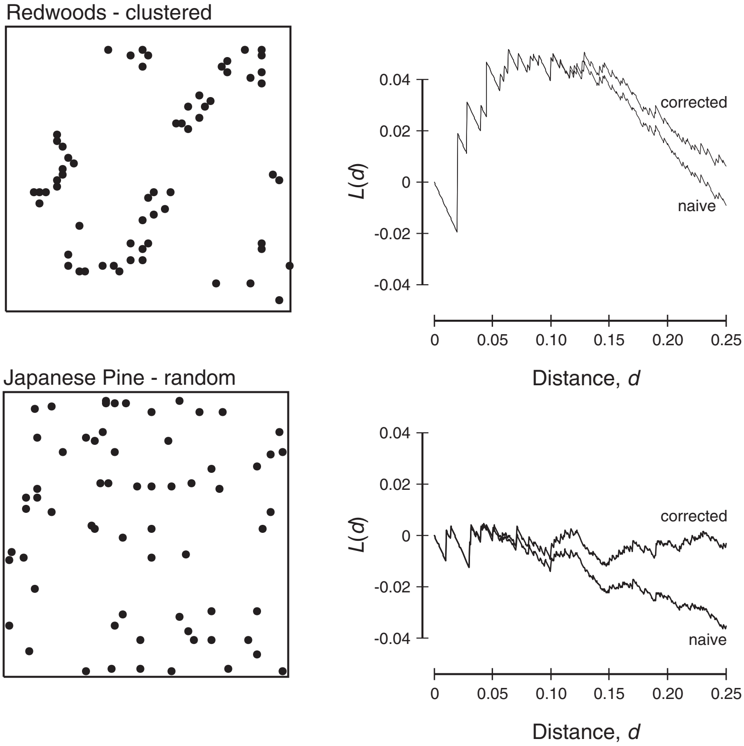


Figure 5.16 *L* functions, naive and corrected for two data sets.

Where $L(d)$ is above zero, there are more events at the corresponding spacing than would be expected under IRP/CSR; where it is below zero, there are fewer events than expected. In the case of the redwoods data, $L(d) > 0$ for d values across the whole range of plotted values, indicating that there are more events at these spacings than expected under IRP/CSR. For the Japanese pine, $L(d)$ is close to zero, until around $d = 0.1$ and then begins to fall continuously.

However, interpretation of the naive L functions is often made difficult by edge effects. In the two cases shown above, where $d \approx 0.1$, $L(d)$ falls continuously. This suggests that there are fewer events at these separations than expected. However, this is simply because many of the circles used in determining the K function at such distances *extend outside the study region* (which is a square of unit size). It is possible to correct calculation of the K and L functions to account for such edge effects, although the mathematics required is complex. In Figure 5.16, the corrected functions have been calculated using Ripley's isotropic correction (see Ripley, 1988). We may also consider using a guard zone, as illustrated in Figure 5.13, or any of the edge correction methods discussed in Section 5.2.

5.4. MONTE CARLO TESTING

While the above plots give us an idea of whether or not, and over what ranges of distances, a pattern is clustered or not, they are still not a statistical assessment of the data, because it remains unclear how far the L functions should depart from zero before we judge them to be unusually high or low values. While analytical results for the range of expected values are available in some cases, it is generally considered more straightforward to use computer simulation to estimate appropriate values First, read the simple example presented in the following box.

Nearest-Neighbor Distances for 12 Events

Figure 5.7 and Table 5.2 presented a simple pattern of 12 events in a 100 by 100 region. Our measured mean nearest-neighbor distance turned out to be 21.62. If we hypothesize that these events are the result of IRP/CSR, what values of $\bar{d}_{\min}$ do we expect?

One way to answer this question would be to do an experiment where we place 12 events in the same region, using random numbers to locate each event, and calculate the value of $\bar{d}_{\min}$ for that pattern. This gives just one value, but what if we do it again? Given the random process we are using, won't we get a different answer? So, why not do the experiment over and over again? The result will be a frequency distribution, called the *sampling distribution*, of $\bar{d}_{\min}$. The more times we repeat the experiment, the more values we get and more we can refine the simulated sampling distribution.

Relax, we are not going to ask you to do this (although it could be done as a class experiment with a large enough class). Instead, we've done it for you. Because the computer feels no pain when asked to do this sort of thing, we've repeated the experiment 1,000 times. The resulting frequency distribution of outcomes for the mean nearest-neighbor distance is shown in Figure 5.17.

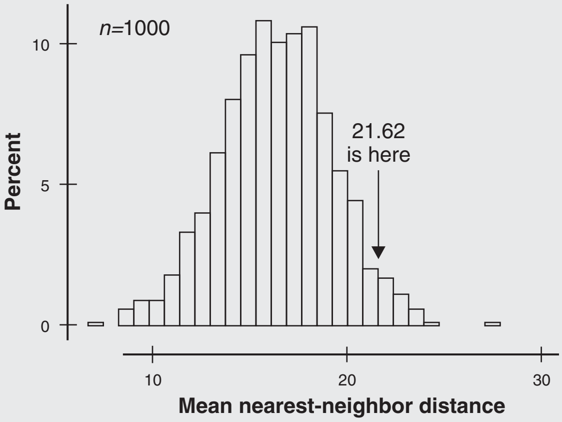


Figure 5.17 Results of a simulation of IRP/CSR for 12 events (compare Table 5.2).

The simulated sampling distribution is roughly normal with mean $\bar{d}_{\min} =$ 16.50 and with a standard deviation 2.93. The observed value given for the pattern, at 21.62, lies some way above this mean and so can be seen to be a moderately uncommon realization, albeit within 2 standard deviations of the mean. Note that in 1000 simulations, the range of values for $\bar{d}_{\min}$ was considerable, from 7.04 to 27.50, and that the theoretical value for an unbounded study area is actually 14.43.

In exactly the same way that we can simulate a point pattern to determine the distribution of a simple statistic such as the mean nearest-neighbor distance, we can simulate to determine the distribution of much more complex measures such as *K(d)* or its associated *L* functions. The procedure is exactly the same: use a computer to generate patterns and measure the quantity of interest each time, generating an expected distribution of values. This approach also allows us to neatly take care of problems like edge effects simply by using the same study region in the simulations as in our observed data. Since each simulation is subject to the same edge effects as the observed data, the sampling distribution we obtain automatically accounts for edge effects without complex adjustments to the calculation. Such a simulation approach is known as a *Monte Carlo procedure* and is widely used in modern statistics.

Typically, a Monte Carlo procedure is used randomly to locate n events in the study area A, perhaps 100 or 500 or, as we did to create Figure 5.17, 1,000 times. Each randomly generated point pattern is then analyzed using the same methods applied to the pattern under investigation. Results for the randomly generated patterns can then be used to construct an *envelope* of results, inside which a pattern generated by IRP/CSR would be expected to sit. Depending on how many simulated patterns are generated, reasonably accurate confidence intervals can be placed on the envelope, so we can determine how unusual the observed pattern is with reasonable accuracy. One of the most frequently used freeware computer programs for point pattern analysis, CrimeStat III (Levine and Associates, 2007), has such a simulation facility built in. Similarly, the *R* statistics package, and its **spatstat** library for point pattern analysis (Baddeley and Turner, 2005), can generate simulation envelopes for any of the many point pattern measures it can calculate. Given the analytical difficulties involved in dealing with many point pattern measures, this approach is becoming increasingly common.

Results for a simulation analysis using 99 simulations for the point data sets from Figure 5.16 are shown in Figure 5.18. In these diagrams, interpretation is much clearer. For the redwood data, it is apparent that over a distance range from around 0.02 to 0.2, the observed pattern is more clustered than we would expect it to be were it generated by IRP/CSR. Even more clearly, in the second panel, we can see that at all distances, the observed L function for the Japanese pine data set lies inside the simulation envelope generated by IRP/CSR, so we must conclude that in terms of the L function at least, this pattern is entirely typical of what we would expect for a pattern produced by IRP/CSR.

The Monte Carlo simulation approach has a number of clear advantages:

- There is no need for complex corrections for edge and study region area effects (although note that, as long as the same calculation is applied to the measurement of both the observed data and the simulated data, any desired correction can be incorporated).
- Although the procedure works by using the same number of events n as in the original pattern, it is not so sensitively dependent on this

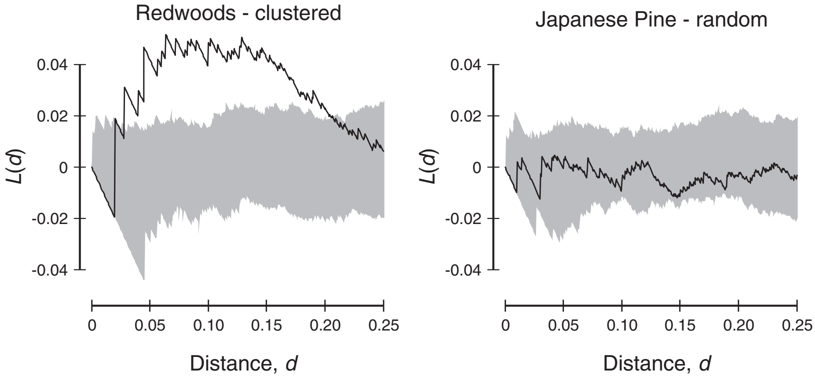


Figure 5.18 *L* functions plotted with simulation envelopes produced by 99 simulation runs.

choice as are approaches based on an equation that includes λ. It is also easy to gauge the importance of this assumption by varying n in the simulated patterns.

- Perhaps the most important advantage of the approach is that spatial process models other than IRP/CSR may be conveniently investigated; indeed, *any* process that fits *any* theory we might have about the data can be simulated and an assessment made of the observed pattern against the theory. This allows us to move beyond investigation of only analytically simple process models. Statistical assessment of measures such as the pair correlation function depends heavily on simulation approaches (see Perry et al., 2006).

A disadvantage of simulation is that it may be computationally intensive. For example, if there are 100 events in the original point pattern and (say) a $p = 0.01$ confidence level is required, it is necessary to run *at least* 99 simulations. Each simulation requires 100 events to be generated. For the K function, distances between all 100 events in each simulation must then be calculated. This involves approximately $100 \times 99 \times 99/2 \approx 500\ 000$ basic calculations. Each distance determination involves two subtraction operations (the difference in the coordinates), two multiplications (squaring the coordinate differences), an addition, and a square root operation—six in total—and the square root operation is not simple. That's a total of 3 million mathematical operations, followed by further sorting operations to build the upper and lower bounds of the envelope.

Nevertheless, this sort of calculation is well within the capacity of any modern desktop computer. The graphs in Figure 5.18 took only a few seconds to produce on a standard 2008-vintage laptop. Even so, just because it is possible does not mean that it is always necessary to proceed in this way. It is important to be sure that it is worthwhile to apply statistics to the problem at all before embarking on such complex analysis. Perhaps using the point pattern measures in a descriptive manner is adequate for the task at hand. On the other hand, if it is important that your analysis is right—for example, in the detection of disease hot spots—then leaving a machine running for an hour or two (on a large problem) may be a small price to pay for the gain in knowledge obtained from the simulations.

5.5. CONCLUSIONS

A lot of detailed material has been presented in this chapter, but the basic messages are readily apparent. We have been concerned with developing a clearer idea of the concept of pattern and how it can be related to process. In principle, any pattern can be described using a variety of measures based on its first- and second-order properties—or, put another way, by looking for departures from first- and second-order stationarity.

In a point pattern, first- and second-order variation can be directly related to two distinct classes of pattern measure: density-based measures and distance-based measures. Among density-based measures, quadrat counts and kernel-density estimation provide alternative solutions to the problem of the sensitivity of any density measurement to variation in the study area. Numerous distance-based measures are also available, from the very simple mean nearest-neighbor distance, through G and F functions, to the full complexity of the K and pair correlation functions, which use information about *all* the interevent distances in a pattern. Perhaps the most important point to absorb is that as a minimum, some preliminary exploration, description, and analysis using these or other measures is likely to be useful. For example, a kernel-density estimated surface derived from a point pattern is helpful in identifying the regions of greatest concentration of a phenomenon of interest, while the G, F, and K functions together may help identify characteristic distances in a pattern—particularly intra- and intercluster distances. In many cases, such information is useful in itself.

However, if we wish, we can go further and determine how well a pattern matches what we would expect if the pattern were a realization of a particular spatial process that interests us. This involves determining for the process in question the sampling distribution for the pattern measure we wish to use. The sampling distribution may be determined either analytically (as in Chapter 4) or by simulation (as in this chapter). Having done this, we can set up and test a null hypothesis that the observed pattern is a realization of the process in question. Our conclusion is either that the pattern is very unlikely to have been produced by the hypothesized process or that there is no strong evidence to suggest that the pattern was not produced by the hypothesized process. Either way, we cannot be sure—that's statistics for you—but we can assign some probabilities to our conclusions, which may represent a useful advance over simple description.

So, what to make of all this? We've come a long way in a short time: has it been worth it? Does comparing a pattern to some spatial process model really help us understand the geography better? This is the core of spatial statistical analysis, and the question we have asked is a very pure one: whether or not an observed point pattern is or is not an unusual realization of IRP/CSR. Most of the time when dealing with point patterns, this isn't the only hypothesis we want to test. Does being able to make the comparison statistical really help? Is it useful to know that "there is only a 5% chance that this pattern arose from this process by chance"? The answer is, "it depends", but experience suggests that practical problems of the type you may be asked to address using a GIS are rarely capable of being solved using pure spatial pattern analysis methods.

The statistical approach becomes important if we want to use spatial patterns as evidence in making important decisions. In the world we live in, *important* usually means decisions that affect large numbers of people or large numbers of dollars—frequently in opposition to one another. A classic example is the conflict of interest between a residential community that suspects that the polluting activities of a large corporation are responsible for apparently high local rates of occurrence of a fatal disease. IRP/CSR is a process model of only limited usefulness in this context. We know that a disease is unlikely to be completely random spatially, because the population is not distributed evenly, and we expect to observe more cases of a disease in cities than in rural areas. In epidemiology, the jargon is that the "at-risk" population is not evenly distributed. Therefore, to apply statistical tests, we have to compare the observed distribution to the at-risk population distribution. A good example of what can be done in these circumstances is provided by the case studies in the paper by Gatrell et al. (1996). Using the simulation approach discussed above, we can create a set of simulated point patterns for cases of the disease based on the at-risk population density, and then make comparisons between the observed disease incidence point pattern and the simulation results using one or more of the methods we have discussed.

In short, even the complex ideas we have discussed in detail in this chapter and the previous one are not the whole story. Some other, more practical issues that might be encountered in this example or other real-world cases are discussed in the next chapter.

CHAPTER REVIEW

- A point pattern consists of a set of *events* at a set of locations in the study region, where each event represents a single instance of the phenomenon of interest.
- We describe a point pattern using various *measures* or *statistics*. The simplest measure is the mean location and standard distance, which can be used to draw a summary circle or ellipse. However, this measure discards most of the information about the pattern, so it is useful only for an initial comparison of different patterns or for recording change in a pattern over time.
- Measures of pattern are broadly of two types: *density measures*, which are *first-order* measures, and *distance measures*, which are *second-order* measures.
- Simple density is not very useful. *Quadrat counts* based on either a census or a sample of quadrats provide a good, simple summary of a point pattern's distribution.
- The simplest distance measure is the nearest-neighbor distance, which records for each event the distance to its nearest neighbor also in the pattern.
- Other distance functions are the *G, F, K,* and pair-correlation functions, which use more of the interevent distances in the pattern to enhance their descriptive power, although interpretation may be difficult.
- If we feel it is necessary to conduct a formal statistical analysis, the general strategy is to compare what is observed with the distribution predicted by a hypothesized spatial process, of which the IRP/CSR is by far the most often used. Tests are available for all the pattern measures discussed.
- In practice, edge effects, and their sensitivity to the estimated intensity of the process, mean that many of these tests are difficult to apply, so computer simulation is often preferred.

REFERENCES

Baddeley, A. and Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12 (6): 1–42.

Bailey, T. C., and Gatrell, A. C. (1995) *Interactive Spatial Data Analysis* (Harlow, England: Addison Wesley Longman).

Clark, P. J. and Evans, F. C. (1954) Distance to nearest neighbour as a measure of spatial relationships in populations. *Ecology*, 35: 445–453.

Crozier, M. J. (1976) On the origin of the Peterborough drumlin field: testing the dilatancy theory. *Canadian Geographer* 19: 181–195.

Davis, J. H., Howe, R. W., and Davis, G. J. (2000) A multi-scale spatial analysis method for point data. *Landscape Ecology*, 15: 99–114.

Diggle, P. (2003) *Statistical Analysis of Spatial Point Patterns* (London: Arnold).

Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S. (1996) Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, NS 21: 256–274.

Greig-Smith, P. (1964), *Quantitative Plant Ecology* (London: Butterworths).

Grünbaum, B. and Shephard, G. C. (1987) *Tilings and Patterns* (New York: W. H. Freeman).

Levine, N.and Associates (2007) *CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Locations* (available at http://www.icpsr.umich.edu/CRIMESTAT/).

Numata, M. (1961) Forest vegetation in the vicinity of Choshi. Coastal flora and vegetation at Choshi, Chiba Prefecture. IV. *Bulletin of Choshi Marine Laboratory*, Chiba University 3, 28–48 (in Japanese).

Perry, G. L. W., Miller, B. P., and Enright, N. J. (2006) A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecology*, 187: 59–82.

Ripley, B. D. (1976) The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13: 255–266.

Ripley, B. D. (1977) Modelling spatial patterns. *Journal of the Royal Statistical Society, Series B*, 39: 172–212.

Ripley, B. D. (1988) *Statistical Inference for Spatial Processes* (Cambridge: Cambridge University Press).

Rogers, A. (1974) *Statistical Analysis of Spatial Dispersion* (London: Pion).

Rose, J. and Letzer, J. M. (1976) Drumlin measurements: a test of the reliability of data derived from 1:25,000 scale topographic maps. *Geological Magazine*, 112: 361–371.

Smalley, I. J. and Unwin, D. J. (1968) The formation and shape of drumlins and their distribution and orientation in drumlin fields. *Journal of Glaciology*, 7: 377–390.

Strauss, D. J. (1975) A model for clustering. *Biometrika*, 63: 467–475.

Thomas, R. W. (1977) An introduction to quadrat analysis. *Concepts and Techniques in Modern Geography*, 12, 41 pages (Norwich, England: Geo Books). Available at http://www.qmrg.org.uk/catmog.

Thompson, H. R. (1956) Distribution of distance to nth neighbour in a population of randomly distributed individuals. *Ecology*, 37: 391–394.

Trenhaile, A. S. (1971) Drumlins, their distribution and morphology. *Canadian Ceographer*, 15: 113–26.

Trenhaile, A. S. (1975) The morphology of a drumlin field. *Annals of the Association of American Geographers*, 65: 297–312.

Wiegand, T. and Moloney, K. A., (2004) Rings, circles and null models for point patterns analysis in ecology. *Oikos*, 1104: 209–229.

Yamada, I. and Rogerson, P. A. (2003) An empirical comparison of edge effect correction methods applied to *K*-function analysis. *Geographical Analysis*, 37: 95–109.

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 6

Practical Point Pattern Analysis

CHAPTER OBJECTIVES

In this chapter, we:

- Review briefly two of the most cogent geographic critiques of this approach to spatial statistical analysis
- Point out that, because of the unrealistic nature of the assumptions made in its derivation, the basic homogeneous Poisson process that we have outlined so far is not often useful in practice
- Describe point process models used as *alternatives to IRP/CSR*
- Outline the controversy over childhood cancer linked to a nuclear reprocessing plant in northwest England, a classic example of spatial point pattern analysis in action; in turn, this problem illustrates a series of issues that almost invariably arise in practical point pattern analysis
- Examine methods that tackle the first of these, the presence of inhomogeneity or heterogeneity, which makes the homogeneous Poisson process, our standard IRP/CSR model, inappropriate for most analyses
- Review approaches to problems involving some hypothesized source, or sources, around which the events might cluster, necessitating what is called a *focused* test
- Note that we are frequently concerned not with a general test that declares a pattern to be clustered or more regular than random, but with *detecting and locating significant clusters*
- Discuss the *Geographical Analysis Machine* (GAM) developed by Openshaw and his colleagues, an example of a system that attempted to address all of these issues simultaneously
- Suggest that the use of *proximity polygons*, a standard GIS geometric transformation of the type suggested in Section 2.3, might enable better characterization of point patterns for analysis

- Note that many of the operations we discuss can be accomplished using simple interevent *distance matrices*.

After reading this chapter, you should be able to:

- Outline the basis of classic critiques of spatial statistical analysis in the context of point pattern analysis and articulate your own views on the issues raised
- Explain why IRP/CSR is usually an unrealistic starting point in spatial point pattern analysis
- Distinguish between *general*, *focused*, and *scan* approaches to spatial point pattern analysis
- Discuss the merits of point pattern analysis in cluster detection and outline the issues involved in real-world applications of these methods
- Outline how proximity polygons could be used in point pattern analysis
- Assemble a simple interevent distance matrix and show how it could be used in point pattern analysis

6.1. INTRODUCTION: PROBLEMS OF SPATIAL STATISTICAL ANALYSIS

In our account of classical spatial statistical analysis in Chapters 4 and 5, we avoided considering its limitations in any extended way. Nevertheless, there are difficulties with the approach, both in aspects of the mode of statistical inference used and in the details of its application to real-world questions and problems. In this chapter, we consider these issues more closely. Much of our treatment focuses on a particular application: identifying clusters of a rare disease. However, before considering the particulars of that application, it is instructive to review effective critiques of spatial analysis from two eminent geographers. Although their views were presented 40 or so years ago, they remain highly relevant and provide a useful introduction to our development of concepts behind contemporary spatial analysis.

Peter Gould's Critique

In a paper entitled “Is *statistix inferens* the geographical name for a wild goose?” Peter Gould (1970) made a number of important criticisms of the use of inferential statistics in geography, and it is good to be aware of them. In summary, Gould suggests the following:

- Geographic data sets are not samples.
- Geographic data are almost never random.
- Because of autocorrelation, geographic data are not independent random.
- Because n is always large, we will almost always find that our results are statistically significant.
- What matters is scientific significance, not statistical significance.

Now that you have read our account, we hope that you will answer these criticisms more or less as follows:

- The process-realization approach enables us to view geographic data as samples in a particular way.
- There is no answer to this criticism; geographic data are not random. The real question is whether or not it is scientifically *useful* to analyze geographic data as if they were the result of a stochastic (random) process; the answer to this question must be "yes."
- Even though data are not independent random, this does not prevent us from using statistics if we can develop better models than IRP/CSR. This is discussed further in Section 6.3.
- Often n is large, but not always, so this point is not convincing, particularly if a commonsense approach to the interpretation of statistical significance is used.
- Gould is right. *Scientific significance is the important thing*. This requires that we have a theory about what is going on and test that theory appropriately, not just use whatever statistics come to hand.

Perhaps the most important point implied by Gould's criticisms is that IRP/CSR is a rather strange hypothesis for geographers to test against. After all, it suggests that the geography makes no difference, something that we don't believe from the outset! *The whole point of IRP/CSR is that it exhibits no first- or second-order effects*, and these are precisely the types of effects that make geography worth studying. In other words, we'd be disappointed if our null hypothesis (IRP/CSR) were ever confirmed, and it turns out that for large n it almost never is. Furthermore, rejecting IRP/CSR tells us virtually, nothing about the process that actually *is* operating. This is a difficulty with inferential statistics applied to spatial processes: whereas a null hypothesis like "Mean tree height is greater than 50 m" has an obvious and meaningful alternative hypothesis ("Mean tree height is less than or equal to 50 m"), IRP/CSR as a null hypothesis admits *any* other process. Thus, rejecting the null hypothesis is arguably not very useful.

David Harvey's Critique

In two papers published many years ago, David Harvey (1966, 1968) also discussed some of these issues. His major point is irrefutable and simple but very important: there are inherent contradictions and circularities in the classical statistical approach we have outlined. Typically, in testing against some process model, we estimate key parameters from our data (often, for point pattern analysis, the intensity λ). The estimated parameters turn out to have strong effects on our conclusions, so much so that we can often conclude anything we like by altering the parameter estimates—which can usually be done by altering the study region. Modern simulation approaches are less prone to this problem, but the choice of the study region remains crucial.

Harvey's disillusionment and frustration with spatial analytic approaches to the understanding of geographic phenomena has to be taken seriously and not dismissed as a failure to understand the approach. On the contrary, it is important to acknowledge that, ultimately, spatial analysis methods are limited in their ability to prove *anything*. This is a point in the philosophy of science that deals with how we assess evidence in developing and establishing scientific theories. Harvey's critique points to the importance of theories for explaining our observations. We hope that we have made it clear that spatial analysis has an important role to play in establishing how well any particular theory about geographic processes fits the evidence of observational data. Both theory and appropriate methods are necessary for advancing understanding. The limitations Harvey notes can be addressed by appropriate and thoughtful use of spatial statistical analysis.

Implications

Gould's critiques point to two important weaknesses in the classical approach we have outlined. First, a hypothesis-testing approach in which we reject the null hypothesis of IRP/CSR tells us nothing we didn't already know; it also fails to tell us anything about the processes that *are* operating. This is a failure associated with the particular approach to statistical reasoning that we have presented. Alternative approaches are available and are becoming increasingly widely used; they are discussed in Section 6.2. The second weakness Gould's critique emphasized is how poor a process model IRP/CSR is. As we hinted in Section 5.4 in considering simulation approaches, it is becoming increasingly common to consider alternative models. We briefly discuss some of these models in Section 6.3.

6.2. ALTERNATIVES TO CLASSICAL STATISTICAL INFERENCE

Some of Gould's and Harvey's concerns are partly addressed by more recent developments in the application of statistical analysis to point patterns. Many of these developments relate to other approaches to statistical inference other than the classical approach that we focused on in earlier chapters. In any case, as you explore spatial analysis further, you will almost certainly encounter these other approaches to statistical inference. It is therefore worthwhile to outline them (very briefly) here.

In Chapters 4 and 5, we presented a perspective based on *classical statistical inference*, sometimes referred to as *frequentism*. This approach is the one covered in almost any standard introduction to statistics. For a null statistical process (such as IRP/CSR) it asks, "How probable would the observed pattern be if the hypothesized null process were operating?" The outcome of such a *hypothesis test* is a p-value that leads us to either:

- *Reject the null hypothesis* if the p-value is low. We can then go on to conclude that our pattern is unlikely to have been produced by the hypothesized process; or
- *Fail to reject the null hypothesis* if the p-value is high. In this case, we conclude that there is insufficient evidence to believe that the observed pattern is not the outcome of the hypothesized process.

Potentially, we could apply this approach sequentially to a series of alternative process models, that is, not just to the independent random process. Thus, we might reject IRP/CSR, and then present a different possible process model and determine if we can reject it as a candidate generating process for our data. Quite apart from the evident difficulty (where do we stop?), this is rarely done. The net result is that classical inference often doesn't allow us to conclude much more than "The data are more clustered than random" or, worse, "The data aren't obviously nonrandom." Neither conclusion is much use beyond the very earliest stages of an investigation.

An approach that is becoming more widely used, and that can allow us to make more progress, is *likelihood-based inference* (see Edwards, 1992). Here, the idea is to use statistical analysis of point pattern measures to assess which of a number of alternative process models is the most likely to have produced the observed pattern. Roughly speaking, this involves assessing, for an observed pattern, which of several alternative process models it most typifies. Whereas classical inference assigns probability to the data in light of a hypothesized process, likelihood statistics assign *likelihood* to each of a number of possible processes given the observed data. For many scientists, this is a more satisfying way to pose the problem of assessing theories based on evidence.

The technical details of likelihood statistics are rather complex and beyond the scope of this book. In outline, the procedure involves first estimating parameters of the alternative process models of the chosen types (based on the observed patterns) and then determining how likely each fitted process model is to have produced the observed pattern. Computer simulation is invariably used. The attraction of the likelihood approach is that it allows us to do more than simply reject IRP/CSR. Potentially, we can draw conclusions about which of several models is most likely to be operating. In the next section, we describe some alternative process models beyond IRP/CSR that may be considered.

Considerable care is required in applying the likelihood approach. Careful specification of the alternative models for consideration is important. It is also important to realize that just because one model is the most likely, this does not preclude the possibility either that *none* of the alternative models is much good or that the differences between the alternatives are not sufficiently marked to justify a strong preference for one model over others. Judicious use of visualizations can be very important in this context to avoid getting carried away with the numerical statistical outputs and placing undue emphasis on the mechanical choice of the "best" model.

Finally, in situations where previous research or prior knowledge strongly suggests that a particular spatial process is in operation, many would advocate a *Bayesian* approach (see Bolstad, 2007). Here, the idea is to use observational data to refine a preexisting statistical model of the process. Although the philosophical underpinnings of the Bayesian approach are attractive to many (not all!) scientists, the statistical analysis of point patterns is not an area where such methods have been widely developed or adopted to date.

Either or both likelihood and Bayesian approaches, in combination with careful use of visualization and (yes) classical hypothesis testing, can allow us to advance far beyond the rather mindless rejection of implausible null hypotheses (rightly) decried by Gould and Harvey. You will almost certainly find yourself using such tools in more advanced work on point pattern analysis.

6.3. ALTERNATIVES TO IRP/CSR

By now, you are probably impatient to know more about the alternatives to IRP/CSR we keep mentioning. Before describing some of these alternatives, it is useful to consider their essential characteristics. As we have mentioned, IRP/CSR is an ideal null process, exhibiting no first- or second-order effects. Alternative processes, then, are simply ones that introduce either or both. We have already noted that it is difficult to distinguish first- and second-order effects in data, but it is possible to devise process models in which the distinction is very clear. First-order effects involve allowing the probability of events to vary from place to place across the study region, while second-order effects involve introducing interactions of some kind between events.

Pause for Thought

Before reading any further, can you think of some distributions of point events where we would expect either first-order intensity variations or second-order interaction effects? For each of the following point object types, suggest what actual mechanisms might introduce such effects into the patterns we observe:

- Trees in a wood
- Cases of asthma among children
- Recorded burglaries in a city
- Rain gauges in a hydrological observation network
- Houses of the customers of a large store
- Domestic fire incidents
- Automobile accidents across a county

Two example applications that are generally easily appreciated are to be found in spatial epidemiology and in the study of distributions of plant species in plant ecology.

In either case, we may expect first-order variations. Home addresses of individual cases of a disease can be expected to vary with the spatial density of the at-risk population. Where more about the risk factors associated with a disease is known, we may be able to introduce further inhomogeneity into our first-order model based on the presence of different population subgroups, housing types, and so on. There will be more cases where there are more people at risk. Similarly, varying suitability either in the physical landscape (land elevation, slope, etc.), climatic conditions (precipitation, solar radiation, etc.), or soil (acidity, chemistry, granularity, etc.) can be expected to produce first-order variation in the observed spatial distribution of a particular species. If you think about it, you will see that these cases are not as different as they may at first appear.

Similarly, both disease and plant distributions might be expected to exhibit second-order interaction effects. To the extent that a disease is infectious, then, given one case in a neighborhood, we might expect further cases to be more common in the same neighborhood rather than in others. Plant distributions provide an even more clear-cut case here. Seed dispersal and vegetative spread (i.e., via root systems) are generally highly localized processes, so that, on average, anywhere we find an individual member of a species, we can expect to find more of the same. So, how are these effects incorporated into process models? A few simple examples are considered below.

The *inhomogeneous Poisson distribution* is a simple extension of the homogeneous Poisson process (which we have been calling IRP/CSR). Instead of assuming a spatially uniform intensity λ, we allow the intensity to vary from place to place. This is illustrated in Figure 6.1, where three realizations are shown. This first of these is a homogeneous case with $\lambda = 100$ across the whole study area. The second and third cases introduce spatial variation in λ indicated by the shading and contours of the variation of intensity. In both cases, the range of values is from 100 to 200, with contours from 110 to 190 at intervals of 10. It is noteworthy that, in spite of a twofold variation in intensity across the region, neither realization is very evidently different from the homogeneous case. There is some impression of an absence of events in the northwest quadrant of the third pattern, but there is little to suggest to the human eye that there is a higher probability of events falling in the center of this pattern. Unless there are strongly marked variations in intensity, this is quite typical of more complex process models.

A process that introduces second-order effects is the *Thomas process* or *Poisson clustering process*. Here, a simple Poisson process (which may be inhomogeneous) produces "parent" events. Each parent then produces a random number of "children" placed around the parent at random. The parent events are then removed to leave the final pattern. Three

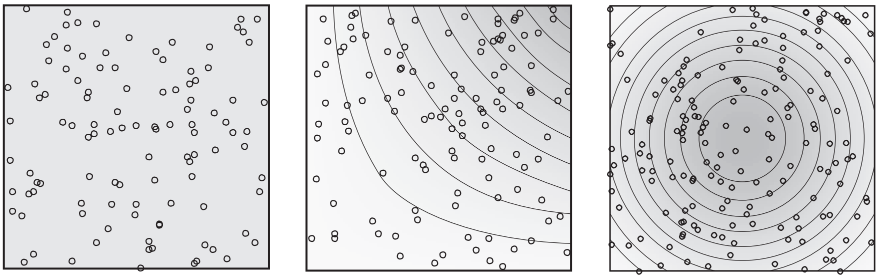


Figure 6.1 Three Poisson process realizations. See text for details.

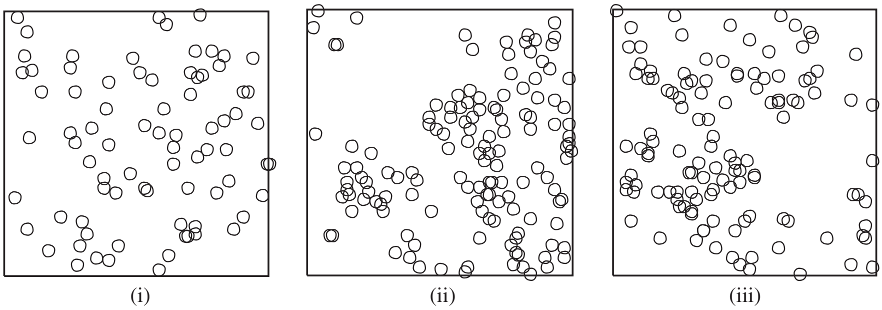


Figure 6.2 Three realizations of the Thomas process in a unit square. Parameters settings are (i) $\lambda = 10$, $\mu = 10$, $\sigma = 0.3$, (ii) $\lambda = 10$, $\mu = 10$, $\sigma = 0.1$, and (iii) $\lambda = 20$, $\mu = 5$, $\sigma = 0.1$.

realizations are shown in Figure 6.2. Three parameters are required to specify this process, namely, the intensity of the original distribution of the parents (λ), the number of children of each parent (μ, itself the mean intensity of a Poisson distribution), and the characteristics of the dispersal of children from the parent locations. Usually the last step proceeds according to a Gaussian kernel, so it is unlikely that children will be very far from the parent locations and the standard deviation of the kernel must be specified. Again, it is noteworthy that one of these patterns (in Figure 6.2i) is not obviously distinguishable from IRP/CSR, at least by eye.

With appropriate choice of parameters, the Thomas process produces patterns that are quite markedly clustered—even by eye. Other process models use *packing constraints* or *inhibition* to make it unlikely that events will occur closer together than some minimum threshold distance, and these produce evenly spaced or dispersed patterns.

From an analysis perspective, the most troublesome feature of any of these process models is that they introduce (perhaps many) more parameters that must be estimated from the observed data before statistical analysis can proceed. Thus, in the same way that we use a simple estimate of a pattern's intensity to condition subsequent statistical analysis for IRP/CSR, with these more complex processes it is necessary to estimate additional parameters from the data before we can apply any statistical procedure. For inhomogeneous processes, a common procedure is to use a kernel density estimate for the spatially varying process intensity. Estimation of other parameters may involve complex statistical fitting procedures. Once a number of estimated best-fit process models have been derived, statistical testing using likelihood methods may allow some assessment of which of a number of models appears to account best for the observed data. Perry et al. (2008) is an example of such a study, which gives a good sense of just how much work is involved in systematically pursuing this approach.

6.4. POINT PATTERN ANALYSIS IN THE REAL WORLD

In practice, such comprehensive analysis and statistical testing of observational data with respect to empirically derived best-fit point process models remains unusual, even in the research literature.

While we can agree that in many cases a more complex process model than IRP/CSR is clearly required, it is time to take a step back again. Assuming that we have adopted some more or less complex model as the most plausible for our data, whether on theoretical grounds or based on model-fitting, we can, as before, develop some expectation of the values of various pattern metrics. But what would we conclude from a study comparing observed disease cases to an inhomogeneous Poisson clustering process (or some other model)? We might infer either that the pattern of incidents matched the process well or not. In the latter case, we would be able to say that observed events were more—or less—clustered than we expected. While this knowledge may be useful, it is still limited in many practical situations. The approach is a *general* technique, concerned with the overall characteristics of a pattern (Besag and Newell, 1991).

There are at least two other geographic questions we might wish to ask of a point pattern. First, we might have a hypothesis that relates the clustering to either a single center or multiple centers. The classic example is the 1854 study by Dr. John Snow of the Soho (London) cholera outbreak, where an obvious pattern of deaths was hypothesized to cluster around a single source of infected water (see Johnson, 2006). Modern versions of this same problem are reviewed by Hills and Alexander (1989) and elsewhere (Diggle, 1990). This is a *focused* rather than a general problem.

Second, in their standard forms, neither general nor focused approaches say much about *where* the pattern deviates from expectations. All the concepts introduced in Chapters 4 and 5 omit this important aspect of point pattern analysis entirely. Today, the problem of *cluster detection* is usually addressed by some form of *scan statistic.*

Thus, we have three related further sets of issues: correcting for inhomogeneity, testing for clustering relative to some assumed focal point, and detecting clusters.

Background: Cancer Clusters Around Nuclear Installations

In the remaining sections of this chapter, we illustrate practical issues in point pattern analysis using the example provided by a cluster of children's deaths from the cancer leukemia around the town of Seascale in northern England. According to Gardner (1989), Seascale saw four cases of childhood leukemia in the 1968–1982 period, whereas just 0.25 might have been expected in this small town. The hypothesis was that this apparent cluster of cases was related to the nearby Sellafield, Britain's oldest nuclear reactor and fuel reprocessing plant. Local family doctors had already expressed concern about higher than expected levels of leukemia, but in November 1983, a TV program *Windscale: The Nuclear Laundry* dramatized the possible link (Sellafield was previously named Windscale). The program was based on evidence assembled by Urquhart et al. (1984) and led to a major public and scientific controversy, a variety of detailed academic and medical studies, and a detailed official report (Black, 1984). The report concluded that the cluster was real, not mere chance, but that evidence linking the cluster to the plant was circumstantial.

Furthermore, there were reasons to doubt a *direct* link between leukemia deaths and ionizing radiation from the Sellafield plant:

- Measured levels of radiation in the area did not seem high enough to cause genetic damage.
- Apparent clusters occur naturally in many diseases for unexplained reasons. Meningitis is a good example of such a clustering disease.
- The actual number of cases in the cluster (four) was much too small to infer that it was unusual.
- If radiation were the cause, then one would expect some correlation in time between the operation of the plant and the occurrence of the cases. No such time correlation was found.
- Similar clusters of cancers have been found around nonnuclear plants and even at places where plants had been planned but were never built.
- Finally, many industries use a number of chemicals whose leukemogenic potential is poorly understood but which may be equally, or even more culpable.

The Black Report led to establishment of the Committee on Medical Aspects of Radiation in the Environment (COMARE), which studied another nuclear plant, at Dounreay in the far north of Scotland. COMARE found that there had been six cases of childhood leukemia around the Dounreay plant when only one would have been expected due to chance spatial variation. In 1987 a report in the *British Medical Journal* suggested that there was another cluster around the British Atomic Energy Research establishment at Aldermaston in southern England. All this research activity led to the publication of a special volume of the *Journal of the Royal Statistical Society, Series A* (1989, vol. 152), which provides a good perspective on many of the issues involved. The debate has rumbled on more or less continuously ever since in the epidemiology literature, with numerous alternative tests suggested (see, for example, Bithell and Stone, 1989; Bithell, 1990; Bithell et al., 2008). Whatever the cause, most studies conclude that there was a cluster of cancer cases at Seascale, but that there is no direct evidence of a causal link between leukemia and any excess ionizing radiation in the area.

If you are skeptical about that conclusion, it is useful to note that other hypotheses might also account for the cluster. Kinlen (1988) argued his *rural newcomer hypothesis*, stating that the cause was an unidentified infectious agent brought by construction workers and scientists moving into previously isolated areas such as those around Sellafield and Dounreay. The infectious agent, he suggested, triggered leukemia in a vulnerable host population that had not built up any resistance to it. Two years later, Gardner et al. (1990) completed a major study examining the family histories of those involved. They suggested that men who received cumulative lifetime doses of radiation greater than 100 mSv, especially if they had been exposed to radiation in the six months prior to conception, had six to eight times the chance of fathering a child who developed leukemia because of mutations to the sperm. However, they were unable to find any direct medical evidence in support of this hypothesis, and the theory seems counter to the results of trials involving the victims at Hiroshima and Nagasaki, which found no pronounced genetic transmission. However, geneticists have pointed out that common acute lymphatic leukemia is a cancer that is possibly transmitted in this way.

Whatever the underlying causes, the search for cancer clusters around Sellafield, Dounreay, and other nuclear installations illustrates well the three issues noted above: inhomogeneity, focused testing, and cluster detection. We consider each in more detail below.

6.5. DEALING WITH INHOMOGENEITY

The first issue that any analysis of this problem has to deal with is inhomogeneity. None of the classical tests outlined in Chapter 5 is appropriate if the background is spatially inhomogeneous. If we want to study possible clustering in a pattern of events such as childhood deaths from leukemia where we know that the at-risk population is not evenly distributed over the study region, what is required is a test for clustering against a Poisson process that allows for this variation. Numerous tests have been proposed, and several modifications of standard pure point pattern statistics have been suggested and used (Cuzick and Edwards, 1990). An excellent introduction to the issues in testing for clustering in the presence of inhomogeneity is the one by Gatrell et al., 1996). A recent review paper by Kulldorff (2006) lists over 100 tests of spatial randomness adjusted for inhomogeneity and cites over 150 references.

Approaches Based on Rates

The simplest approach, used by the first people to explore the problem around Sellafield, is to express the counts of events as some rate of incidence in the at-risk population. At first sight, this seems straightforward. All we need to do is determine the ratio of the observed events to the at-risk population. Of course, this is not so simple. The notion of an at-risk population implies that we have aggregate information for both the cases and the at-risk population over some appropriate spatial units. Almost certainly our base population figures will come from an official census of population, and typically these will be in regions delineated for the convenience of the census agency. They are unlikely to bear any relationship to the problem at hand, and often they may also refer to a particular "snapshot" of the population at a date some time before or after the disease incidents. Any rates we estimate will obviously be conditional on the units chosen, and the MAUP examined in Section 2.2 is an obvious and very unwelcome consequence.

The severity of the problem is illustrated by some of the early letters to the editor of *The Lancet* concerning the Seascale cluster. Commenting on claims made in the television program, Craft and Birch (1983) used case and at-risk population data for all cancers and leukemia in children under 15 years of age in a series of date bands from 1968 to 1982 for five fairly large arbitrary regions spanning the entire northwest of England. They concluded that at this scale of examination there was no evidence of elevated rates, as claimed by the program, and they pointed out the well-known property of a Poisson distribution, that apparent clusters of events often occur by chance. On the other hand, using data on death rates from leukemia in people under 25, together with population data from the 1961 and 1971 U.K. Census of Population, Gardner and Winter (1984) computed observed and expected rates across 14 smaller Local Authority Areas in Cumbria for 1959–1967 and 1968–1978, showing that the estimated rate for the second time period for one rural district next to the Sellafield plant was around 9.5 times what would have been expected by chance. This appears to confirm the suspicion of a cancer cluster. In another letter to the editor of *The Lancet'*, the originators of the study featured in the TV program suggested that clusters would appear at a finer level of spatial aggregation than Craft and Birch used. They also stated that the contention that some clusters would occur by chance “fails to meet” the point that a cluster of cases may also be a sign that some specific cause is at work” (Urquhart et al., 1984). This echoes Lloyd et al.’s (1984) comment that epidemiologists may be too prone to translate “cause unknown” as meaning “by chance.” These criticisms led Craft et al. (1984) to recompute rates using the smallest available spatial regions, which in 1981 were the 675 Census Wards. The result was that Seascale, the village closest to the Sellafield field plant, came out with the highest ranked Poisson probability ($p = 0.0001$). However, as the authors pointed out, using these data, many small areas of the region studied could be claimed to have an excess of childhood cancer. They argued that ”these variations are almost inevitable for a group of diseases with an average incidence of 106 per million of total population.”

What is happening here is just as logic would suggest: the finer the spatial resolution of the units chosen, the more apparent clusters of cases appear. A secondary issue appears at finer spatial resolutions: the numbers of cases used in the numerator of any ratio gets smaller and smaller, so that decisions about which cases to include and which to omit become increasingly important. In response to Craft et al. (1984), Urquhart and Cutler (1985) updated the study period and “found” six or seven additional cases. Given the low numerators involved, addition of these cases seemed to them to change the area ratios, but both Craft et al. (1985) and Gardner (1985) challenged these additions. The detail is unimportant to us here. What matters is that simply changing the categories of cancer used and/or the time period over which the case data are aggregated can change the picture that emerges dramatically, particularly at high spatial resolutions.

So, attempting to correct for inhomogeneity by estimating rates based on arbitrary spatial regions may not be as good (or as simple) an idea as it appears, since it introduces problems associated with the modifiability of the areal units, the choice of an appropriate spatial scale/resolution at which to operate, the time periods over which the data are aggregated, and in many cases, instabilities due to the small numbers involved. From a geographical perspective an even greater problem is that the approach discards most, if not all, of the locational information available in the distribution of the cases/events.

Approaches Based on KDE

In Section 3.6 we outlined *kernel density estimation* (KDE), which for a given bandwidth and kernel shape produces an *estimate* of the local intensity of a process. An obvious question is whether we can use this method to correct for variations in an underlying population at risk. Clearly, for the diseases cases, which are represented by point objects, we can use standard KDE to estimate the spatial intensity of cases. It is harder to estimate this for the underlying at-risk population, where we usually only have area-aggregated totals. A possibility discussed by Bailey and Gatrell (1995, pp. 126–128) is to "locate" each area total at a summary point in the area, such as its centroid, and use KDE to estimate the at-risk population intensity. Using this approach, the validity of the assumption that the population can be centered on a single point is, of course, critical. Even so, the ratio of kernel estimates for the events and the population density provides a relatively easily calculated estimate of the population-corrected disease intensity.

In the case of cancer clusters around Sellafield, Bithell (1990) used KDE on both the cases and the at-risk population data, taking the ratio of the two surfaces, for cases and background, as an estimate of the relative risk. Taking the ratio of two areal density estimates cancels out the "per unit of area" term in both, but it does not cancel out any influences that the choice of bandwidth has on both sets of density estimates. Although it might seem desirable to use the same kernel (form and bandwidth) for both estimates, Bailey and Gatrell (1995, p.127) recommend oversmoothing the denominator using a larger bandwidth and Bithell (1990) displays a series of possible relative risk maps based on different bandwidths. There are numerous other possible approaches to this problem, many of them readily implemented in a GIS environment where the distribution of the variable that creates the inhomogeneity can be established (see, for example, Baddeley et al., 2000; Schabenberger and Gotway, 2005; Perry et al., 2006; Bivand et al., 2008).

Approaches Based on Cases/Controls

Where the underlying at-risk population is available as a second set of pure point events, the use of ratios of density estimates is greatly simplified. The archetype example is where we have a set of n_1 "cases" making up the first point pattern and n_2 randomly selected "controls" making up the second. If there is no clustering of the cases relative to the controls, we can argue that the cases are indistinguishable from a random sample of the cases and controls. Here the null hypothesis is that all of the individuals, $n_1 + n_2$, are randomly assigned as a case or a control. If this is so, their K functions (see Section 5.2, especially Figures 5.11 and 5.12) should be identical, giving:

$$K_{11}(d) = K_{22}(d) = K_{12}(d) \tag{6.1}$$

We can only estimate these functions from the data, but a plot of the difference between the estimated functions, $D(d) = K_{11}(d) - K_{22}(d)$, should show peaks and troughs where the cases are more and less clustered than the controls. An empirical significance test is straightforward: all that is necessary is to label cases and controls randomly and repeat the test as many times as necessary to establish upper and lower simulation envelopes of $D(d)$. For a case control study of the Sellafield problem, see Gardner et al. (1990). In their study of childhood leukemia in west-central Lancashire, which studiously avoided looking at the Sellafield issue, Gatrell et al. (1996) used this approach to show that, although the pattern appears clustered on a simple pin/dot map, there is no statistically significant clustering relative to the controls.

6.6. FOCUSED APPROACHES

A second problem in practical point pattern analysis occurs where we are looking for clustering around some specific point, line, or area source (or sources if there are more than one). In our example, the hypothesized source is clearly the relevant nuclear installation and the interest is in the increased incidence of the cancer around it, with or without any correction for inhomogeneity, as discussed above.

The original method used by Heasman et al. (1986) in evidence given during the public inquiry into the Dounreay cluster used a *focused* test. Circles were centered on the plant, and for each distance band (e.g., <12.5 km, $12.5 - 25$ km, and "rest of Scotland") and for three time periods—1968–1973, 1974–1978, and 1979–1984—the number of events was counted. The at-risk population was estimated using the two nearest dated census records for 1971 and 1981, with individual census areas classified into the same distance bands. The idea is illustrated in Figure 6.3.

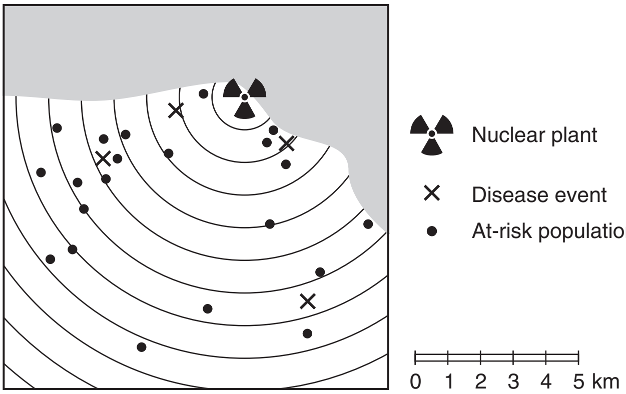


Figure 6.3 Schematic illustration of a focused cluster technique, as used in the Dounreay public inquiry.

Table 6.1 Observed and Expected Leukemia Cases in Distance Bands around the Dounreay Reprocessing Plant (After Heasman et al., 1986)

Time period	*Area and period*	*Observed leukemia cases*	*Expected cases*
1968–1973	<12.5 km	0	0.17
	12.5–25 km	0	0.17
	Other mainland	2	0.41
1974–1978	<12.5 km	0	0.50
	12.5–25 km	0	0.44
	Other mainland	0	1.12
1979–1984	**<12.5 km**	**5**	**0.51**
	12.5–25 km	1	0.45
	Other mainland	1	1.15

The results for the observed cases and the expected numbers are shown in Table 6.1. The table shows that the only evidence of any clustering close to the plant occurred in the 1979–1984 period, with almost 10 times the expected rate of incidence of the disease.

Although this seems reasonable at first sight, from both a geographic and a statistical point of view it is unsatisfactory, a fact that the original researchers were well aware of. First, geographically, the boundaries given by the distance bands are arbitrary and, because they can be varied, are subject to the MAUP, like any other boundaries drawn on a map. Second, quantitative geographers will recognize that rather than assuming an equal risk as one moves away from the focal point, some form of distance decay in the effect will be expected and should be allowed for explicitly (Diggle, 1990). Third, a more serious problem is that the test is *post hoc*. We already have the data, and in using these data to select the focal location, the investigator is being unfair. What would happen if we chose some other center? In an ideal world, the only way around this problem is to postulate the location before collecting the data, but this is almost always unrealistic.

6.7. CLUSTER DETECTION: SCAN STATISTICS

Our third issue in practical point pattern analysis is that often the main interest is not in a general, global test for clustering, or even in relating a pattern to a specific focus, but in detecting *where* there is significantly greater clustering than expected. This process known as *cluster detection*.

The Geographical Analysis Machine

To detect significant clusters in a point pattern, several prerequisites must be in place. First, there has to be some mechanism by which properties of the point pattern can be assessed to determine whether clustering occurs and, if so, at what spatial scale. Second, a mechanism by which a correction can be made for first-order background inhomogeneity is required. Third, there has to be some way of evaluating the statistical significance of the results relative to some null hypothesis.

The *Geographical Analysis Machine* (GAM) of Openshaw et al. (1987, 1988) was an attempt to address these desiderata that drew heavily on GIS technology and data and was applied to the distribution of childhood leukemia over the whole of northern England. At the time of its proposal, the approach caused considerable controversy—much of it unnecessary with the benefit of hindsight. The GAM approach was primarily computational, rooted in the developing GIS technology of the time rather than in pure approaches to statistical hypothesis testing. Nowadays, this type of computational approach is used more often, and the objections raised seem less pertinent than they did at the time. The use of computation to address spatial analytical problems is a topic we return to in Chapter 12.

In its basic form, GAM was an automated cluster detector for point patterns that included elements of geovisualization (Section 3.2), naive kernel density mapping with varying bandwidths (Section 3.6), and Monte Carlo significance testing (Section 5.4). Importantly, it also explicitly rejected the notion of a focused test. The basic GAM conducts an exhaustive search using an approximation to *all* possible centers of all possible clusters over the entire study region. The basic procedure is as follows:

1. Lay out a two-dimensional grid over the study region, in this case the whole north of England. This provides an approximation to the idea of all possible cluster foci.
2. Treat each grid point as the center of a series of search circles.
3. Generate circles of a defined sequence of radii (e.g., 1.0, 2.0, . . . , 20 km), giving an approximation to the idea of all possible bandwidths and therefore all scales of potential clustering. (In total, some 704,703 circles were tested.)
4. For each circle, count the number of events falling within it, for example, 0 to 15 year-old deaths from leukemia, 1968–1985, geolocated to 100 m spatial resolution by unit post codes. This is a standard naive KDE of the type discussed in Section 3.6.
5. Determine whether or not this exceeds a specified density threshold using some population covariate. The published study used the 1981

U.K. Census of Population Small Area Statistics at the Enumeration District level. These aggregated data were treated as if they were located at the centroid of each enumeration district and so could be used to form a count of the at-risk population of children inside each circle. In the region studied, there were 1,544,963 children in 2,855,248 households, spread among 16,237 Enumeration Districts, and the centroids of these districts were geolocated to 100 m resolution by an Ordnance Survey Grid Reference. This provides a correction for inhomogeneity.

6. If the incidence rate in a circle exceeds some threshold, draw that circle on a map. The appropriate threshold was determined by generating 199 sets of synthetic data in which 853 children in the total at-risk population in 1981 of 1.54 million were randomly assigned to have leukemia, and expected values under this null hypothesis were computed. At the 99% confidence level, 3602 circles (0.5%) were isolated as having a higher incidence than expected and were drawn using a pen plotter.

Circle size and grid resolution were linked such that the grid size was 0.2 times the circle radius, so that adjacent circles overlapped. The result is a dense sampling of circles in a range of sizes across the study region. The general arrangement of a set of circles (of one size only) is shown in Figure 6.4. Note that to keep the diagram (almost) readable, these circles are only half as densely packed as in an actual GAM run.

The end result of this procedure is a map of "significant circles," as indicated in Figure 6.4, where six circles with high disease incidence rates

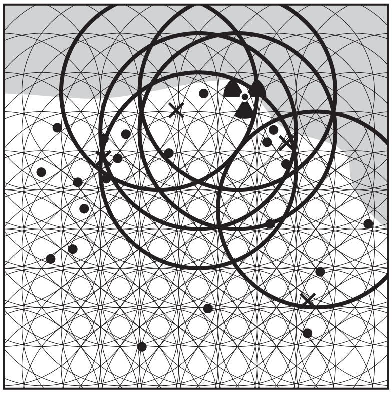


Figure 6.4 The pattern of circles used by GAM. Six circles with high rates of incidence of the disease are highlighted. Note that this diagram is illustrative only.

Table 6.2 Summary Results from GAM: Childhood Leukemia in Northern England

Circle radius (km)	*No. of circles drawn*	*No. of significant at level:*	
		99%	99.8%
1	510,367	549	164
5	20,428	298	116
10	5,112	142	30
15	2,269	88	27
20	1,280	74	31

are drawn with a heavier line. In the original GAM investigation there were many significant circles, as shown in Table 6.2.

The results at the more rigorous 99.8% level test (which requires five times as many simulated patterns to be generated) confirmed the suspected cluster at Sellafield but also identified a much larger cluster in Tyneside, centered on the town of Gateshead, where there is no known source of ionizing radiation. In fact, very few significant circles were drawn outside of these two clusters.

There are some major statistical problems with the GAM approach. The most important one relates to the difficulty of carrying out numerous interrelated significance tests. It is a simple fact, rooted in the logic of significance testing, that at the 99% significance level, we would expect 1% of all circles drawn to be significant if they were nonoverlapping. With 510,000 circles, we would therefore expect 5000 or so to be labeled significant, regardless of the pattern of occurrence of the disease. Thus, the GAM analysis actually detects fewer suspicious clusters than expected at small scales but many more when larger scales are considered. The reasons for this are not entirely clear, although the overlapping of the circles used by GAM complicates matters and may account for the very large number of significant circles listed in Table 6.2 for larger radii. Significance tests using overlapping circles are not independent of one another, so the GAM may give an exaggerated impression of the severity of a cluster. One response to this is, of course, not to treat the significance level as statistically valid per se (which it is not), but instead to think of it as a sensible way of setting a variable threshold across the study region relative to the simulated results. This view encourages us to think of the results from GAM as exploratory and indicative only.

It should also be noted that the approach is computationally intensive. The original GAM, running on a 1987 supercomputer (an Amdahl 5860), took over 6.5 hours for the cancer study, and using very small circle radius increments with large overlap, it could run for as long as 26 hours. Since that time, computers have increased in speed and flexibility, so that what seemed extraordinary a couple of decades ago is now commonplace. Using this increase in computer power has led to a series of approaches that in spirit, if not in detail, mirror the GAM approach and go under the general name of *scan statistics* (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997). Rushton's DMAP takes a similar approach, but keeps the circle size constant to enable the investigator to specify a scale for the clustering (Rushton and Lolonis, 1996).

More recent work at the Centre for Computational Geography at Leeds University has pursued the geocomputational angle using *genetic algorithms* (see Section 12.3) to generalize the GAM idea in devices called the MAP Explorer (MAPEX) and the STAC (Space Time Attribute Creature). Instead of blindly testing all options, MAPEX and STAC are vaguely "intelligent" in that, if they find evidence of a cluster, they adapt their behavior to zero in on it. However, the basic operation is much the same as that of the original GAM, using cheap computer power to test all possible options.

In conclusion, we would argue that a typical GIS that features a toolkit of functions for point pattern analysis is not very useful for serious analysis, except for the simplest and purest problems. As the Seascale/Sellafield example shows, detecting clustering and locating clusters in data when spatial variation is expected anyway *is a very difficult problem*. Doing the science properly will almost invariably require use of specialist software designed by spatial statisticians such as that available in the *R* programming environment (Bivand et al., 2008), which includes Baddeley's *SpatStat* system (Baddeley and Turner, 2005). Levine's *CrimeStat III* offers a less flexible but more approachable alternative to *R* (Levine, 2004).

6.8. USING DENSITY AND DISTANCE: PROXIMITY POLYGONS

Many of the issues that we have been discussing revolve around the fact that geographic space is nonuniform, so different criteria must be applied to identify clusters at different locations in the study region. For example, this is the reason for the adoption of a variable "significance level" threshold in the GAM approach. In this context, it is worthwhile to briefly discuss a recent development in the analysis of point patterns that has considerable potential for addressing this problem.

The approach in question is the use of *proximity polygons* and the *Delaunay triangulation* in point pattern analysis. This idea is most easily explained starting with the construction of the proximity polygons of a point

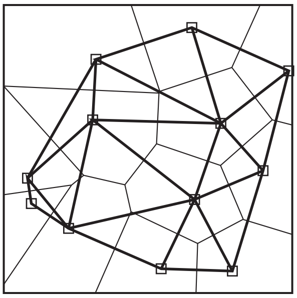


Figure 6.5 Proximity polygons and the Delaunay triangulation for a point pattern.

pattern. Recall from Section 2.3 that the proximity polygon of any entity is that region of space closer to the entity than to any other entity in the space. This idea is readily applied to the events in a point pattern and is shown in Figure 6.5 for the point pattern of Figure 5.7. Also shown is the Delaunay triangulation derived from the proximity polygons by joining pairs of events whose proximity polygons share a common edge.

The idea of using these constructions for point pattern analysis is that the proximity polygons and the Delaunay triangulation have measurable properties that may be of interest. For example, the distribution of areas of the proximity polygons provides an indication of how evenly spaced (or not) the events are. If the polygons are all of similar sizes, then the pattern is evenly spaced. If there is wide variation in polygon sizes, then points with small polygons are likely to be in closely packed clusters, and those in large polygons are likely to be more remote from their nearest neighbors. The number of neighbors that an event has in the Delaunay triangulation may also be of interest. Similarly, the lengths of edges in the triangulation give an indication of how evenly spaced (or not) the pattern is. Similar measures can also be made on the proximity polygons themselves. These approaches are detailed in Okabe et al. (2000), and there is a geographic example in Vincent et al. (1976).

There are two other constructions derived from the Delaunay triangulation whose properties can also be of interest in point pattern analysis. These are shown in Figure 6.6. In the left-hand panel, the *Gabriel graph* has been constructed. This is a reduced version of the Delaunay triangulation, where any link that *does not intersect the corresponding proximity polygon edge* is removed. The proximity polygons have been retained in the diagram, and by comparison with Figure 6.5, you should be able to see how this works.

In the right-hand panel, the *minimum spanning tree* of this set of points is shown. This is the set of links from the Delaunay triangulation, with *minimum total length* that together joins all the events in the pattern. This construction includes all the links between nearest-neighbor pairs.

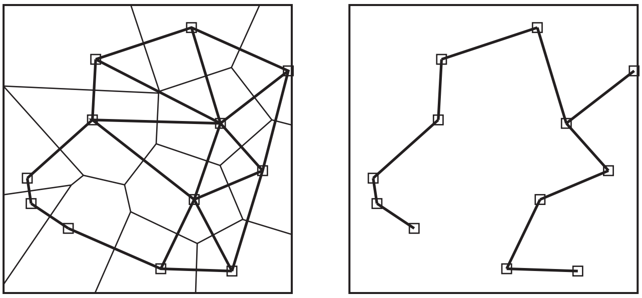


Figure 6.6 The Gabriel graph (left) and the minimum spanning tree (right) of a point pattern.

The minimum spanning tree is much more commonly seen than the Gabriel graph. Its total length may be a useful summary property of a pattern and may provide more information than the simple mean nearest-neighbor distance. You can see this by thinking about what happens to each measure if the clusters of connected events in Figure 5.7 are moved farther apart, as shown in Figure 6.7. This change does not affect the mean nearest-neighbor distance, since each event's nearest neighbor is the same, as indicated by the solid lines. However, the minimum spanning tree is changed, as indicated by the dotted lines now linking together the clusters of near neighbors, because it must still join together all the events in the pattern. You will find it instructive to think about how this type of change in a pattern affects the other point pattern measures we have discussed.

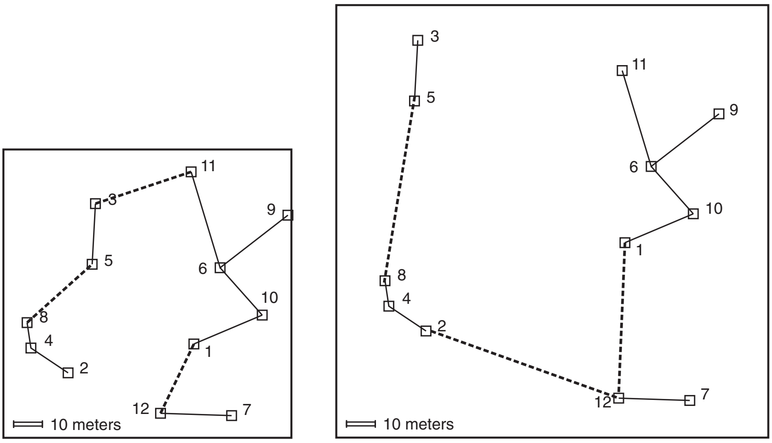


Figure 6.7 The effect of ''exploding'' a clustered point pattern. The point pattern on the left (from Figure 5.7) is changed by moving its constituent clusters only.

In contrast with the point pattern measures we have reviewed, where local inhomogeneity creates major problems, neighborhood relations determined from proximity polygons are defined with respect to local patterns and not using fixed criteria like “nearest neighbor” or “within 50 m.” It seems likely that this property of proximity polygons and related constructions may allow the development of cluster detection techniques that have a “natural” mechanism for determining locally high concentrations. This idea has not yet been developed into a working geographic cluster detection method (although machine vision researchers have been interested in the technique for many years; see Ahuja, 1982). The key question that must be addressed in any development of this idea will be how the background or the at-risk population is linked to locally varying properties of the proximity polygon tessellation.

6.9. A NOTE ON DISTANCE MATRICES AND POINT PATTERN ANALYSIS

In this short section, we consider how the distance-based methods in point pattern analysis can be calculated using a distance matrix, as introduced in Section 2.3. First, assume that we have a distance matrix $\mathbf{D}(S)$ for our point pattern S. Each entry in this matrix records the distance between the corresponding pair of events in the point pattern. The distance matrix for the simple point pattern of Figure 5.7 is

$$\mathbf{D}(S) = \begin{bmatrix} 0 & 44.9 & 59.6 & 56.8 & 44.9 & 27.9 & 28.1 & 58.5 & 55.2 & \underline{25.6} & 59.6 & 26.8 \\ 44.9 & 0 & 59.6 & \underline{15.6} & 38.6 & 64.1 & 58.6 & 22.6 & 93.9 & 70.2 & 81.7 & 34.8 \\ 59.6 & 59.6 & 0 & 55.0 & \underline{21.1} & 48.7 & 87.5 & 47.6 & 67.0 & 69.6 & 35.0 & 76.2 \\ 56.8 & 15.6 & 55.0 & 0 & 36.1 & 71.4 & 73.6 & \underline{9.0} & 100.5 & 81.1 & 82.7 & 50.3 \\ 44.9 & 38.6 & \underline{21.1} & 36.1 & 0 & 44.4 & 71.4 & 30.3 & 70.1 & 61.6 & 47.0 & 56.8 \\ 27.9 & 64.1 & 48.7 & 71.4 & 44.4 & 0 & 51.4 & 69.6 & 29.8 & \underline{21.9} & 34.6 & 54.6 \\ 28.1 & 58.6 & 87.5 & 73.6 & 71.4 & 51.4 & 0 & 78.0 & 72.2 & 36.4 & 85.6 & \underline{24.8} \\ 58.5 & 22.6 & 47.6 & \underline{9.0} & 30.3 & 69.6 & 78.0 & 0 & 97.9 & 81.6 & 77.2 & 55.9 \\ 55.2 & 93.9 & 67.0 & 100.5 & 70.1 & \underline{29.8} & 72.2 & 97.9 & 0 & 35.8 & 36.7 & 81.7 \\ 25.6 & 70.2 & 69.6 & 81.1 & 61.6 & \underline{21.9} & 36.4 & 81.6 & 35.8 & 0 & 55.3 & 49.1 \\ 59.6 & 81.7 & 35.0 & 82.7 & 47.0 & \underline{34.6} & 85.6 & 77.2 & 36.7 & 55.3 & 0 & 84.3 \\ 26.8 & 34.8 & 76.2 & 50.3 & 56.8 & 54.6 & \underline{24.8} & 55.9 & 81.7 & 49.1 & 84.3 & 0 \end{bmatrix} \tag{6.2}$$

Even this small pattern generates a large amount of data—although you will note that the matrix is symmetrical about its main diagonal. This is because the distance between two events is the same regardless of the direction in which we measure it.

In each row of the matrix we have underlined the shortest, or nearest-neighbor, distance. Thus, the nearest-neighbor distance for event 1 (row 1) is 25.6. You may wish to compare these values to those in Table 5.2. Aside from rounding, they are the same. Therefore, the 12 underlined values in the distance matrix may be used to determine both the mean nearest-neighbor distance for this point pattern and its *G* function.

Note that in practice, if we were only interested in nearest-neighbor-based measures, we would not calculate all the distances, as has been done here. It is generally better for a larger data set to make use of efficient spatial data structures that allow the nearest neighbor of a point to be rapidly determined. The need for such efficient data structures should be clear if you imagine the distance matrix for a 100-event pattern—there are 4950 interevent distances—or for a 1000-event pattern, with 499,500 distinct distances. The types of data structure required are discussed in GIS texts (see, for example, Worboys, 1995, pp. 261–267).

Of course, some pattern measures, such as the *K* function, require that all interevent distances be calculated anyway. In this case, we can think of the determination of $K(d)$ as being equivalent to converting $\mathbf{D}(S)$ to an adjacency matrix $\mathbf{A}_d(S)$, where the adjacency rule is that any pair of events less than the distance d apart are regarded as adjacent. For the above matrix, at distance $d = 50$, we would obtain the adjacency matrix

$$\mathbf{A}_{d=50}(S) = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \qquad (6.3)$$

Now, if we sum the rows of this matrix, we get the number of events within a distance of 50 m of the corresponding event. Thus, event 1 has six events within 50 m, event 2 has five events within 50 m, and so on. This is precisely the information required to determine $K(d)$, so we can see the usefulness of the distance matrix summary of the point pattern.

Variations on this general idea may be required for determination of other pattern measures.

For example, the pair correlation function requires that the standard distance matrix be analysed in a different way, while the F function requires a distance matrix where rows represent the random set of points in the empty space, and columns represent the events in the point pattern.

One important fact to note here is that the matrix representation is not convenient for humans (all those horrible rows of numbers!), but it is very conveniently handled by computers, which perform the required calculations.

CHAPTER REVIEW

- In academic geography, there has been significant and sensible criticism of classical point pattern analysis approaches. Some of these criticisms are partly addressed by considering alternative approaches to statistical inference such as *likelihood*.
- Another important set of innovations relates to alternative point processes, such as the *inhomogeneous Poisson process* or *Poisson clustering*.
- None of the point pattern measurement techniques discussed in previous chapters indicate *where* there is clustering in a pattern. This is a significant omission in practical applications where the identification and explanation of clusters is of the utmost importance.
- The incidence of childhood leukemia close to nuclear installations in the United Kingdom provides a very practical example of some of these real-world problems.
- We can distinguish general, focused, and scan approaches to cluster detection. *General* tests detect global clustering in a point pattern with or without a background at-risk population. *Focused* tests are those in which proximity to some assumed source is relevant, whereas *scan statistics* attempt to locate significant clusters of cases.
- The difficulty in identifying clusters in real data is that clusters must be found against a background of expected variations due to the uneven distribution of the *at-risk population*, which means that the null model often involves an *inhomogeneous Poisson process*.
- A simple way to correct for inhomogeneity is to compute areal rates of incidence, but this introduces the MAUP. Alternatives that use KDE and a variation of Ripley's K offer a partial solution to this problem.
- The *Geographical Analysis Machine* (GAM) was developed to address many of the complex issues involved in cluster detection. It works by exhaustively sampling the study area in an attempt to find "significant

circles" where more cases of a disease have occurred than might be expected as indicated by simulation.

- The original GAM method was relatively "dumb" and required enormous computing power by the standards of the time. It is now possible to run it on a standard desktop PC or remotely over the Internet. More recent versions of the GAM idea attempt to apply more intelligent search procedures.
- A family of measurement methods based on the *geometric properties of proximity polygons*, the *Delaunay triangulation*, and related constructions such as the *Gabriel graph* and the *minimum spanning tree* of a point pattern, can be developed. These methods are not often used at present, but they hold out the possibility of developing cluster detection methods that are sensitive to local variations in pattern intensity.
- The *minimum spanning tree* demonstrates an important limitation of nearest-neighbor-based measures when a clustered pattern is "exploded."
- The distance and adjacency *matrices* discussed in Chapter 2 can often be used in calculating distance-based point pattern measures.

REFERENCES

Ahuja, N. (1982) Dot pattern processing using Voronoi neighbourhoods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI 3: 336–343.

Baddeley, A. J., Moller, J., and Waagespetersen, R. (2000) Non- and semi-parametric estimation of interaction in inhomogengeous point patterns. *Statistica Neerlandica*, 54: 329–350.

Baddeley, A. and Turner, R. (2005) Spatstat: an *R* package for analyzing spatial point patterns. *Journal of Statistical Software*, 12: 1–42. (For a comprehensive set of resources related to this package, see http://school.maths.uwa.edu.au/homepages/adrian/.)

Bailey, T. C. and Gatrell, A. C. (1995) *Interactive Spatial Data Analysis* (Harlow, Essex, England: Longman).

Besag, J. and Newell, J. (1991) The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, 154: 143–155.

Bithell, J. F. (1990) An application of density estimation to geographical epidemiology. *Statistics in Medicine*, 9: 691–701.

Bithell, J. F., Keegan, T. J., Kroll, H. E., Murphy, M. F. G., and Vincent, T. J. (2008) Childhood leukemia near British nuclear installations: methodological issues and recent results. *Radiation Protection Dosimetry*, 132: 191–197.

Bithell, J. C. and Stone, R. A. (1989) On statistical methods for analyzing the distribution of cancer cases near nuclear installations. *Journal of Epidemiology and Community Health*, 43: 79–85.

Bivand, R. S., Pebesma, E. J., and Gomez-Rubio, V. (2008) *Applied Spatial Data Analysis with R* (New York: Springer).

Black, Sir D. (1984) *Investigation of the Possible Increased Incidence of Cancer in West Cumbria* (London: Her Majesty's Stationery Office).

Bolstad, W. M. (2007) *Introduction to Bayesian Statistics*, 2nd ed. (Hoboken, NJ: Wiley).

Craft, A. W. and Birch, G. M. (1983) Childhood cancers in Cumbria. *The Lancet*, 322: 1299.

Craft, A. W. and Openshaw, S., and Gardner, M. J. (1985) Childhood cancers in West Cumbria. *The Lancet*, 325: 403–404.

Craft, A. W, Openshaw, S., and Birch, J. (1984) Apparent clusters of childhood lymphoid malignancy in Northern England. *The Lancet*, 324: 96–97.

Cuzick, J. and Edwards, R. (1990) Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society, Series B*, 52: 73–104.

Diggle, P. J. (1990) A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society, Series A*, 153: 349–362.

Edwards, A. W. F. (1992) *Likelihood*, Expanded Edition (Baltimore, MD: Johns Hopkins University Press).

Gardner, M. J. (1985) Childhood cancer in West Cumbria. *The Lancet*, 325: 403–404.

Gardner, M. J. (1989) Review of reported increases of childhood cancer rates in the vicinity of nuclear installations. *Journal of the Royal Statistical Society, Series A*, 152: 307–325.

Gardner, M. J., Snee, M. P., Hall, A. J., Downes, S., Powell, C. A., and Terrell, J. D. T. (1990) Results of case-control study of leukemia and lymphoma in young persons resident in West Cumbria. *British Medical Journal*, 300(6722): 423–429.

Gardner, M. J., and Winter, P. D. (1984) Mortality in Cumberland during 1959–78 with reference to cancer in young people around Windscale. *The Lancet*, 323: 216–218.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S. (1996) Spatial point pattern analysis and its application in geographical epidemiology, *Transactions of the Institute of British Geographers*, NS 21: 256–274.

Gould, P. R. (1970) Is *statistix inferens*, the geographical name for a wildgoose? *Economic Geography*, 46: 439–448.

Harvey, D. W. (1966) Geographical processes and the analysis of point patterns. *Transactions of the Institute of British Geographers*, 40: 85–95.

Harvey, D. W. (1968) Some methodological problems in the use of Neyman type A and negative binomial distributions for the anaysis of point patterns. *Transactions of the Institute of British Geographers*, 44: 85–95.

Heasman, M. A., Kemp, I. W., Urquhart, J. D., and Black, R. (1986) Childhood leukaemia in Northern Scotland. *The Lancet*, 327: 266.

Hills, M. and Alexander, F. (1989) Statistical methods used in assessing the risk of disease near a source of possible environmental pollution: a review. *Journal of the Royal Statistical Society, Series A*, 152: 353–363.

Johnson, S. (2006) *The Ghost Map* (New York: Riverhead; London: Penguin Books).

Kinlen, L. (1988) Evidence for an infective cause of childhood leukemia—comparison of a Scottish New Town with nuclear reprocessing sites in Britain, *The Lancet* 332: 1323–1327.

Kulldorff, M. (1997) A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26: 1481–1496.

Kulldorff, M. (2006) Tests of spatial randomness adjusted for an inhomogeneity: a general framework. *Journal of the American Statistical Association*, 101: 1289–1305.

Kulldorf, M. and Nagarwalla, N. (1995) Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14: 799–810.

Levine, N. (2004) *CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Incident Locations* (version 3.0) (Houston, TX: National Institute of Justice; Washington, DC: and Ned Levine & Associates. (see also http://www.nedlevine.com/nedlevine17.htm).

Lloyd, O. L., MacDonald, J., and Lloyd, M. M. (1984) Mortality from lymphatic and haematopoietic cancer in Scottish coastal towns. *The Lancet*, 324: 95–96.

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. (2000) *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd ed. (Chichester, England: Wiley).

Openshaw, S., Charlton, M., Craft, A. W., and Birch, J. M. (1988) Investigation of leukaemia clusters by use of a geographical analysis machine. *The Lancet*, 331 (8580): 272–273.

Openshaw, S., Charlton, M., Wymer, C., and Craft, A. (1987) Developing a mark 1 Geographical Analysis Machine for the automated analysis of point data sets, *International Journal of Geographical Information Systems*, 1: 335–358.

Perry, G. L. W., Enright, N. J., Miller, B. P., and Lamont, B. B. (2008) Spatial patterns in species-rich sclerophyll shrublands of southwestern Australia. *Journal of Vegetation Science*, 19: 705–716.

Perry, G. L. W., Miller, B. P., and Enright, N. J. (2006) A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecology*, 187: 59–82.

Rushton, G. and Lolonis, P. (1996) Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine*, 15: 717–726.

Schabenberger, O. and Gotway, C. A. (2005) *Statistical Methods for Spatial Data Analysis* (London: Chapman & Hall).

Urquhart, J. and Cutler, J. A. (1985) Incidence of childhood cancer in west Cumbria. *The Lancet*, 325: 172.

Urquhart, J., Palmer, M., and Cutler, J. (1984) Cancer in Cumbria: the Windscale connection. *The Lancet*, 323: 217–218.

Vincent, P. J., Howarth, J. M., Griffiths, J. C., and Collins, R. (1976) The detection of randomness in plant patterns. *Journal of Biogeography*, 3: 373–380.

Worboys, M. F. (1995) *Geographic Information Systems: A Computing Perspective* (London: Taylor & Francis).

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 7

Area Objects and Spatial Autocorrelation

CHAPTER OBJECTIVES

In this chapter, we:

- Outline the types of area object of interest
- Show how area objects can be recorded and stored
- Show how area can be calculated from digital data
- Define some of the properties of area objects, such as their *shape, centroid,* and *skeleton*
- Introduce a range of measures of *spatial pattern*
- Return to the concept of a *spatial weights matrix* and describe alternative approaches to building one
- Describe the most widely used measure of *spatial autocorrelation,* Moran's *I*
- Briefly present some alternative measures

After reading this chapter, you should be able to:

- List the general types of area object
- Explain how these can be recorded in digital form
- Outline what is meant by the term *planar enforcement*
- Suggest and illustrate a method for finding polygonal areas using the coordinates of their vertices
- Summarize basic measures of the geometry of areas
- Compute Moran's *I* for a study area and explain how statistical significance can be ascribed to the computed value
- Outline some alternatives to Moran's *I*

7.1. INTRODUCTION: AREA OBJECTS REVISITED

Revision

You can increase your understanding of the materials in this chapter if you take a few minutes to revisit relevant sections of Chapters 1–3 and revise the following:

- How area objects fit into the entity-attribute typology introduced in Chapter 1 and summarized in Figure 1.2
- Autocorrelation, the MAUP, and the ecological fallacy, all discussed in Chapter 2

7.2. TYPES OF AREA OBJECT

Areas are some of the more complex object types commonly analyzed. Before starting, we must distinguish *natural areas*—entities modeled using boundaries defined by natural phenomena such as the shoreline of a lake, the edge of a forest stand, or the outcrop of a particular rock type—from those areas *imposed* by human beings. Natural areas are *self-defining* insofar as their boundaries are defined by the phenomena themselves. Sometimes, as in the case of a lake, the boundary of an object is crisp and its "inside" (the water) is homogeneous. Frequently, however, natural areas are the result of subjective, interpretative mapping by a field surveyor and, as discussed in Chapter 1, may be open to disagreement and uncertainty.

Contrast such natural areas with those imposed by human beings, such as countries, provinces, states, counties, or census tracts. These have been called *fiat*, or *command*, regions as distinct from bona fide regions (see Smith and Varzi, 2000). Here, boundaries are defined independently of any phenomenon, and attribute values are enumerated by surveys or censuses. Imposed areas are common in GIS work that involves data about human beings. Such imposed areas are a form of sampling of the underlying social reality and can be misleading in several ways. First, the imposed areas might bear little relationship to the underlying patterns. In two now classic accounts, Coppock (1955, 1960) showed that imposed civil parish boundaries are totally unsuitable for reporting U.K. Agricultural Census data, which are collected at the farm level. Individual farms can stretch over more than one parish, and parish boundaries frequently include dissimilar agricultural areas and vary widely in size. Second, imposed areas are arbitrary or *modifiable*, and some care is required to demonstrate that *any* analysis is not an artifact of the particular boundaries used. This is the MAUP (see Section 2.2). Third, because data for area objects are often aggregations of individual-level information, the danger of ecological fallacies, where we assume that a relationship at a macrolevel of aggregation also exists at the microlevel, is very real.

A third type of area object arises where space is divided into small, regular grid cells called a *raster*. Unlike natural and imposed areas, which are usually irregularly shaped, with different spatial extents, in a raster the area objects are identical and together cover, or *tessellate*, the region of interest. Because grids are often used to record pictorial information, they are also called *pixels* (from "picture elements"). In GIS, the term refers to a data structure that divides a study area into a regular pattern of cells in a particular sequence and records an attribute for the part of the Earth's surface represented by each cell. Each cell is therefore an area object, but the major concept is that of a continuous *field* of information. In any given database there might be many raster *layers* of information, each representing an individual field. Typically, a raster data structure makes use of square, or nearly square, pixels, but there is nothing sacrosanct about this. Square cells have the advantage that they may be nested at different resolution levels, but, against this, they do not have uniform adjacency. The distance from the center of each cell to its neighbors is greater for diagonally adjacent cells than for those in the same row or column. At the cost of losing the ability to nest pixels, hexagons have uniform adjacency, and triangular meshes also have some attractive properties. Relative to the complexities of storing polygon areas, the advantage of a raster data structure is that once the raster is registered to real-world coordinates, further georeferencing at the individual pixel level is unnecessary. The real-world (x, y) coordinates are implicit in the (*row, column*) position of each pixel within the raster.

Finally, area objects are often created in a GIS analysis using polygonal Voronoi/Thiessen regions around every event in a pattern of point objects (see Section 2.3). These regions are defined such that each contains all locations closer to the generating object than to any other object in the pattern.

Thus, there are several types of area object. Four of these, two natural and two imposed, are shown in Figure 7.1. Area objects have a range of geometric and topological characteristics that can make them difficult to analyze. It may be that the entities are isolated from one another or perhaps overlapping. If the latter is true, then any location can be inside any number of entities, and the areas do not fill or *exhaust* the space. The pattern of areas in successive forest fire burns is an example of this (see Figure 7.1i). Areas may sometimes have holes or areas of different attributes wholly enclosed within them. Some systems allow area entities to have islands. An alternative is that all regions that share the same attribute are defined to be one single

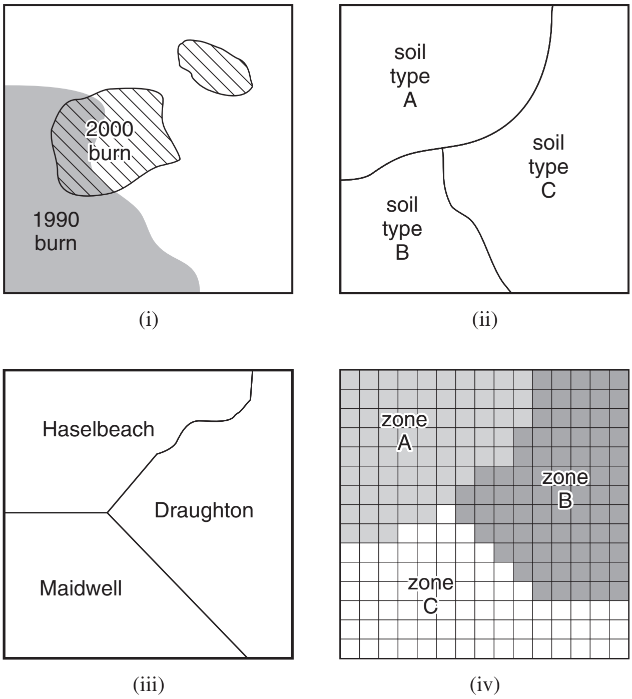


Figure 7.1 Types of area objects: (i) a pattern of forest burns over natural areas that overlap and are not planar enforced; (ii) natural area objects resulting from an interpreted mapping, in this case soil type; (iii) imposed areas, in this case three civil parishes; and (iv) an imposed raster recording three types of soil.

area so that each area object is potentially a collection of areas—like an archipelago. This is perfectly usual when dealing with attributes such as geology or land use and may require special attention. Different from either of these is the case where area objects all mesh neatly together and exhaust the study region, so that there are no holes and every location is inside just a single area. Such a pattern of areas is termed *planar enforced*, and this concept is a fundamental assumption of the data models used in many GISs. Figures 7.1(ii)–(iv) show planar enforced regions.

Early GISs and some simple computer mapping programs store area objects as complete polygons, with one polygon representing each object. The polygon approximates the outline of the area and is recorded as a series of coordinates. If the areas don't touch and if, like forest fire burns, they can overlap, this is a simple and sensible way to record them. However, for many distributions of interest, most obviously census tracts, areas are planar enforced by definition, so using polygon storage will mean that, although nearly all the boundaries are shared between adjacent areas, they are all

Thought Exercise

Obtain a 1:50,000 or similar scale topographic map. If you don't have access to a paper map, then visit your National Mapping Agency's Web site (for example, in Great Britain this is http://leisure.ordnancesurvey.co.uk). For each of the following area objects represented, decide whether they are natural or imposed and show how they might be recorded in a data model that assumes planar enforcement:

1. The map grid
2. The woodland areas
3. Any parklands
4. National, county, or district boundaries.

input and coded twice, once for each adjacent polygon. With two different versions of each internal boundary line stored, errors are likely; it can also be difficult to dissolve boundaries and merge adjacent areas stored this way. The alternative is to store every boundary segment just once, as a sequence of coordinates, and to build areas by linking boundary segments either implicitly or explicitly. Variations on this approach are used in many current vector GISs (see Worboys and Duckham, 2004, pp. 177–185, for a description of common data structures). These more complex data structures can make transferring data between systems problematic. However, GIS analysis benefits from the ready availability of the adjacency information, and the advantages generally outweigh the disadvantages.

7.3. GEOMETRIC PROPERTIES OF AREAS

No matter how they arise, area objects have a number of properties that we may need to measure. These include their two-dimensional *area*, *centroid* and *skeleton*, *shape*, *spatial pattern*, and *fragmentation*. There are a number of geometric properties and analyses provided in GISs that we describe in the sections that follow.

Area

In a GIS, we might wish to estimate the area of a single specified class of object (for example, woodland on a land-use map) or the average area of each

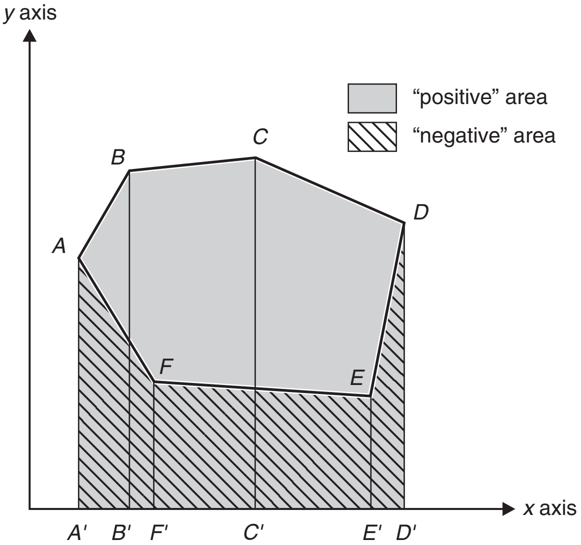


Figure 7.2 Finding the area of a polygon from the coordinates of its vertices.

parcel of land. It is often necessary to find areas as a basis for density calculations. Measuring area is superficially obvious but more difficult in practice (see Gierhart, 1954; Wood, 1954; Frolov and Maling, 1969). In a GIS, the most frequently used algorithm finds the area of a number of *trapezoids* bounded by a line segment from the polygon, two vertical lines dropped to the x-axis, and the x-axis itself, as shown in Figure 7.2.

For example, the area of the trapezoid $ABB'A'$ is given by the difference in x coordinates multiplied by the average of the y coordinates:

$$\text{Area of } ABB'A' = (x_B - x_A)(y_B + y_A)/2 \tag{7.1}$$

Since x_B is greater than x_A, this area will be a positive number. Moving to the next two vertices, B and C, we use the same approach to get the areas of $BCC'B'$ and $CDD'C'$, both also positive numbers. Now, consider what happens as we continue around the polygon and calculate the area of $DD'E'E$. In this case, the x coordinate of E is less than that of D, and the computed area value is a negative number. The same is true for all three trapezoids formed by the lower portion of the polygon. If, as we work around the polygon vertex by vertex, we keep a running total of the area, first we add three positive areas ($ABB'A'$, $BCC'B'$, and $CDD'C'$), obtaining a larger area than required. As we calculate the areas of the three lower trapezoids, $DD'E'E$, $EE'F'F$, and $FF'A'A$, these are negative and are subtracted from the grand total. Inspection of the diagram shows that the result is the area of the polygon $ABCDEF$, that is, the area of the gray-shaded area with the hatched part subtracted. Provided we work clockwise around the polygon and make sure to come back to the starting vertex, the general formula is simply

$$\text{Polygon area}, A = \sum_{i=1}^{n} (x_{i+1} - x_i)(y_{i+1} + y_i)/2 \tag{7.2}$$

where (x_{n+1}, y_{n+1}) is understood to bring us back to the first vertex (x_1, y_1). This is the *trapezoidal rule* for numerical integration and is widely used in science when it is necessary to find the area enclosed by a graph. The algorithm works when the polygon has holes, but not for all polygons, since it cannot handle polygons whose boundaries self-cross. Note also that if the polygon coordinates are stored in a counterclockwise sequence, the area will be negative, so the method relies on a particular sort of data structure.

How Big Is Mainland Australia?

This exercise can be done using a semiautomatic digitizer or on-screen using standard graphics software, but it is useful to do it by hand. Alternatively, you can, of course, also do it in a GIS, using a high-resolution map as a background to work from.

Trace the shoreline of Australia from a map of the continent, taking care to ensure that the source is drawn on an equal-area map projection.

Record the shoreline as a series of (x, y) coordinates. How many vertices do you need to represent the shape of Australia so that it is instantly recognizable? What is the minimum number you can get away with? How many do you think you need to ensure that you get a reasonable estimate of the total area of the continent?

Use the method outlined above to compute its area. This is easily done using any spreadsheet program. Enter your coordinates into the first two columns and copy these from row 2 on into the next two columns, displacing them upward by one row as you do so. Copy the first coordinate pair into the last row of the copied columns. The next column can then be used to enter and calculate the trapezoid formula. The sum of this column then gives your estimate of the continent's area. You will have to scale the numbers from coordinate units into real distances and areas on the ground. Compare your

(*continues*)

(*box continued*)

result with the ''official'' value, which is 7,617,930 km^2, according to a fact sheet produced by the Australian government (see http://www.dfat.gov.au/geo/fs/aust.pdf).

We think that the minimum number of coordinate pairs needed to make the result recognizably Australia is nine. Using a 1:30,000,000 map as a source, we recorded just 45 coordinates for the shoreline. We got an area of 7,594,352 km^2, which is about 1.3% too low. Mind you, almost 100,000 km^2 is a lot of land—almost as much as Iceland, South Korea, or the state of Kentucky. In fact, the closeness of this result is likely to be a happy accident.

What conclusions do you draw from this exercise? In Section 1.3 we pointed out that most superficially exact geometric calculations on spatial data yield quantities that are really estimates of some true but unknown value, and area is no exception.

Calculation of the area of a given polygon is therefore straightforward. However, this approach can only determine the area defined by the stored polygon vertices. Strictly, the result is only an estimate of the true area, its accuracy dependent on how representative the stored vertex coordinates are of the real outline and, hence, on the resolution of the input data. What happens if even the real boundaries are in some way uncertain? Likewise, if we are computing the area of a fuzzy object, or one that isn't internally homogeneous, how can we take account of this in the area measure we obtain? Again, we have an estimate of the true area, but there are circumstances where the "error bars" around this estimate might be very large indeed. It can be very important to recognize the uncertainty. An example might be where we are finding the area of a particular soil type or of a forest stand as a basic input in some resource evaluation. Similarly, controversy surrounding the rate at which the Amazon rain forest is being cut down is ultimately an argument about area estimates and has important implications for the debates over climate change (see Nepstad et al., 1999; and Houghton et al., 2000).

Finally, suppose that we want to calculate the total area of many polygons, such as the total area of a scattered land use type across a geographic region. The details of how this is done depend on how the data are structured, but any method is effectively the repeated application of the trapezoid procedure. In a raster structure, areas may be determined more simply, by counting pixels and multiplying by the pixel area. For a fragmented set of area objects such as a land cover map, it may therefore be more efficient to use raster coding to estimate areas.

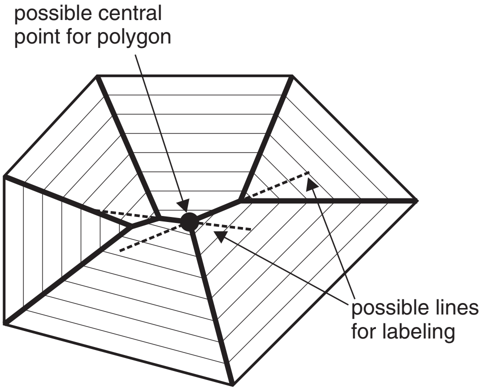

possible central point for polygon
possible lines for labeling

Figure 7.3 The skeleton and resultant central point of a polygon.

Skeleton and Centroid

The *skeleton* of a polygon is a network of lines inside a polygon constructed so that each point on the network is equidistant from the nearest two edges in the polygon boundary. Figure 7.3 shows the idea.

The skeleton is constructed by shrinking the polygon outline inward, with each boundary moving inward at the same rate. As this proceeds, vectors and arcs merge, forming a tree-like structure, and, as the polygon gets smaller, it may form two or more isolated "islands." Ultimately, the polygon is reduced to a possible central point that is farthest from the original boundary and is also the center of the largest circle that could be drawn inside the polygon. This center point may be preferable to possible polygon centers calculated by other means. For example, the more easily calculated *mean center* of the polygon vertices sometimes lies outside the polygon area and may therefore be unsuitable for some purposes. In contrast, a center point on the skeleton is guaranteed to lie inside the polygon.

The polygon skeleton is useful in computer cartography and provides potential locations for label placement on maps. The skeleton center point also has potential uses in analysis when we want a representative point object location for an area object. As noted in Table 1.1, the center point also offers a possible way of transforming between two of the basic geometric object types, from an area to a point object.

Shape

Areal units all have a two-dimensional *shape*, that is, a set of relationships of relative position between points on their perimeters, which is unchanged by changes in scale. Shape is a property of many objects of interest in geography, such as drumlins, parks or reserves, coral atolls, and central business districts. Some shapes, notably the hexagonal market areas of central place theory, are the outcomes of postulated generating processes. Shape may also have important implications for processes. In ecology, the shapes of patches of a specified habitat are thought to have significant effects on what happens in and around them. In urban studies, the traditional monocentric city form is considered very different in character from the polycentric sprawl of Los Angeles or the edge cities of the contemporary world (Garreau, 1992).

In the past, shape was described verbally, using analogies such as "streamlined" (drumlins), "ox-bow" and "shoestring" (lakes), "armchair" (cirques), and so on, although there was often little agreement on what terms to use (see Clark and Gaile, 1975; Frolov, 1975; Wentz, 2000). While quantifying the idea of shape therefore seems worthwhile, in practice, attempts to do this have been less than satisfactory. The most obvious quantitative approach is to devise indices that relate the given shape to a regular geometric figure of a well-known shape, such as a circle, hexagon, or square. Most attempts to date use the circle.

Figure 7.4 shows an irregular shape together with a number of possible shape-related measurements that could be taken from it, such as the perimeter P, the area a, the longest axis L_1, the second axis L_2, the radius of the largest internal circle R_1, and the radius of the smallest enclosing circle R_2. In principle, we are free to combine these values in any reasonable way, although not all combinations will produce a good index. A good index should have a known value if the shape is circular, and to avoid dependence on the measurement unit adopted, it should also be dimensionless.

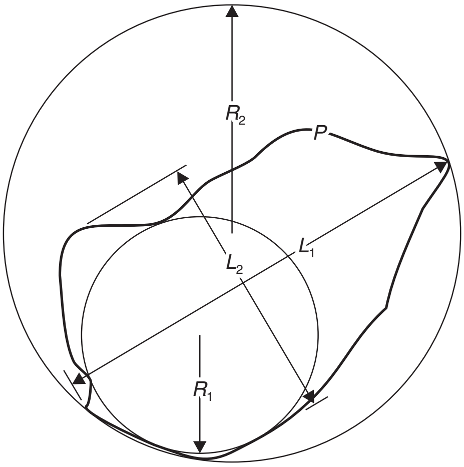


Figure 7.4 Measurements used in shape analysis.

One such index is the *compactness ratio*, defined as

$$\text{compactness} = \sqrt{(a/a_2)} \tag{7.3}$$

where a_2 is the area of the circle having the same perimeter (P) as the object. The compactness is 1.0 if the shape is exactly circular and it is also dimensionless. Other potentially useful and dimensionless ratios are the *elongation ratio* or *eccentricity*, L_1/L_2, and the *form ratio*, $a/{L_1}^2$.

Boyce and Clark (1964) proposed a more complicated measure of shape. Their *radial line index* compares the observed lengths of a series of regularly spaced radials from a node at the center of the shape with those that a circle would have and that would obviously be a fixed value equal to the circle radius. Although this index has been used in a number of studies, reviewed in Cerny (1975), it suffers from three sources of ambiguity. First, no guidance is given on where to place the central point, although most investigators use the shape's center of gravity. Second, the choice of the number of radii is important. Too few make the index open to strong influence from extreme points on the perimeter; with too many, the work of calculation may become excessive. Third, it is apparent that a great many visually different shapes could give the same index value. Alternatives have been developed by Lee and Sallee (1970), Medda et al. (1998), and Wentz (2000)

The area of shape analysis is currently seeing rapid development as a result of content-based image retrieval applications, which aim to automate the task of searching large image databases for items of interest. A wide range of techniques have been developed, many of which are reviewed by Zhang and Lu (2004). In general, shape analysis remains a challenging area, and multiple measures prove useful in characterizing even relatively simple shapes.

Spatial Pattern and Fragmentation

So far, we have concentrated solely on the measurable properties of areas as individual units of study without reference to the overall pattern that they create. Sometimes, as in geomorphology and biogeography, the patterns made by areas are of interest in their own right, irrespective of the values that might be assigned to them. Such patterns can be as regular as a chessboard, honeycomb, or contraction cracks in basalt lavas or as irregular as the counties of England and the states of the United States. A simple approach to this problem is to assemble the frequency distribution of *contact numbers*, that is, the number of areas that share a common boundary with each area (Boots, 1977). An example is given in Table 7.1, which shows the

Table 7.1 Contact Numbers for the Counties of England and the Lower 48 States of the United States

Contact number, m	*Percentage of lower 48 U.S. states*	*Percentage of English counties, n = 46*	*Percentage expected in an independent random process*
1	2.0	4.4	N/A
2	10.2	4.4	N/A
3	18.4	21.7	1.06
4	20.4	15.2	11.53
5	20.4	30.4	26.47
6	20.4	10.9	29.59
7	4.1	13.0	19.22
8	4.1	0	8.48
9	0	0	2.80
10	0	0	0.81
Totals	100.00	100.00	100.00
Mean contact number	4.45	4.48	6.00

frequency distribution of contact numbers for the contiguous states of the United States and the counties of England.

It is evident that very regular patterns, like honeycombs, will have frequency distributions with a pronounced peak at a single value, while more complex patterns will show spreads around a central modal value. The independent random process introduced in Section 4.2 can be used to generate polygonal areas and, in the long run, produces the expected distribution given in the last column of the table. The modal value is for areas with six neighbors. It is apparent that these administrative areas have lower contact numbers than expected, implying a more regular than random patterning. Note, however, that random expected values cannot be compared directly with the observed case for two reasons. First, the method of defining the random process areas ensures that the minimum contact number must be three. Second, the procedure does not have edge constraints, whereas both the United States and England have edges. Furthermore, as with point pattern analysis, the usefulness of this finding is open to debate, since we know to begin with that the null hypothesis of randomness is unlikely to hold.

Perhaps more useful are measures of *fragmentation*, or the extent to which the spatial pattern of a set of areas is broken up. Fragmentation indices and other measures of the spatial pattern or configuration of a set of areas, or *patches*, are widely used in ecology (see, for example, Turner et al., 2001). In this context, the way in which a landscape is broken up into areas that represent different kinds of habitat is often considered important. While occasionally only the total available area of a particular habitat is of interest, more often the number and size of the habitat patches or the length of the boundary between habitats are significant. Often, only areas above some minimum size represent a viable habitat, but equally, for the avoidance of catastrophic events, it may be important that there be some minimum number of habitat patches and that they not be too widely dispersed (so that populations can move from one patch to another). How much edge adjoining other habitat types patches have may also be of interest, as this may affect habitat quality and in some cases may make a habitat more prone to invasion by exotic species. In different circumstances, patches with long edges may be desirable as corridors, or perhaps as obstacles; fire breaks are an example of the latter.

Many measures of these and related patch characteristics are provided by a useful tool called FRAGSTATS (Berry et al., 1998; McGarigal et al., 2002). In its most recent version, this software works on categorical raster data. It can be set up to treat one category as a matrix and other categories as patches and corridors in the matrix. For each category, the numbers and sizes, along with a range of other metrics, are calculated for every patch, along with their averages and standard deviations relative to the overall distribution of patch sizes, and relative to patch sizes within that category. Other metrics measure the degree of separation or proximity of patches to similar patches and the relative abundance of different patch types, as well as many other features of a landscape. Full details of the outputs available are provided in the above works. A more recent program with similar capabilities is IAN, also developed in an ecology context (DeZonia and Mladenoff, 2004).

7.4. MEASURING SPATIAL AUTOCORRELATION

In the remainder of this chapter, we develop the idea of spatial autocorrelation, first introduced in discussing the problems with spatial data in Section 2.2; we also describe ways of measuring it. You will recall that *spatial autocorrelation* is a technical term for the fact that spatial data from near locations are more likely to be similar than data from distant locations. More correctly, any spatial data set is likely to have characteristic distances or lengths, or *lags*, at which it is correlated with itself, a property known as *self-correlation* or *autocorrelation*. Furthermore, according to Tobler's (1970) first law of geography that "Everything is related to everything else, but near things are more related than distant things," autocorrelation is likely to be most pronounced at short distances. If the world were not spatially autocorrelated in this way, then geography would have little point, so autocorrelation is extremely important to the discipline and to spatial analysis. The ubiquity of spatial autocorrelation is the reason why spatial data are special. As a result of autocorrelation, samples from spatial data are not truly random, with consequences for statistics that are a major theme of this book.

As geographers, we are predisposed to spatial patterns in data, and because of autocorrelation, patterns very often appear to be there. One reason for developing analytical approaches to spatial autocorrelation is to provide a more objective basis for deciding whether or not there really is a spatial pattern, and if so, how unusual that pattern is. The by-now familiar problem is to decide whether or not any observed spatial autocorrelation is significantly different from random. Could the apparent pattern have occurred by chance? Arguably, one of the tests for autocorrelation discussed in the remainder of this chapter should *always* be carried out before we start developing elaborate theories to explain the patterns we think we see in a map just in case those apparent patterns are no more than a chance occurrence.

The degree to which data are similar or different over short or long ranges is fundamental to all branches of geographic information analysis, and the autocorrelation concept is correspondingly applicable to all the types of spatial objects we have recognized (point, line, area, and field) but, for pedagogic convenience and with an eye on tradition, we introduce the idea in the context of patterns in the attributes of area objects. Traditionally, spatial autocorrelation has been thought of as a statistical property of a spatial pattern, but in the context of the measures of pattern discussed in the previous section, it is also possible to think of it as just another pattern metric. Many of the point pattern measures already considered in Chapters 4, 5, and 6 can be considered as measures of autocorrelation for the occurrence of point events. Similarly, the semivariogram, which is fundamental to more advanced methods of interpolation (discussed in Chapter 10), is another approach to characterizing autocorrelation in a continuous field.

Spatial Structure and the Spatial Weights Matrix

The essental idea of any approach to autocorrelation is to assess how similar or different attribute values at geographic locations are relative to how spatially close or distant are the associated locations. In broad terms, it is easy to see how we can assess similarity in attribute values using some simple calculation based on the difference in the attribute values. The real research question is how to incorporate spatial proximity into a measure of autocorrelation. In fact, we have already encountered the conceptual tools needed to do this, in Section 2.3, where the concepts of distance, adjacency, interaction, and neighborhood were introduced. Each of these is a way to represent spatial relationships between locations.

In the measurement of autocorrelation, we need to capture the spatial relationship between all pairs of locations, and this is done using a *spatial weights* or *spatial structure matrix* generally denoted **W**. In the first row of the matrix, we record the spatial relationship between the first location and every other location in the map in turn, so that the value in the first row, second column of the matrix represents the relationship between the first and second locations in the map. More generally, the element in row i, column j of the weights matrix, denoted w_{ij}, represents the relationship between location i and location j, so that we have

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ w_{n1} & \cdots & \cdots & w_{nn} \end{bmatrix} \tag{7.4}$$

Each w_{ij} value is dependent on the spatial relationship between locations i and j and on how we choose to represent that relationship. Note that while the order of the locations is arbitrary, the order must be the same for both the rows and columns of the matrix.

With this framework in place, we need to assign values to each matrix element. Most simply, if we use *adjacency*, the w_{ij} values will be 1 if two locations are adjacent and 0 if they are not. Even this simple case is not as straightforward as it seems, as we may choose to require areas to share an edge in order to consider them adjacent (the Rook's case), or we may consider it sufficient that they only meet at a corner vertex (the Queen's case). These cases are shown in Figure 7.5(i) and (ii), respectively, where the extra adjacencies introduced in the Queen's case between polygons that meet at a corner only are apparent. These four polygons have been extracted from the maps of Figure 7.6 (i) and (ii), where the same adjacencies are applied to the 103 Census Area Units of Auckland City, New Zealand.

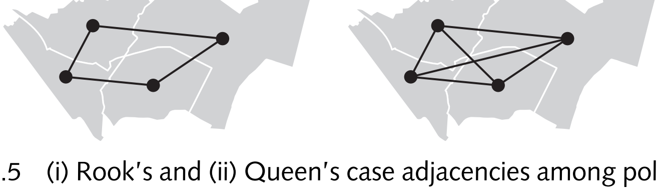


Figure 7.5 (i) Rook's and (ii) Queen's case adjacencies among polygons.

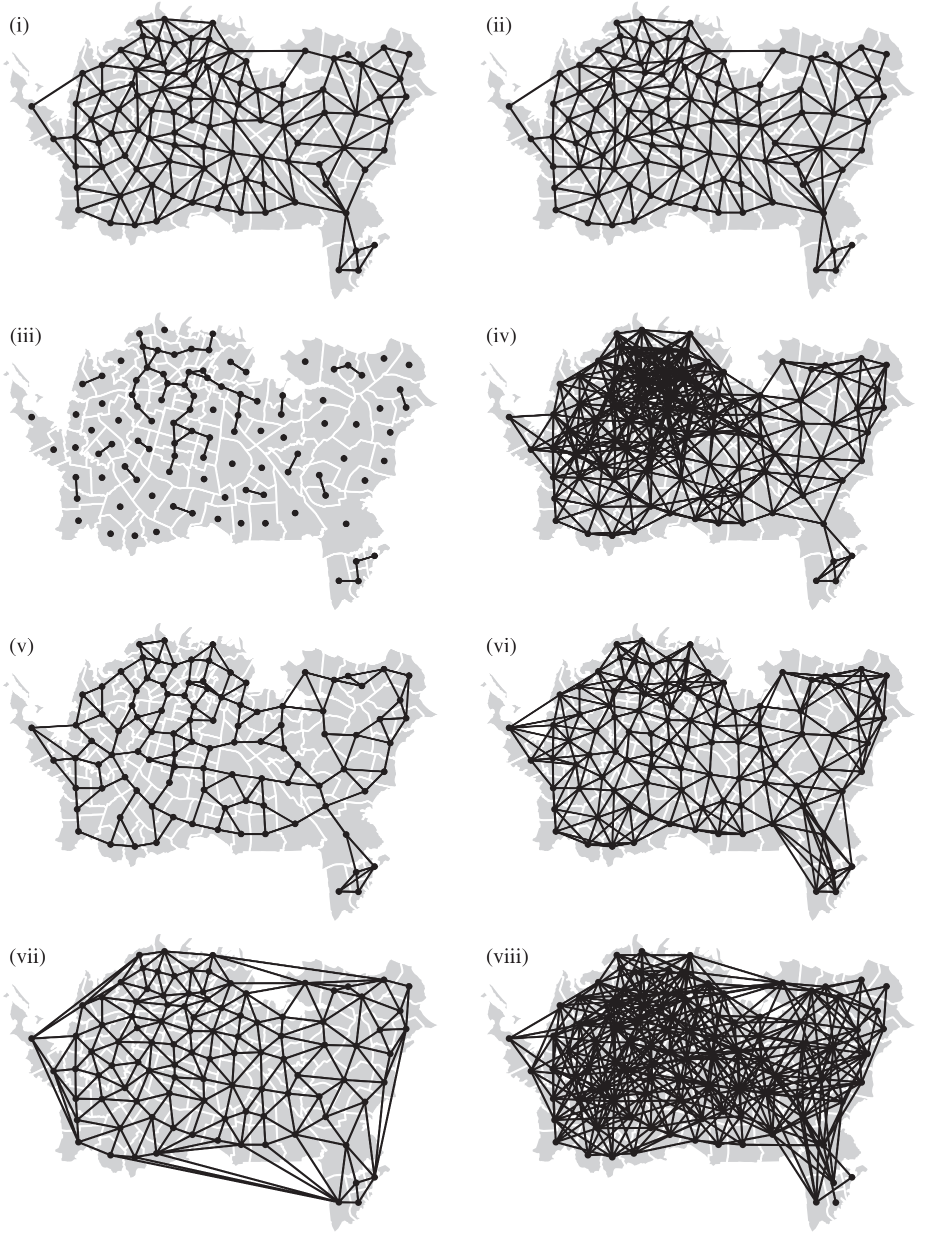


Figure 7.6 Eight alternative spatial structures for the 103 Census Area Units in Auckland City, New Zealand: (i) Rook's case adjacency, (ii) Queen's case adjacency, (iii) center-to-center distance less than 1 km, (iv) center-to-center distance less than 2.5 km, (v) three nearest neighbors, (vi) six nearest neighbors, (vii) Delaunay triangulation, and (viii) lag two Rook's case adjacency.

Alternatively, we may ignore contiguity between the polygons entirely and instead use some measure of distance between polygons. Often, this involves representing the polygon areas as points at the polygon centroid or at some central point on the skeleton and then measuring the distances between the points. Then, based on some distance threshold d, we consider two cases adjacent if $d_{ij} < d$ and not otherwise. In Figure 7.6(iii) and (iv), the adjacencies produced with a distance threshold of 1 km and 2.5 km are shown. The sparse connectivity of the lower distance threshold case is clear. Alternatively, we may wish to include only the nearest neighbors. Figures 7.6(v) and (vi) show the connectivity for three and six nearest-neighbor cases.

More complex cases are also possible. A number of adjacency rules based on the Delaunay triangulation introduced in Section 2.3 can be developed. These are discussed in Bivand et al. (2008, pp. 244–246), and the simplest case is shown in Figure 7.6(vii). The variants discussed by Bivand et al. remove the troublesome longer-distance links that appear around the boundary of the study area when this approach is adopted. Finally, in Figure 7.6(viii), we show the connectivity produced when adjacencies at lag two are used. Zones adjacent at lag two are those that are neighbors "once removed" across an intervening zone. Often the lag two adjacency matrix is denoted $\mathbf{W}^{(2)}$, and, for clarity, its elements are denoted ${w_{ij}}^{(2)}$. It is a trivial matter to find the adjacency matrix at any desired lag using matrix multiplication operations or network shortest path algorithms, although some care is required to avoid counting relationships multiple times at different lags. However, in practice, it is not clear how meaningful analyses based on more remote lags are likely to be. For discussions of yet more ways of constructing $\mathbf{W}$ matrices, see Bavaud (1998) and Getis and Aldstat (2004).

In the cases discussed above, adjacency remains a binary quantity, so that w_{ij} may only take on the values 1 (connected) or 0 (not connected). We may also consider some relationships to be stronger than others, and then the w_{ij} values will range from 0 (for weak interaction) to 1 (for strong interaction). Common ways of weighting the strength of interaction between two locations use an inverse-power relationship and their separation distance. Further complexity can be introduced by considering the length of shared boundaries between adjacent locations, so that

$$w_{ij} \propto \frac{l_{ij}}{d_{ij}^{z} l_i} \tag{7.5}$$

where z is a power factor, l_{ij} is the length of the shared boundary between zones i and j, and l_i is the length of the perimeter of zone i. With this approach, it is necessary to scale the weights so that they all lie in the range 0 to 1. A typical method is to ensure that each row of the matrix sums to 1, as discussed in Section 2.3.

Two important considerations in the construction of the weights matrix are how we deal with the relationship between each location and itself and how we enforce symmetry. Because we are not interested in the relationship between each location and itself, elements on the main diagonal of the matrix (i.e., w_{11}, w_{22}) are usually set to zero. Symmetry in the weights matrix is generally required so that $w_{ij} = w_{ji}$ in all cases. Some methods for constructing the matrix do not guarantee symmetry. For example, in the k nearest-neighbor approach, area A may have areas B, C, and D as its three nearest neighbors, while the three nearest neighbors of B are C, D, and E and do not include A. In this case, $w_{AB} \neq w_{BA}$. To resolve this situation, we can enforce symmetry by setting

$$\mathbf{W}_{\text{final}} = \tfrac{1}{2}\left(\mathbf{W} + \mathbf{W}^{\mathrm{T}}\right) \tag{7.6}$$

so that each pairwise two-way relationship is the average of the two one-way relationships.

Intuitively, we might suppose that the information in a **W** matrix will tell us quite a lot about the spatial patterning of the areas from which it was derived, and recent work has explored aspects of this situation. Although developed in a slightly different context, a key early paper is that by Tinkler (1972), who was concerned with the structure present in connected networks, in which the structure is expressed using a binary connectivity matrix, **C**, recording the presence or absence of a connection between any two specified nodes on a network. The eigensystem (see the Appendix) of such a matrix characterizes the network connectivity. The principal eigenvalue of a **C** matrix gives an overall *connectivity* measure of the set of nodes, and the elements of its eigenvector indicate the centrality within the network of each node in turn (see also Boots, 1983, 1984). Similar interpretations are possible for our **W** matrices (Griffith, 1996; Boots and Tiefelsdorf, 2000).

The important points to appreciate are that a wide variety of spatial weights matrices are possible in any given situation and the choice of spatial weights for use in autocorrelation measurement is a key step in the analysis. In a sense, the choice of **W** represents a hypothesis about the phenomenon being studied, so that ideally, the spatial structure represented in the weights matrix will correspond to some aspect of the problem that is meaningful in terms of the processes under consideration. This is not always easy to arrange, however, and in the study of social processes in particular, census units or other administrative units are often used in the absence of any other convenient approach. Developing a spatial weights matrix that relates to the posited underlying processes will also be difficult where those processes are not well understood. In such cases, it is advisable to work with simple adjacency-based approaches at least in the exploratory phase of the analysis.

Moran's *I*, an Index of Spatial Autocorrelation

Once the spatial structure for analysis has been determined, any particular measure of autocorrelation can be constructed by defining a way of measuring the difference between location attribute values. The most widely used measure is Moran's *I*, which is a translation of a non-spatial correlation measure to a spatial context and is usually applied to areal units where numerical ratio or interval data are available (Moran, 1950). The easiest way to present the measure is to dive straight in with the equation for its calculation and to explain each component in turn.

I is calculated from

$$I = \left[\frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right] \times \left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}\right] \tag{7.7}$$

This equation is fairly formidable, so let's unpick it piece by piece. The important part of the calculation is the second fraction. The numerator on top of this fraction is

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y}) \tag{7.8}$$

which you should recognize as a covariance term. The subscripts i and j refer to different areal units or zones in the study, and y is the data value in each. By calculating the product of two zones' differences from the overall mean ($\bar{y}$), we determine the extent to which they co vary. If both y_i and y_j lie on the same side of the mean (above or below it), then this product is positive; if one is above the mean and the other below, then the product is negative and the absolute size of the resulting total will depend on how close to the overall mean are the zone values. The covariance terms are multiplied by w_{ij}, an element from the spatial weights matrix **W**, which has the effect that the covariance elements are weighted according to how closely related they are spatially. When **W** is an adjacency matrix with $w_{ij} = 1$ if zone i and zone j are adjacent and 0 otherwise, then the covariance term is included in the calculation only for pairs of adjacent locations.

Everything else in the equation normalizes the value of *I* relative to the number of zones being considered, the number of adjacencies in the problem, and the range of values in y. The divisor $\Sigma\Sigma w_{ij}$ accounts for the total spatial weights in the map, and the multiplier

$$\frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7.9}$$

is actually *division* by the overall data set variance, which ensures that I is not large simply because the values and variability in y are high.

The net result of the calculation in Equation (7.7) is that if the data are positively autocorrelated, then most pairs of adjacent locations will have values on the same side of the mean and Moran's I will have a positive value. On the other hand, if the data are negatively auto correlated, most adjacent locations will have attribute values on opposite sides of the mean, and the overall result will be negative. Thus, as for a conventional nonspatial correlation coefficient, a positive value indicates a positive autocorrelation and a negative value a negative or inverse correlation. The value is not strictly in the range -1 to $+1$, as it is impossible for a map to be perfectly autocorrelated, whether positively or negatively, except in very unusual situations. Generally speaking, an index score of 0.3 or more, or of -0.3 or less, is an indication of relatively strong autocorrelation. However, some attention must be paid to the statistical significance of any measure index value, and we discuss this in more detail below.

7.5. AN EXAMPLE: TUBERCULOSIS IN AUCKLAND, 2001–2006

Figure 7.7 shows reported cases of tuberculosis per 100,000 population for Auckland City, New Zealand, in 2001–2006 (note that these are not annual rates, but rates accumulated over the whole six-year period relative to the 2006 census population). As a preliminary stage in the analysis of these data, determining how strongly autocorrelated they are is of interest. Examination of the map suggests that there is a tendency for census areas in the southwest of the city (toward New Windsor) to have experienced higher rates of incidence of tuberculosis. These areas form an arc from near Waterview to Onehunga. There is also a more isolated group of areas around Tamaki in the east, which also have higher incidence rates.

Using the Rook's and Queen's case spatial weights matrices shown in Figure 7.5(i) and (ii), we can determine Moran's I for these data. This would be an arduous calculation to perform by hand, but it is readily carried out in a number of current software packages, such as *ArcGIS*™, *GeoDa*, and packages available in the *R* statistics environment. The calculated result in the *R*

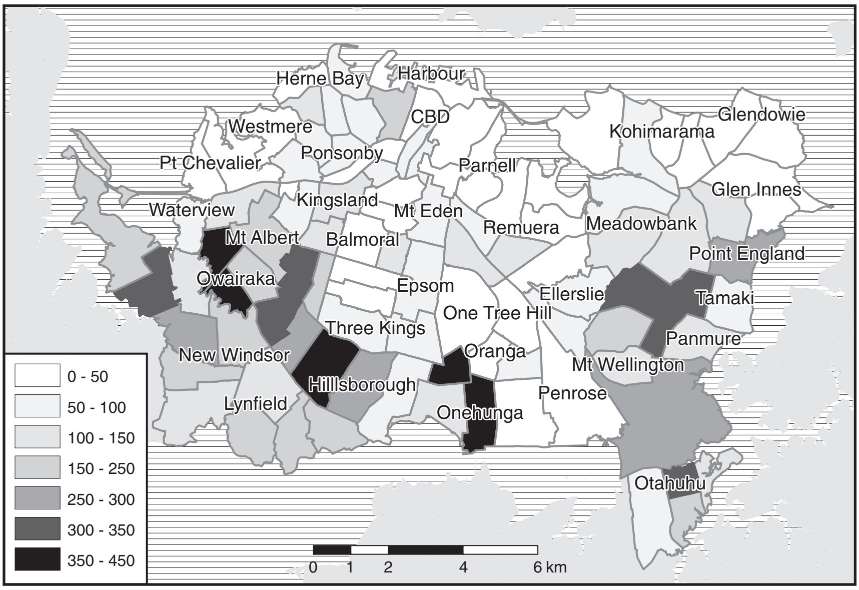


Figure 7.7 Reported cases of tuberculosis per 100,000 population, Auckland City, 2001–2006. The polygons are New Zealand Statistics census area units.

package **spdep** is 0.383, for the Rook's case, and 0.394, for the Queen's case. As visual inspection of the map suggests, both results are evidence for positive spatial autocorrelation.

Before considering more closely how we can associate a level of statistical significance with these results, it is instructive to study the scatterplot in Figure 7.8. This is a *Moran scatterplot* showing the relationship between the attribute values themselves (horizontal axis) and the local mean attribute value (i.e., the mean value of the adjacent locations). This graph has four regions: the lower-left quadrant contains cases where the attribute value in each polygon and the mean attribute value of neighboring polygons are less than the global mean; the upper-right quadrant contains cases where both the attribute value and the local mean are greater than the global mean; and the other two quadrants contain cases where the attribute value and the local mean lie on opposite sides of the global mean. Locations that lie in the lower-left and upper-right quadrants are those that contribute to overall positive autocorrelation, since they have an attribute value similar to that of their neighbors, while locations in the other two quadrants contribute to a negative autocorrelation. If, as here, most locations are in the lower-left and upper-right quadrants, then the overall outcome will most likely be a positive value of Moran's I, indicating positive autocorrelation.

In Figure 7.8, particular locations have been identified that contribute strongly to the measured positive autocorrelation. You should be able to

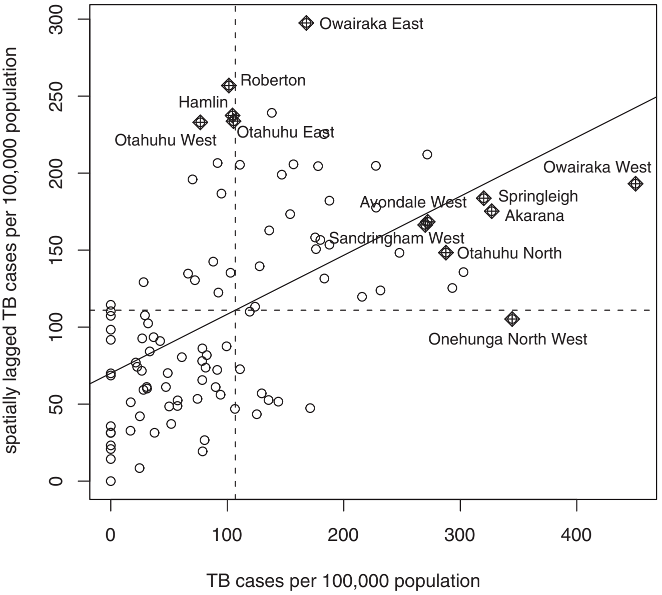


Figure 7.8 Moran scatterplot for the tuberculosis data.

deduce from the map, and from the plotted values, which census area units are Owairaka West and East and also Onehunga North West. In one of the most widely used programs for performing this type of analysis, *GeoDa*, this exploration can be performed interactively via *linked brushing* (see Section 3.4), with a selection area in the Moran scatterplot highlighting associated regions in a map view.

It is worth noting that Moran's *I* is effectively the correlation coefficient for the relationship between the attribute values and the local mean attribute values. If you are familiar with the statistical theory behind regression, this is clear if the equation for Moran's *I* is rewritten in matrix form as

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\mathbf{y}^T \mathbf{W} \mathbf{y}}{\mathbf{y}^T \mathbf{y}} \qquad (7.10)$$

where $\mathbf{y}$ is the column vector whose entries are each $(y_i - \bar{y})$. In statistics, this is a common formulation of this type of calculation and you may encounter it in the literature (see Anselin, 1995).

This insight allows the standard diagnostic statistics from linear regression to be used to associate p-values with observed values of Moran's I. However, because the spatial structure of the map is also a parameter in the analysis, a more usual approach is based on a Monte Carlo procedure (see Section 5.4). The location attribute values can be permuted any required

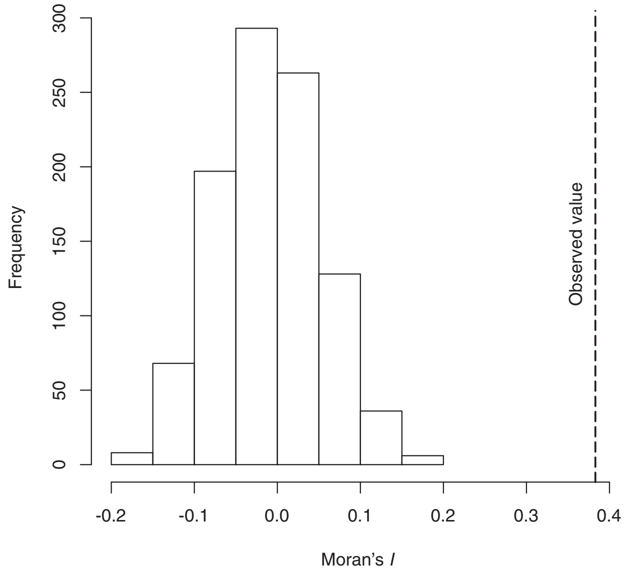


Figure 7.9 Comparison of the observed value of Moran's *I* with the values produced under 999 random permutations of the data.

number of times (999 is typical), that is, the attribute values observed in the map are randomly assigned to the map locations, and Moran's *I* is recalculated each time for the "scrambled" map. This gives an empirical sampling distribution for the index and allows the observed Moran's *I* to be assessed in terms of how unusual it is relative to this randomized benchmark. Figure 7.9 is a histogram of the distribution of 999 random permutations of the data in this map, with the vertical line showing the value of Moran's *I* observed for the actual data. It is clear that the observed value is very unusual with respect to the randomized data, which means that we regard the finding of positive autocorrelation as statistically significant.

In Table 7.2 the Moran's *I* results are shown for a number of the other spatial weighting schemes presented in Figure 7.6. As we might expect, the results are consistent across all weighting schemes, since they all emphasize immediate neighboring locations. The only slight exception is the distance threshold of 2.5 km, which has a somewhat lowered value.

If we determine the index value for a range of spatial lags based on Rook's adjacency and plot these results, with error bars (determined by a Monte Carlo procedure), we obtain the result shown in Figure 7.10. This confirms the strength of the finding of positive autocorrelation, since it holds at lags of up to three. This figure also suggests that if you move four "steps" away from a census area unit, then you are likely to come to an area where the incidence rate of tuberculosis is unrelated to that in the original location. After two

Table 7.2 Moran's *I* for Various Spatial Weighting Schemes: TB in Auckland, 2000–2006

Spatial weighting scheme	*Figure*	*Moran's I*
Rook's adjacency	7.5(i)	0.3830
Queen's adjacency	7.5(ii)	0.3941
$d < 2500$ m	7.5(iv)	0.3510
$k = 3$ nearest	7.5(v)	0.3780
$k = 6$ nearest	7.5(vi)	0.4014
Delaunay triangulation	7.5(vii)	0.3846

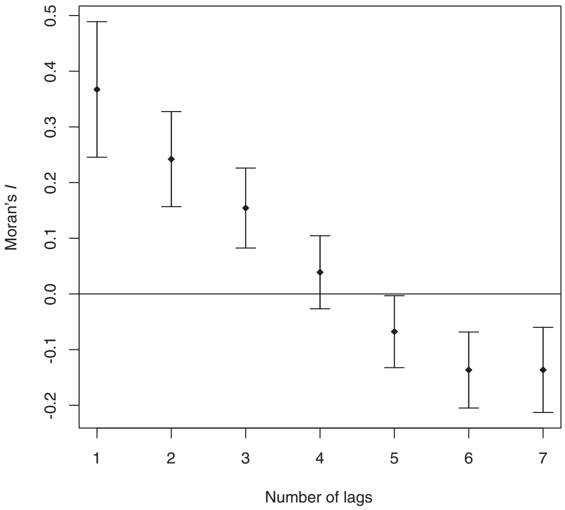


Figure 7.10 Moran's *I* at different spatial lags for the Auckland data.

further steps, the new location's incidence rate is likely to be opposite to the original locations, so that there is a slight negative autocorrelation at this spatial lag. Another way of thinking about this is that the scale of the regions of higher or lower relative rates of incidence is such that they do not extend more than about four census area units in any direction. This is in accordance with the pattern we see in the map. This type of analysis is closely related to semivariogram analysis, which is a preliminary step in the advanced interpolation methods considered in Chapter 10.

7.6. OTHER APPROACHES

Although it is the measure most frequently used and has possibly the most attractive properties, Moran's *I* is not the only spatial autocorrelation measure. An alternative is Geary's C. This is similar to I and is calculated from

$$C = \left[\frac{n-1}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \right] \times \left[\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - y_j)^2}{2 \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \right] \tag{7.11}$$

As with Moran's I, the first term is a variance normalization factor to account for the numerical values of y. The second term has a numerator based on the square of the *difference* in y between the two areas under consideration, and is greater when there are large differences between adjacent locations. The denominator, $2\ \Sigma\Sigma w_{ij,}$ normalizes for the combined spatial weights in the map. Geary's C can be confusing in one respect: a value of 1 indicates *no* autocorrelation; values *less* than 1 (but greater than or equal to 0) indicate positive autocorrelation, and values *more* than 1 indicate negative autocorrelation. The reason for this is clear if you consider that the $\sum w_{ij} (y_i - y_j)^2$ term in the calculation is always positive but gives smaller values when similar values are neighbors. Geary's C can easily be converted to the more intuitive ±1 range by subtracting the value of the index from +1.

In situations where interval or ratio data are not available, or where some threshold value of the attribute is of particular interest, so that areas above and below the threshold can be treated as binary outcomes, another possible approach is the *joins count test*. This approach is based on counting the number of occurrences of neighboring pairs of polygons in the various different possible categories. In the binary case, where we can characterize the two available states as "black" and "white," we arrive at counts of the number of black–black, white–white, and black–white neighbor joins. The observed counts can be compared to the expected numbers to assess the type and strength of autocorrelation present. Positively autocorrelated maps will have more black–black and white–white joins than expected, while negatively autocorrelated maps will have fewer such joins and more black–white joins than expected. Cliff and Ord (1973) and Unwin (1981) provide a full account of this approach.

Joins counting is very similar to one of the measures of spatial pattern provided by FRAGSTATS, which is a count of the number of neighboring like pairs of grid cells in a raster. Joins counts methods are rather limited, however, since they only apply to categorical data, and are not easy to handle when there are more than a small number of categories (2 or 3) because of the large number of possible types of join that quickly arise (for example, with 6 categories, 15 join types are possible, and with 12 there are 66!

CHAPTER REVIEW

- Area objects of interest come in many guises, with a useful—though sometimes ambiguous—distinction between those that arise naturally and those that are arbitrarily imposed for the purposes of data collection.
- Area objects have geometric and topological properties that can be useful in description. The most obvious of these is the object's *area*, but we can also find the *skeleton* and *centroid*, and attempt to characterize their *shape*. If there are many area objects in a pattern, measures of *fragmentation* and other spatial pattern metrics may also be used.
- *Autocorrelation* is a key concept in geography, so much so that, arguably, a test for autocorrelation should always be carried out before further statistical analysis of geographic data.
- Any autocorrelation measure must be based on both the *spatial structure* of the geographic objects in the study and the *similarity or difference of attribute values* at locations near one another.
- The spatial structure of the study area is usually represented by constructing a *spatial weights matrix* based on the *adjacency* or *interaction* between locations.
- A wide variety of methods for defining spatial weights is available. Binary outcomes (0 or 1) can be produced using *Rook's or Queen's adjacency*, as well as *distance thresholds* or *nearest-neighbor* rules. The Delaunay triangulation is the basis of several other approaches.
- Adjacency-based matrices can also be calculated with *spatial lags* that record locations that neighbor one another via a number of intervening connected locations.
- *Continuously variable spatial weights* between 0 and 1 may also be used. They are usually based on some combination of distance and, optionally, on the length of the shared boundary between polygons.
- The most widely used measure of spatial autocorrelation is Moran's I, which employs a *covariance* term between each areal unit and its neighbors. A value of zero indicates random arrangement, a positive value positive autocorrelation, and a negative value negative autocorrelation.
- Geary's C uses the sum of squared differences between each areal unit and its neighbors. A value of 1 indicates no autocorrelation, values between 0 and 1 indicate positive autocorrelation, and values between 1 and 2 indicate negative autocorrelation.

- For many autocorrelation measures, Monte Carlo simulation is the most appropriate way to determine statistical significance. In this context, simulation consists of randomly shuffling the observed areal unit values and recalculating the autocorrelation measure(s) of interest to determine a sampling distribution.

REFERENCES

Anselin, L. (1995) Local indicators of spatial association—LISA. *Geographical Analysis*, 27: 93–115.

Bavaud, F. (1998) Models for spatial weights: a systematic look. *Geographical Analysis*, 30: 152–171.

Berry, J. K., Buckley, D. J., and McGarigal, K. (1998) Fragstats.arc: Integrating ARC/INFO with the Fragstats landscape analysis program. *Proceedings of the 1998 ESRI User Conference*, San Diego, CA.

Bivand, R. S., Pebesma, E. J., and Gomez-Rubio, V. (2008) *Applied Spatial Data Analysis with R* (New York: Springer).

Boots, B. N. (1977) Contact number properties in the study of cellular networks. *Geographical Analysis*, 9: 379–387.

Boots, B. N. (1983) Comments on using eigenfunctions to measure structural properties of geographic networks. *Environment and Planning*, *Series A*, 14: 1063–1072.

Boots, B. N. (1984) Evaluating principal eigenvalues as measures of network structures. *Geographical Analysis*, 16: 270–275.

Boots, B. and Tiefelsdorf, M. (2000) Global and local spatial correlation in bounded regular tessellations. *Journal of Geographical Systems*, 2: 319–348.

Boyce, R. and Clark, W. (1964) The concept of shape in geography. *Geographical Review*, 54: 561–572.

Cerny, J. W. (1975) Sensitivity analysis of the Boyce-Clark shape index. *Canadian Geographer*, 12: 21–27.

Clark, W. and Gaile, G. L. (1975) The analysis and recognition of shapes. *Geografiska Annaler*, 55B: 153–163.

Cliff, A.D. and Ord, J.K. (1973) *Spatial Autocorrelation* (London: Pion).

Coppock, J. T. (1955) The relationship of farm and parish boundaries: a study in the use of agricultural statistics. *Geographical Studies*, 2: 12–26.

Coppock, J. T. (1960) The parish as a geographical statistical unit. *Tijdschrift voor Economische en Sociale Geographie*, 51: 317–326.

DeZonia, B. and Mladenoff, D. J. (2004) IAN—raster image analysis software program. Department of Forest Ecology and Management, University of Wisconsin, Madison, WI. Available at http://landscape.forest.wisc.edu/projects/IAN/.

Frolov, Y. S. (1975) Measuring the shape of geographical phenomena: a history of the issue. *Soviet Geography*, 16: 676–687.

Frolov, Y. S. and Maling, D. H. (1969) The accuracy of area measurement by point counting techniques. *Cartographic Journal*, 6: 21–35.

Garreau, J. (1992) *Edge City* (New York: Anchor).

Getis, A. and Aldstat, J. (2004) Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, 36: 90–104.

Gierhart, J. W. (1954) Evaluation of methods of area measurement. *Survey and Mapping*, 14: 460–469.

Griffith, D. (1996) Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Canadian Geographer*, 40: 351–367.

Houghton, R. A., Skole, D. L., Nobre, C. A., Hackler, J. L., Lawrence, K. T., and Chomentowski, W. H. (2000) Annual fluxes or carbon from deforestation and regrowth in the Brazilian Amazon. *Nature*, 403(6767): 301–304.

Lee, D. R. and Sallee, G. T. (1970) A method of measuring shape. *Geographical Review*, 60: 555–563.

McGarigal, K., Cushman, S. A., Neel, M. C., and Ene, E. (2002) *FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps*. Computer software program produced at the University of Massachusetts, Amherst. Available at www.umass.edu/landeco/research/fragstats/fragstats.html.

Medda, F., Nijkamp, P., and Rietveld, P. (1998) Recognition and classification of urban shapes. *Geographical Analysis*, 30(4): 304–14.

Moran, P. A. P. (1950) Notes on continuous stochastic phenomena. *Biometrika*, 37: 17–33.

Nepstad, D. C., Verissimo, A., Alencar, A., Nobre, C., Lima, E., Lefebvre, P., Schlesinger, P., Potter, C., Moutinho, P., Mendoza, E., Cochrane, M., and Brooks, V. (1999), Large-scale impoverishment of Amazonian forests by logging and fire. *Nature*, 398(6727): 505–508.

Smith, B. and Varzi, A. C. (2000) Fiat and bona fide boundaries. *Philosophy and Phenomenological Research*, 60(2): 401–420.

Tinkler, K. (1972). The physical interpretation of eigenfunctions of dichotomous matrices. *Transactions of the Institute of British Geographers*, 55: 17–46.

Tobler, W. R. (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46: 234–240.

Turner, M. G., Gardner, R. H., and O'Neill, R. V. (2001) *Landscape Ecology in Theory and Practice: Pattern and Process* (New York: Springer-Verlag).

Unwin, D. J. (1981) *Introductory Spatial Analysis* (London: Methuen).

Wentz, E. A. (2000) A shape definition for geographic applications based on edge, elongation and perforation. *Geographical Analysis*, 32: 95–112.

Wood, W. F. (1954) The dot planimeter: a new way to measure area. *Professional Geographer*, 6: 12–14.

Worboys, M. F. and Duckham, M. (2004) *GIS: A Computing Perspective* (London: Taylor & Francis).

Zhang, D. and Lu, G. (2004) Review of shape representation and description techniques. *Pattern Recognition*, 37(1): 1–19.

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 8

Local Statistics

CHAPTER OBJECTIVES

In this chapter, we:

- Explain the concepts underlying the emerging array of *local statistics*
- Account for the relatively late arrival of local statistics on the spatial analytic scene
- Review the various approaches that can be used to construct *localities* for the development of local statistics
- Discuss how the popular Getis-Ord family of G statistics are calculated and interpreted
- Outline the local version of Moran's I statistic
- Explain why inference based on local statistics is challenging and describe current approaches to dealing with the difficulties
- Provide an overview of the increasingly popular method *geographically weighted regression*
- Explain how many other spatial analysis methods can be considered as local statistics even if this was not the intent behind their original development

After reading this chapter, you should be able to:

- Explain what is meant by local statistics and suggest reasons for their current popularity
- Provide some explanations for the slow adoption of local statistical approaches
- Review a number of bases on which localities for local statistics can be constructed

- Define the Getis-Ord G and local Moran's I statistics and discuss how they should be interpreted
- Give an account of why inference for local statistics is difficult and outline approaches for addressing the various problems
- In general terms, describe how geographically weighted regression works
- Redescribe a selection of other spatial analytical methods as local statistics

8.1. INTRODUCTION: THINK GEOGRAPHICALLY, MEASURE LOCALLY

We now consider one of the most important innovations in geographic information analysis in recent years, namely, the development and use of a variety of *local statistics*. As we will see, local statistics arise naturally out of any of the methods for measuring spatial autocorrelation, discussed in the previous chapter. Once the connection is noted and generalized, the way is open to development of localized variants of almost any standard summary statistic, with a particularly interesting recent innovation being *geographically weighted regression*. In the same way, many older spatial analysis methods can also be reinterpreted as local statistics. Thus, this chapter is also a useful precursor to the interpolation methods discussed in more detail in the next chapter, because many estimates produced by spatial interpolation are based on local statistics.

What do we mean by a local statistic? A *local statistic* is any descriptive statistic associated with a spatial data set whose value varies from place to place. In the broadest sense, any spatial data set is a collection of local statistics, in that the recorded attribute values are different at each location. A local statistic is different in that it is derived by considering a subset of the spatial data *local* to the spatial location where it is being calculated. A simple example is a localized mean, calculated by determining the mean value of an attribute based on attribute values in the data set near the location of interest. In the next chapter, we will see that such a localized mean is the underlying basis for many simple methods of spatial interpolation. It is also worth noting that a localized mean is exactly equivalent to one kind of smoothing filter that may be applied to image or raster data. In Section 9.5, we will see that this concept has been generalized as *map algebra*, specifically in the form of focal operations on raster data. Thus, the concept of a local statistic is widely deployed in spatial analysis, although it goes by different names in different contexts. For now, the important point is to realize how central the concept of a local statistic is for many spatial analysis methods.

It is perhaps surprising that such a key idea has only gained currency since the mid-1990s. Two review papers in *Progress in Human Geography* by Unwin (1996) and Fotheringham (1997) were the first to highlight explicitly the importance of local statistics. With this in mind, it is useful to consider why the idea has only taken off recently. One important consideration is the mapping capability provided by GIS tools. As we shall see, many local statistics are a natural by-product of the calculation of summary global statistics. Prior to the possibility of easily mapping any data set, the summary result would have been reported and the local statistics used in its calculation discarded. It was the advent of readily available mapping tools that led to the exploration of the potential of local statistics as an analytical output in their own right. These developments parallel the increasing importance of exploratory data analysis (Tukey, 1977), an approach where indentifying outliers and the overall structure in data are important aims and visualization methods are central.

A second technical reason for the recent increase in the popularity of local statistics is that (perhaps paradoxically) the statistical evaluation of local statistics is more challenging than the statistical assessment of related global measures. This parallels the difficulties faced in identifying local clusters in point patterns relative to simply determining if a point pattern is clustered or not. The difficulty lies in assessing analytically the statistical significance of particular localized patterns, and in this context, Monte Carlo simulation approaches are often used to generate pseudo-significance results. The computational burden of simulation-based methods is significant, so local statistics have only become practical as substantial computing power has become generally available.

A third reason for the increased interest in local statistics is recognition of the importance of geographic variation in phenomena. This is itself a side-effect of the widespread adoption and use of GIS tools and the accompanying increase in data availability. As more data have become available, this has allowed studies both to expand their spatial range and to focus in at higher spatial resolution. Both developments have prompted the realization that the idea of a single global process or model being a realistic explanation is not always very plausible.

Finally, interest in local statistics reflects developments in the spatial sciences more generally that increasingly recognize the importance of local contexts in understanding the global patterns of phenomena. While the methods deployed by many human geographers in their research have become increasingly qualitative since the 1980s, this is largely a response to the realization that local contexts matter. Qualitative methods such as interviews and focus groups recognize the multilayered richness of local contexts and the importance of multiple interpretations by people in those contexts. By their very nature, quantitative or statistical data tend to simplify and flatten out some of that complexity. This is particularly the case if we restrict ourselves to global summary statistics and a search for broad generalizations about the whole of a large study area. Local statistics, on the other hand, emphasize the variety among local contexts and focus attention on what is different from one place to another rather than on what is similar. Thus, local statistics can be seen as a response from the quantitative end of geography to increased recognition of the importance of local context.

8.2. DEFINING THE LOCAL: SPATIAL STRUCTURE (AGAIN)

In Section 7.4, as a necessary precursor to the development of global spatial autocorrelation statistics, the construction of a wide variety of spatial weights matrices among a set of polygons was described. The local neighborhood of a particular location is fully described by a single row in such spatial weights matrix. Thus, if the weights matrix is

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ w_{n1} & \cdots & \cdots & w_{nn} \end{bmatrix} \tag{8.1}$$

then a row matrix

$$\mathbf{W}_i = [\, w_{i1} \quad w_{i2} \quad \cdots \quad w_{in} \,] \tag{8.2}$$

describes the local neighborhood of each location i. As has been discussed in Chapters 2 and 7, there is a wide range of possible bases on which the weights matrix and thus localities may be defined.

Some More Revision

It is useful at this point to revisit Section 2.3 and revise the materials on definitions of distance, adjacency, neighborhood, and proximity, as well as how these can be summarized in adjacency A or weights matrices **W**.

When you have done this, revisit Section 7.4 to remind yourself of how these matrices are used in the definition of global spatial autocorrelation indices.

For polygon data, adjacency either immediately or indirectly via intervening neighboring polygons is a common basis for construction of localities. Using polygon centroids, localities may be constructed based on distance criteria, and the same approach can also be applied to point data sets. In this case, it is also possible to introduce additional constraints so that all locations have some minimum number of neighbors in their locality.

It is important to keep in mind that the choices made in constructing localities prior to determining local statistics are a critical aspect of the analysis. Local statistics may point to patterns of a particular kind when localities are constructed based on adjacency, but they may reveal completely different patterns when localities are constructed based on a distance criterion. The important points are, first, that where possible, a number of different weights matrix constructions be examined, and second, that consideration be given to which method makes the most sense in substantive terms. For example, it is easy to assume that simple spatial adjacency based on contiguity among a set of polygons is somehow the “natural” approach to constructing localities. However, when we are interested in some phenomenon whose patterns are likely to be related to transport accessibility, it may be much more relevant to connect locations via the transport network, and thus to base adjacency on estimated distances over road or other networks. Such options have become much more readily explored using the capacity of GIS to relate spatial data in a wide range of ways.

8.3. AN EXAMPLE: THE GETIS-ORD G_i AND G_i^* STATISTICS

The goal of the Getis-Ord family of local statistics developed in Getis and Ord (1992) and Ord and Getis (1995) is to enable detection of local concentrations of high or low values in an attribute, and it nicely illustrates the concept of a local statistic. The statistic is simple to calculate. For a location i, the value is given by

$$G_i(d) = \frac{\sum_j w_{ij}(d) x_j}{\sum_{j=1}^{n} x_j} \quad \text{for all } i \neq j \tag{8.3}$$

where $w_{ij}(d)$ are weights from the spatial weights matrix and x_j denotes the attribute values at locations j. The dependence on a particular set of assumptions about spatial dependence is denoted for both G_i and w_{ij} by their functional dependence on d. Note that the numerator in this fraction is the sum of the x_j values in the locality of the location of interest i but not including x_i itself, and that the denominator is the sum of all the x values in the whole study area. Thus, G_i is simply the proportion of the sum of all x values in the study area accounted for by just the neighbors of i. In a location where high values are clustered, G_i will be relatively high; conversely, in a location where low values are concentrated, G_i will be low. The closely related statistic G_i^* is defined similarly to G_i, the only difference being that the attribute value at location i itself is included in both the numerator and denominator summations in Equation (8.3). Due to the dependence of G_i (and G_i^*) on the ratio of two sums of x values, it is important that the attribute under consideration be a ratio-scaled variable with a natural origin. Another way to think about this is that the value of G_i will be different if we add a constant value to every location or if we transform the variable by taking logarithms.

It is relatively easy to derive expected values and the variance for the G_i statistic under an assumption of random spatial distribution of the attribute values. The expected value is given by

$$E(G_i(d)) = \frac{\sum_j w_{ij}(d)}{n-1} \tag{8.4}$$

which states that the expected value of G_i is that proportion of the study region accounted for by the neighborhood of location i where we assume 0/1 valued adjacency weights. Calculation of the statistic's variance is more complex (see Getis and Ord, 1992, p. 191 for details).

In all the equations above, d denotes the fact that we can calculate the value of the statistic for various distances or, more generally, under a variety of definitions of locality. Thus, as discussed in the previous section, the spatial weights matrix is chosen by the analyst under some assumption of what constitutes a meaningful set of localities for the purposes of the analysis. In essence, the choice of a **W** encapsulates a hypothesis about both the range and nature of any likely local geographic effects. Typically, this will be distance-dependent in some way, insofar as we are interested in knowing if concentrations of high or low data values occur only at short distances, over a wide range of distances, or perhaps only at large distances.

Because expected values and variances for the Getis-Ord statistic are known, a z score can be determined for each location's G_i value. The map in Figure 8.1 shows calculated G_i values as z scores based on the tuberculosis incidence data considered in the previous chapter, using the Rook's case adjacency weights matrix. The notable difference between this map and the map of incidence rates themselves shown in Figure 7.7 is that the

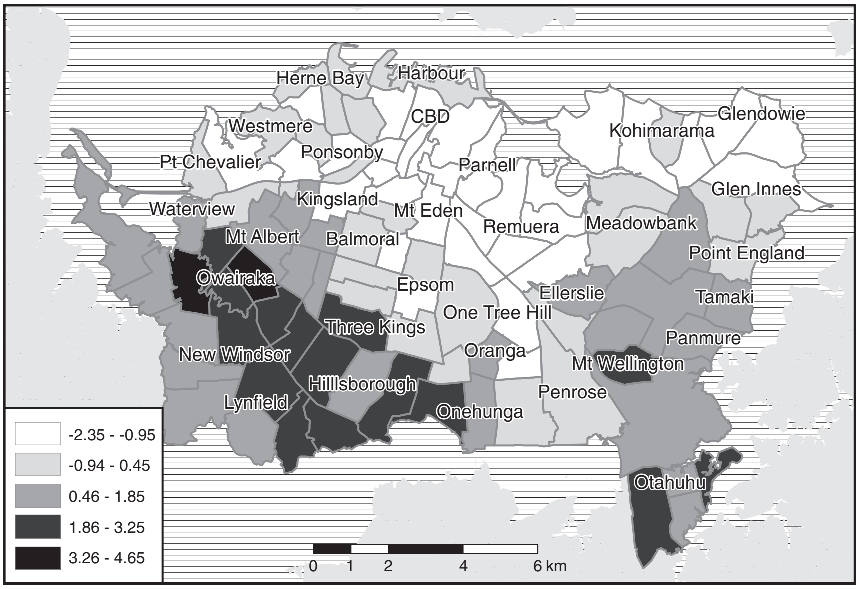


Figure 8.1 Map of the Auckland tuberculosis data *z* scores determined from the calculated G_i values.

highest-incidence locations are not the ones with the highest associated G_i values. Instead, the census area units neighboring high-incidence areas are highlighted. This is most obvious in the case of Mt. Wellington, toward the southeast of the area, which has a high z score in spite of having a relatively low incidence rate. The two particularly high Getis-Ord statistic scores in the western part of the map are also not locations among the highest incidence rates in the original map, but they do have neighboring locations with high incidence rates.

Although we might normally interpret z scores outside the range -1.96 to +1.96 as unusual cases and single out these parts of the map for particular attention, more care is required in making inferences from local statistics. This is because an assumption of normality in the distribution of most local statistics is problematic, particularly where the localities under consideration are small, so that the statistic is being calculated based on small numbers of cases. If d is increased so that localities have larger numbers of cases, this is less of a problem, but, of course, the localities under consideration are no longer quite so local either! It is particularly important to be cautious about overinterpreting high or low z score values toward the edges of the study area, as these may be based on very small numbers of locations. These considerations limit the usefulness of analytical approaches to the statistical assessment of the statistic, leading to a need for simulation-based approaches to inference. This difficulty is common in the interpretation and analysis of local statistics and is considered in more detail in Section 8.4.

Other Local Statistics

In principle, almost any standard statistic can be turned into a local statistic. Instead of summarizing over a whole data set, we summarize over only the data in the locality of each data point. The calculations required to determine a global value of Moran's I provide another example. In this case, at each location, the following quantity is calculated:

$$I_i = z_i \sum_j w_{ij} z_j \tag{8.5}$$

where the z values are z scores determined from the values of the attribute of interest for the whole data set. Positive values of I_i result where either low or high values of the attribute are near one another, while negative values result where low and high values are found in the same area of the map. Thus, local Moran's I gives an indication of data homogeneity and diversity. This statistic is fully developed in a paper by Luc Anselin (1995), which presents the more general concept of *local indicators of spatial association* (LISA) statistics.

When working with the local version of Moran's I, the *Moran scatterplot* of Figure 7.8 comes into its own as an analytical tool. The four quadrants of the plot defined by the global mean attribute value (or by $z_i = 0$ and $\sum_j w_{ij} z_j = 0$) each correspond to the different possible combinations of the value at i and among its neighbors. We can identify these in shorthand as "low–low" or "high–high" cases that contribute to positive autocorrelation and "low–high" or "high–low" cases that contribute to negative autocorrelation. For example, "high–high" cases are ones where the value of i is high and neighboring values are also high. While all cases are in one of the four quadrants, in general, we are only interested in cases that are statistically unusual in some sense. How these cases are identified is discussed in Section 8.4 when we consider inference for local statistics.

Getis and Ord (1992, pp. 198–199) suggest that both the G_i and I_i statistics should be used in any exploration of a spatial data set, as they measure different things and may point to different driving processes underlying the observed spatial distributions of attribute values. While Moran's global statistic measures spatial autocorrelation without distinguishing between patterns dominated by concentrations of high or low values, a global version of the G_i statistics enables these cases to be distinguished. This is clearer when the global G statistic is written out in full:

$$G(d) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(d)x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j} \quad \text{for all } i \neq j \tag{8.6}$$

This statistic tends to have a high value when the locations where high values are located near one another outweigh the locations where low values are located near one another (and vice versa). Thus, G helps to determine whether it is clusters of high values ("hot spots") or low values ("cold spots") that contribute most to an overall finding of positive spatial autocorrelation.

While a local form of almost any statistic could be developed in principle, in practice relatively few have been formalized as local statistics per se, although, as we shall see, many statistics can be usefully considered as local statistics even if they are not used in the exploratory manner discussed here. The likely reason for this is that the most pertinent feature of any spatial data set is the degree to which attribute values exhibit spatial dependence, and this is precisely the aspect of the data on which the Getis-Ord and Moran's statistics focus. Another reason is that local statistics are most useful as exploratory tools in the early stages of an investigation. More formal statistical analysis calls for some method by which significance can be determined, and this presents difficult problems in the context of the small number of cases included in each localized calculation.

8.4. INFERENCE WITH LOCAL STATISTICS

We have seen how the simplifying assumption of spatial randomness can allow analytical results for the expected values of the G_i local statistics to be determined. Such simplifying assumptions treat local statistics as simple random samples from the total population of all the attribute values in the study area. In many cases, because of the central limit theorem, this results in the expectation that local statistics will be normally distributed, and that unusual cases can be identified where calculated z values are less than -1.96 or greater than $+1.96$, which is the range of values associated with standard 95% confidence intervals.

However, this approach is problematic. It is evident that this is the case when Figure 8.1 is reviewed. Six census areas have z score values less than -1.96, while 14 have values greater than 1.96. On a conventional interpretation, this would suggest that almost 20% of the locations (20 out of 103 census area units) are unusual in a statistical sense. The difficulty here is that the data are evidently not well accounted for by a null model that assumes complete spatial randomness. We already know this from the calculations detailed in Section 7.5, which show significant positive spatial autocorrelation in this data set. If we know that the data are positively autocorrelated, it makes little sense to identify statistically unusual cases based on a null model that assumes complete spatial randomness!

Another difficulty is that this is a situation where repeatedly applying a statistical test to the same data, when that test assumes independence of the observations, leads to problems (as was mentioned in relation to the GAM in Section 6.7). Consider any two locations, A and B, that are neighbors in a spatial data set. If A has an unusually high value of the G_i statistic, then, given that it shares many of the same neighbors as B, it is highly likely that B will also have a high value of the G_i statistic. Thus, statistical tests of local statistics are inherently nonindependent, and we must make some adjustment to the criteria we use to determine which observations are unusually high or low. This is known as the *multiple testing problem*, which can be addressed by adjusting the probability threshold used to determine which results are considered statistically significant. Where n tests are conducted, with a desired statistical significance (i.e., a p-value) of α, one possible corrected significance level suggested by Sidak (1967) and endorsed by Ord and Getis (1995) and by Anselin (1995) is

$$a' = 1 - (1 - a)^{1/n} \tag{8.7}$$

An alternative simpler approach, known as the *Bonferroni correction*, is to set $a' = a/n$. In practice, the two approaches produce very similar corrected significance levels. In Figure 8.1, where $n = 103$, for a 0.05 significance level, applying the correction of Equation (8.7) gives an adjusted p-value of 0.000498, while the Bonferroni correction produces a similar value of 0.000485. The former value is associated with a z score of ±3.29.

Applying this new threshold to the mapped data results in only the two highest-value census area units (those located east and west of Owairaka) being considered statistically significant cases. We would interpret this as meaning that even given the known positive autocorrelation in these data, those locations exhibit unusually high similarity to their neighbors. Some writers consider these corrections for multiple testing to be too conservative for the *semi-independent* tests actually applied in the context of local statistics, and also believe that they result in too few determinations of statistically unusual cases. Anselin (1995, p. 96) discusses this matter in some detail. The crux of the argument is that the standard corrections suggested above are for cases where *exactly the same data* are tested n times. For local statistics, overlapping subsets of the same data are tested a number of times, *but never exactly the same subsets*. A rough estimate of the "effective" number of multiple tests actually occurring might be $\sum_i \sum_j w_{ij}/n$, where n is the number of locations in the data set. However, no detailed research results have been reported in this area.

A more recent (and liberal) approach to the statistical assessment of local statistics is to apply a Monte Carlo simulation procedure to produce pseudosignificance values. This is the same approach adopted in assessing many point pattern measures discussed in Section 5.4. For local statistics, the approach is typically repeated using *conditional permutation* (or "shuffling") of the attribute values in the spatial data among the locations in the data set. Each time the data are shuffled, the value at the location of interest is held constant (this is what makes the permutation conditional), the calculations for the statistic in question are performed on the shuffled data, and the resulting value of the local statistic is determined. The permutation procedure is repeated a large number of times (say, 999), and the value of the local statistic associated with the actual distribution of the attribute is ranked relative to the list of values produced by the permutation procedure. Measured actual values of the local statistic that are either very low or very high relative to the list of results produced by the shuffling procedure are then judged to be of interest. A pseudosignificance value can be determined by noting the rank of the actual local statistic relative to the permuted results. For example, if the actual local statistic is the highest recorded among 999 permutations, then it is estimated to be a 1 in 1000 occurrence with a pseudosignificance of $p \sim 0.001$.

While this approach is more computationally intensive than the results derived from analytical expected values and variances, it is conceptually more satisfying and has become routine given contemporary computational resources on the desktop. With appropriate adjustment of the permutation procedure, the simulation approach also has the potential to be used to explore how unusual are the values of local statistics given the presence of known levels of global spatial autocorrelation. This is an aspect of the results reported by Anselin (1995, pp. 108–111) that deserves more attention, since his results clearly demonstrate that the distributional characteristics of local statistics can be expected to depend strongly on global levels of autocorrelation. An appropriately designed permutation procedure that maintains levels of global autocorrelation in the permuted data sets similar to those observed in the actual data would, at least in theory, enable identification of those local patterns that are unusual even in the context of the observed global patterns. We are not aware of any reported results of this kind, and it is clear that such analysis would present substantial challenges in both execution and interpretation. Given such challenges, it seems likely that local statistics of the kind discussed in this section will continue to be used primarily in an exploratory and descriptive mode for the foreseeable future. Many examples of the use of these measures can be found in the research literature in a wide range of subject areas, in spite of the lack of a completely satisfactory inferential framework. As evidence of this, consider the fact that at the time of writing, the three papers by Anselin (1995) and Getis and Ord (1992, 1995) have been cited a combined total of almost 1000 times!

8.5. OTHER LOCAL STATISTICS

Geographically Weighted Regression

Another popular local statistic developed in the last decade or so is geographically weighted regression. In a simple multivariate regression model, we model the relationship between one *dependent variable* and one or more *independent variables*. The mathematical model underlying multiple regression is

$$\begin{aligned} y_i &= b_0 + b_1 x_{i1} + b_2 x_{i2} \cdots + b_m x_{im} \cdots + \varepsilon_i \\ &= b_0 + \sum_{j=1}^{m} b_j x_{ij} + \varepsilon_i \end{aligned} \tag{8.8}$$

so that the value of the independent variable at each location y_i is modeled as the sum of a constant b_0, a sum of products of each independent variable value x_{ij}, and a coefficient b_j, along with an error term ε_i. The model is fitted to the observed data using a least squares regression procedure, which ensures that the sum of the squared errors at all locations in the data set is minimized. The mathematics underlying regression is comparatively simple, but beyond the scope of this book, and is covered in numerous introductions to statistics. For the present purpose, it is convenient to express the full regression model in matrix terms as follows:

$$\begin{aligned} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & & \ddots & \\ 1 & x_{n1} & & x_{nm} \end{bmatrix} \begin{bmatrix} b_0 \\ \vdots \\ b_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \\ \varepsilon_n \end{bmatrix} \\ \mathbf{y} &= \mathbf{Xb} + \mathbf{e} \end{aligned} \tag{8.9}$$

so that $\mathbf{b}$ is a vector containing the estimated regression model coefficients, and $\mathbf{y}$ and $\mathbf{X}$ contain the observed data for the dependent and independent variables, respectively. It turns out that the least squares estimates of the regression coefficients can be calculated for this model from

$$\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{8.10}$$

Ordinary least squares regression is frequently applied to data that are spatially distributed. This involves creating a *global* regression model so that the relationship between the variables is assumed to apply with the same coefficients at all locations. An important step in the construction and evaluation of any regression model is to examine closely the model *residuals*, or errors, for evidence of any trends relative to any of the variables included in the model. For a model where the data are geographically distributed, a natural next step is to map the residuals. When any trend is discernible in the residuals, a regression model is said to be *misspecified*. This can be interpreted in a number of ways, but it generally requires that the analyst consider including additional variables in the model, removing variables from the model, or otherwise adjusting the model to address the problem. A subtle point is that such misspecification doesn't mean that the model is of no use. For example, the model is still the least squares best fit to the data. However, it does mean that the regression diagnostic statistics used to evaluate the statistical model are unreliable.

In a geographic setting, when we observe spatial structure in model residuals (which is almost always the case), this implies that either (1) spatial dependence of the variables should be included in the model or (2) it may be reasonable to allow the model to vary spatially. Both approaches have been developed to a considerable degree in the last two decades or so. The first option is adopted in various forms of *spatial regression*, which include spatially lagged versions of each model variable as additional variables in the model and provide an array of new diagnostic statistics for assessing the quality of the model. Many of these methods are discussed in detail by Luc Anselin in *Spatial Econometrics* (Anselin 1988) and in subsequent collections by Anselin et al. (1995, 2004). We do not cover spatial regression methods of this type in this book, as their interpretation is a rather advanced topic.

Although such spatial regression approaches explicitly include spatial dependence in the model, they generally consider spatial dependence itself to be uniform across the whole study area. Thus, the same estimates for the spatial dependence in the various variables included in the model are based on global estimates of the spatial dependence. In presenting *geographically weighted regression* (GWR), Fotheringham et al. (2002) suggest that this means that those types of spatial regression are "semilocal" rather than truly local statistics. GWR, by contrast adopts the second possible approach above and allows the regression coefficients in the model to vary from place to place. Thus, in the same way that the variables themselves change from place to place, we assume that the relationships among them may vary from place to place. This idea is made explicit in the subtitle to the definitive introduction to GWR in Fotheringham et al. (2002): *The Analysis of Spatially Varying Relationships*. That book provides a comprehensive overview of the basic features of GWR and is recommended, particularly the second chapter, which sets out the idea in a very clear, direct way. We borrow heavily from that chapter in the description of the approach below.

Now that we have noted the spatial autocorrelation in the residuals of our regression model, the idea in GWR is to build many local models instead, as a way of better understanding the spatial structure in the model. At its simplest, this concept involves simply partitioning the data set into a number of regions and estimating a local regression model for each region individually. Thus, instead of modeling (say) school truancy rates based on socioeconomic variables across the whole of a large urban region, we might develop a set of models, one for each of the school districts in the region. It is worth noting here that a related approach is *multilevel modeling*. In this framework, a series of nested models are developed in which the variance in the dependent variable accounted for by a set of independent variables at one level is removed and the remaining variance is then modeled using another set of independent variables. Such a model may include several scales. A number of geographers have explored the application of this approach where levels are defined by a series of geographic scales from the whole study area down to a highly localized level (see Jones, 1991). Since multilevel modeling was originally developed in the context of understanding educational outcomes based on school district, school, and classroom-level variables, this is a natural approach, which may make a lot of sense in many applications.

Returning to GWR, a further development of the idea of a collection of local models is to construct a “moving window” collection of models, where at any chosen location a local subset of the data is used to estimate a regression model. The obvious next step in this progression, adopted in GWR, is to construct a local model at every location in the study area such that observed data are included in each local model and *spatially weighted*, depending on their proximity to the location. In this approach, as with other local statistics, nearby data points are weighted more heavily than those from more remote locations using a kernel function in exactly the same way as KDE (see Section 3.6). This development relies on using *weighted linear regression* for each local model rather than ordinary least squares regression. In weighted regression, a weight is associated with each observation in the data set, so that the regression coefficient estimates are now given by

$$\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{W}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{8.11}$$

where $\mathbf{W}$ is a diagonal matrix of weights for each case in the data set. Off-diagonal elements in $\mathbf{W}$ are zero, and the diagonal elements have values $\sqrt{w_{ii}}$, where w_{ii} is the weight we wish to associate with each observation. In GWR, the elements in $\mathbf{W}$ are based on the spatial association between the location at which the local regression is performed and the available points at which data are available, so that we have a local version of weighted regression.

Clearly, how the regression weights $\mathbf{W}_i$ are determined for each local model is critical. The approach described by Fotheringham et al. (2002) involves using either a Gaussian or biweight kernel function at each location to assign weights to nearby observations in the data set. As with kernel functions in other applications, the critical aspect is not the mathematical form of the kernel but its bandwidth. In applications that support GWR, the bandwidth can be a value chosen by the user, which is fixed for all locations, or it can be an *adaptive* variable bandwidth, which is different at every location. The latter approach accommodates data sets that include significant variation in the intensity of the data points, although it involves a considerable amount of computation. A complex method for automatically choosing each local bandwidth has been implemented based on running multiple models at each location, omitting one point at a time and setting the bandwidth such that it produces the best estimates of the omitted location.

The end result of all this computation is a set of estimated regression coefficients that vary across the study area. These estimates can be mapped so that varying relationships between variables across space can be investigated. Diagnostic statistics associated with the method address the question of whether or not any estimated spatial variation in coefficients is simply a random sampling effect or is actually indicative of spatially varying relationships in the data. The approach is similar to the drift analysis of regression parameters developed by econometricians (Casetti and Can, 1999), as well as the ideas of kernel and nearest-neighbor regression developed by statisticians (Cleveland, 1979; Cleveland and Devlin, 1988).

In the case presented by Fotheringham et al. (2002, pp. 27–64), interpretation of GWR results is straightforward. Their example is house prices in London, modeled as dependent on a range of property characteristics, such as floor space, number of bedrooms, availability of a garage, and so on. In this context, geographic variation in the regression model parameters is readily interpreted as variation in the market value of the various property features included in the model. Thus, for example, the market valuation placed on a garage may be different in different parts of London. Unfortunately, variation in model parameters in other contexts can be much harder to interpret. This is particularly so when the variation in a regression coefficient is extreme enough to shift from positive to negative values across a study area. In such cases, it may be reasonable to assume that other variables in the model are confounding the results or that important variables are missing. Less extreme but still significant variation in the association between variables is typically easier to interpret (see the example discussed below). In all cases, it is vital to map the results, as with other local methods.

Invariably, GWR models produce better fits to observed data than global regression models. This is not surprising, given that a GWR model is not a single model, but a (potentially very large) number of local models. Using the approach to inference proposed by Fotheringham et al. (2002), even allowing for the additional degrees of freedom, GWR models are typically better than the associated global model based on Akaike's Information Criterion (AIC).

A Simple Example of GWR in Action

GWR can be illustrated by summarizing a paper by Brunsdon et al. (2001) that examines spatial variations in the relationship between average rainfall and altitude across Great Britain. In the past, workers typically used standard linear regressions to show that as one ascends, the average annual rainfall increases, a phenomenon called *orographic enhancement*, which results from a combination of meteorological processes. The simple linear model of this relationship, usually fitted by ordinary least squares (OLS) methods, is

$$P = b_0 + b_1 H + \varepsilon \tag{8.12}$$

where

P = rainfall (mm)
b_0 = rainfall at sea level (mm)
b_1 = rate of increase in rainfall with altitude, or height coefficient (mm/m)
H = station altitude (m above sea level)
ε = an error term

Many years ago, for some 6500 stations across Britain, Bleasdale and Chan (1972) found an overall relation in which the estimated average annual rainfall was given by

$$\hat{P} = 714 + 2.42(H)\,\mathrm{mm} \tag{8.13}$$

This implies a *constant* sea level rainfall of 714 mm across the entire country and a *constant* rate of increase in average annual rainfall with height of 2.42 mm/m. There are at least three reasons that led Brunsdon et al. to suggest that this relationship cannot be the same across the entire country:

- In the United Kingdom, the orographic effect is most pronounced at warm fronts and in the warm sector of depressions but is not important in cold frontal rain. It is known that there is spatial variation in the mixture of rain-producing events across the country that would be expected to produce nonstationarity in the relationship.
- Analysis of the results of this model shows that it systematically overpredicts rainfall in the east and underpredicts it in the west, giving a strongly spatially autocorrelated pattern of residuals.
- Armed with data from over 10,000 rain gauge sites, a preliminary visualization exercise in which spatial subsets of the data were isolated and the rain/height relation examined showed enormous variation in the estimates for both b_0 and b_1. Typically, subsets of the data from the south and east showed lower values for both of these coefficients than subsets from the west and north.

The results of the GWR analysis for these data are shown in Figure 8.2, which presents contour maps for the spatially varying b_0 and b_1 estimates. To produce these estimates, Brunsdon et al. had a problem in choosing both the form and the bandwidth of the kernel to be used. While a narrow bandwidth captures many rain gauges in the lowlands, it risks finding too few in the highlands and along the coast. Too wide a bandwidth risks smoothing out important variations in the relationship. Based on a complex cross-validation exercise, a two-dimensional Gaussian function with a 2-km bandwidth (i.e., standard deviation) was chosen. This choice implies that all the gauges in an effective neighbourhood of area 113 km^2 around each estimation point were used, with a weighting that drops sharply with distance until, at 6 km, it is effectively zero.

The results are summarized in two maps shown in Figure 8.2, one for the estimated rate of increase in average annual rainfall in millimeters per meter of ascent, the other for the intercept term, equivalent to the estimated rainfall in millimeters at sea level.

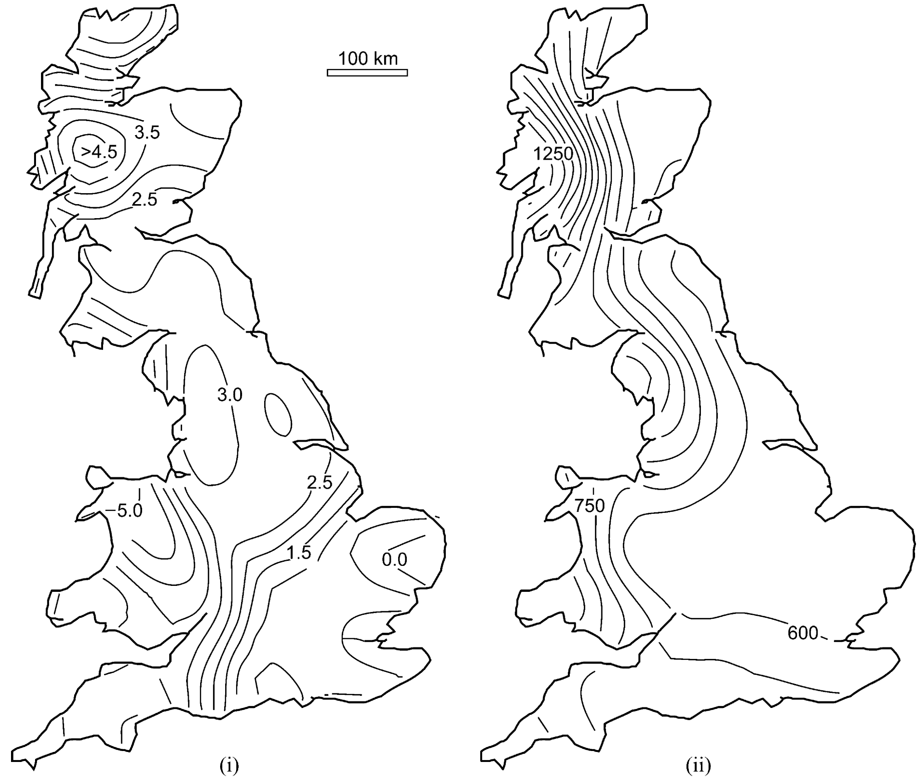


Figure 8.2 Estimation results of GWR for (i) the height coefficient b_1 in mm/m contoured at intervals of 0.5 mm/m and (ii) the intercept constant b_0 for Great Britain contoured at 50-mm intervals from 600 to 1250 mm.

(*Source*: Brunsdon et al., 2001)

Figure 8.2(i) shows the results for the height coefficient. The expectation of spatial nonstationarity in the model is confirmed, with a variation from a value of 0.0 in the east (implying that there is *no* increase with height), rising rapidly in a band running from south to northeast across the country to values in excess of 5.0 mm/m over the mountains of northwest Scotland. Figure 8.2(ii) shows the results for the intercept constant, b_0, which varies from less than 600 mm over much of the east of the country to more than 1200 mm in the far northwest. In conclusion, and perhaps a little tongue in cheek, Brunsdon et al. concluded that perhaps the real relationship between average annual rainfall and altitude across the United Kingdom should be rewritten as follows:

$$\hat{P} = (<600 \text{ to } >1250) + (0.0 \text{ to } >4.5)H \text{ mm} \tag{8.14}$$

Criticism of GWR

Critics of GWR raise valid concerns about a number of aspects. Some focus on how inferences are made about whether or not the observed variations in regression coefficients are statistically significant, and much debate revolves around how to assess the dependence of the model on the automatically determined kernel bandwidths: how are degrees of freedom involved in determining the bandwidths to be accounted for? Such difficulties are similar to those encountered in inference about other local statistics. In the same vein, there may be concerns in GWR about how many observations are included in each local model, and also with the characteristics of those observations in terms of their suitability as input variables to a regression model. In brief, stable regression coefficients depend on the independent variables in a model being uncorrelated with one another (if they are not, then *multicollinearity* problems arise), and also on their having well-behaved distributional characteristics without too many extreme values. For the best results, the independent variables should be approximately normally distributed. Given that the local models are each based on subsets of the total data set, in GWR it is quite likely that some of the local models will suffer from one or both of these problems. Since a symptom of either problem is unreliable estimation of the regression coefficients, a side effect of such issues could be a tendency for GWR to overestimate how much the regression coefficients vary in space.

These are clearly important issues to consider when using GWR. However, many of these concerns are primarily about the inferences that can be made in GWR. These become less important if the method is treated as essentially an exploratory one. That this is the preferred approach to GWR is made clear in the title of the one of the first papers discussing the method, “A Method for Exploring Spatial Nonstationarity” (Brunsdon et al., 1996).

Density Estimation

Density estimation has already been discussed in Section 3.6. Here we suggest that you consider the simple idea that a density estimation is essentially a local statistic. For the set of point events or point-located count data, density estimation produces a locally weighted count in order to estimate the intensity of the point process at every location. In this case, the kernel function fulfills the role of defining each locality by determining the weight to be associated with each event of a point-located count based on its distance from the location at which an estimate of event intensity is required. Structurally, this is very similar to the G_i statistic, although here we make the density estimates at locations that are not part of the point data set.

Interpolation

Spatial interpolation methods may also be regarded as local statistics. These are considered in some detail in the next two chapters. However, it is worth noting in advance that, in essence, all spatial interpolation methods estimate values at unsampled locations in a field based on some locally weighted sum of the values at sampled locations. In the simplest case, all the sampled locations are equally weighted in the sum, and the interpolation is simply a local mean. In more widely used approaches, this procedure is modified slightly by introducing a distance weighting component so that nearer observations matter more than distant ones. The similarity to density estimation is striking, although the intention and the outcomes are quite different. The intention is different in that the surface being estimated at each location is not an intensity surface, but rather the surface of (unmeasured) attribute values. The outcome is different largely because of a seemingly minor but important difference in the two methods. In density estimation, the weights associated with each event or point-located count are unaffected by the number of points included. Where there are many events in a local area, the resulting estimates will increase with each additional event. In interpolation, since the underlying basis of the weighting is an averaging procedure, additional observations included in the local statistic result in the weight associated with each observation being reduced. In fact, in spatial interpolation, the sum of the weights used for each local estimate is 1. Here, additional observations in the locality refine the calculation of the local mean but do not necessarily increase it. Instead, they may increase or decrease the outcome, depending on whether they are low- or high-value observations.

8.6. CONCLUSIONS: SEEING THE WORLD LOCALLY

This chapter is in some ways a review of many of the methods considered in previous chapters. We hope that by now you can see just how central to spatial analysis the concepts of adjacency, distance, interaction, and neighborhood introduced in Section 2.3 really are. All the local statistics discussed here make use of these concepts as a preliminary step in the development of the analysis. At the same time, this chapter anticipates developments in the next two chapters on spatial interpolation methods, because there too, concepts of locality and the associated spatial weighting of observations are of central importance.

In spite of their importance, local statistics still present considerable challenges for the spatial analyst. Foremost among these is the difficulty of drawing inferences based on them. We have discussed the reasons for this difficulty, with the problems of small, nonrandom sampling and multiple testing being the root causes. The current preferred solution, based on computer simulation, is well established and provides one reason for the relatively slow diffusion of local statistics: their seemingly paradoxical dependence on substantial computational resources. This and the importance of mapping local statistics are major reasons why it is only the advent of GIS that has seen local statistics come into their own.

The challenges associated with drawing inferences based on local statistics have led to a strong tendency to treat them as primarily exploratory approaches. That is considered to be a major failing of the approach by some, who prefer more formal statistical approaches over the exploration of data. We believe that seeing the world locally using these methods is a powerful approach that has a place in the toolkit of all would-be spatial analysts.

CHAPTER REVIEW

- *Local statistics* are an important new approach to spatial analysis that has gained in popularity since the 1990s.
- Among the reasons for the slow rise to prominence of local statistics are their dependence on the easy design and creation of maps with which to visualize them and the need for considerable computational resources for assessing the statistical significance of local statistics.
- Less technically, local statistics have also become more popular as the importance of spatial variation in phenomena has become more widely recognized as a result of the diffusion of GIS and other geospatial technologies.
- The earliest local statistics specifically developed for the exploration of data sets are the *Getis-Ord* G_i *and* G_i^* *statistics*, which allow exploration of the degree to which high or low values in a data set are spatially clustered.
- A local version of Moran's I statistic is readily derived from the development of the global statistic.
- Both the G *and* I statistics are required to obtain a thorough understanding of the spatial dependence structure in a data set.
- Inference about the local G and I statistics (and other local statistics) is made difficult by the twin problems of multiple testing and small sample sizes. The best solutions to these difficulties lie in computer simulation of multiple permutations of the original data and derivation of *pseudosignificance tests*.
- Several local forms of regression are available under the general heading of *spatial econometrics*.

- *Geographically weighted regression* (GWR) is a local form of weighted linear regression that allows the standard regression coefficients to vary from place to place and provides approaches to inference concerning the variability in the resulting surfaces of local coefficients.
- Many standard spatial analysis techniques can be usefully reinterpreted as local statistics, an approach that emphasizes the importance of the key concepts of adjacency, interaction, and neighborhood (or locality).

REFERENCES

Anselin, L. (1988) *Spatial Econometrics: Methods and Models* (Dordrecht, Netherlands: Kluwer Academic).

Anselin, L. (1995) Local indicators of spatial association—LISA. *Geographical Analysis*, 27(2): 93–115.

Anselin, L. and Florax, R. J. G. M., Eds. (1995) *New Directions in Spatial Econometrics* (Berlin and New York: Springer-Verlag).

Anselin, L., Florax, R. J. G. M., and Rey, S. J., Eds. (2004) *Advances in Spatial Econometrics: Methodology, Tools and Applications* (Berlin and New York: Springer-Verlag).

Bleasdale A. and Chan, Y. K. (1972) Orographic influences on the distribution of precipitation. In: *Distribution of Precipitation in Mountainous Areas* (Geneva: World Meteorological Office), pp. 322–333.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4): 281–298.

Brunsdon, C., McClatchey, J., and Unwin, D. J. (2001) Spatial variations in the average rainfall/altitude relationship in Great Britain: an approach using geographically weighted regression. *International Journal of Climatology*, 21: 455–466.

Casetti, E. and Can, A. (1999) The econometric estimation and testing of DARP models. *Journal of Geographical Systems*, 1: 91–106.

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74: 829–836.

Cleveland, W. S. and Devlin, S. J. (1988) Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83: 596–610.

Fotheringham, A. S. (1997) Trends in quantitative methods I: stressing the local. *Progress in Human Geography*, 21(1): 88–96.

Fotheringham, A. S., Charlton, M. E., and Brunsdon, C. (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* (Chichester, England: Wiley).

Getis, A. and Ord, J. K. (1992) The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3): 189–206.

Jones, K. (1991) Multi-level models for geographical research. *Concepts and Techniques in Modern Geography*, 54, (48 pages Norwich, England: Environmental Publications). Available at http://www.gmrg.org.uk/catmog.

Ord, J. K. and Getis, A. (1995) Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, 27(4): 286–306.

Sidak, Z. (1967) Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62: 626–633.

Tukey, J. W. (1977) *Exploratory Data Analysis* (Reading, MA: Addison-Wesley).

Unwin, D. J. (1996) GIS, spatial analysis and spatial statistics. *Progress in Human Geography*, 20(4): 540–551.

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 9

Describing and Analyzing Fields

CHAPTER OBJECTIVES

In this chapter, we attempt to:

- Show how important fields are in many practical problems
- Show how field data can be recorded and stored in a GIS
- Introduce the concept of interpolation as *spatial prediction* or *estimation* based on point samples
- Emphasize the importance of the first law of geography in interpolation
- Demonstrate how different conceptions of *near* and *distant* or *neighborhood* result in different interpolation methods that produce different results
- Explore some of the *surface analysis* methods that can be applied to fields

After reading this chapter, you should be able to:

- Outline what is meant by the term *scalar field* and differentiate *scalar fields* from *vector fields*
- Devise an appropriate model for such data and understand how the choice of model will constrain any subsequent analysis
- *Interpolate point data* by hand to produce a field representation
- Describe how a computer can be programmed to produce repeatable contour lines across fields using *proximity polygons*, *spatial averages*, or *inverse distance weighting*
- Explain why these methods are to some extent arbitrary and should be treated carefully in any work with a GIS

- Understand the ideas of the *slope* and *aspect* as a vector field, given by the first derivative of height
- List and describe some typical processing operations using height data

9.1. INTRODUCTION: SCALAR AND VECTOR FIELDS REVISITED

Revision

In order to fit what follows into our framework, you should review earlier material as follows:

- Section 1.2 on spatial data types, noting what is meant by the field view of the world
- Section 2.3, especially Figures 2.3, 2.4, and 2.5 and the related text on proximity polygons
- Section 3.8 on visualizing fields

In Chapter 1 we drew attention to a basic distinction between an *object* view of the world that recognizes point, line, and area objects, each with a bundle of properties (attributes), and a *field* view where the world consists of attributes that are continuously variable and measurable across space. The elevation of the Earth's surface is the clearest and easiest-to-understand example of a field, since, almost self-evidently, it forms a continuous surface that exists everywhere. In a more formal sense, it is an example of a *scalar field* (as we shall see shortly, it turns out that this apparently obvious fact is arguable). A *scalar* is any quantity characterized only by its *magnitude* or amount *independent of any coordinate system in which it is measured*. Another example of a scalar is air temperature. A single number gives its magnitude, and this remains the same no matter how we transform its spatial position using different map projections. A *scalar field is a plot of the value of such a scalar as a function of its spatial position*.

Scalar fields can be represented mathematically by the very general equation

$$z_i = f(\mathbf{s}_i) = f(x_i, y_i) \tag{9.1}$$

where *f* denotes "some function." This equation simply says that the surface height varies with the location.

Notation and Terminology

By convention, z is used to denote the value of a field. When we think of a field as a surface, z is equivalent to the surface height above some datum level. In our notation, $\mathbf{s}_\mathrm{i}$ denotes a spatial location whose coordinates are (x_i, y_i). By *height*, we mean the scalar value of the variable that makes up the field. This could be a quantity such as temperature, rainfall, or even population density. We often use *height* as a general term for any field variable we are interested in.

Just by writing down this equation, we have already made some important assumptions about the scalar field. Depending on the phenomenon represented and on the scale, these assumptions may not be sustainable for real fields. First, we assume *continuity*: for every location, $\mathbf{s}_i$, there is a measurable z_i at that same place. Although this seems intuitively obvious, it is an assumption that is not always satisfied. Strictly, mathematicians insist that continuity also exists in all the *derivatives* of z, that is, in the rates of change of the field value with distance. For a field representing the height of the Earth's surface, this implies that there are no vertical cliffs. Try telling that to anyone who has gone rock climbing in, say, Yosemite or, less dramatically, the Derbyshire Peak District! Second, we assume that the surface is *single-valued*. For each location, there is only *one* value of z. This is equivalent to assuming that there are no caves or *overhangs* of the sort that (a few!) rock climbers greatly enjoy.

An Example of a Field and Its Usefulness

The best example of field data in geography is the height of the Earth's surface, usually expressed in meters above sea level. Such data might be used:

- To produce maps and other visualizations of the relief of the Earth's surface for navigation and general interest
- In hydrology to detect features of interest in the landscape such as river watersheds and drainage networks
- In ecology for the computation of attributes of the land surface that have ecological significance, such as its slope and aspect
- In studies of radio and radar propagation by the mobile telephone industry or military

(*continues*)

(*box continued*)

- In photorealistic simulations both in arcade games and in serious applications such as flight simulation
- In terrain-guided navigation systems to steer aircraft or, more notoriously, Tomahawk cruise missiles
- In landscape architecture to map areas visible from a point (or viewsheds)
- In geoscience, as an input into systems to predict the fluxes of energy (for example, sunlight) onto and across (for example, water) the landscape

The land elevation field has a concrete existence—you stand on it almost all the time. Other scalar fields are less concrete in that we often can't see, touch, or perhaps even feel them, but they are still measurable in a repeatable way.

As an exercise, how many scalar fields of interest in geography can you list? How might these fields be manipulated in studies using GIS?

In fact, scalar fields are found and analyzed in virtually all the sciences. This has its benefits in that there are well-developed theories, methods, and algorithms for handling them. There is also a downside in that the same concepts are often reinvented and given different names in different disciplines, which can lead to confusion.

A Note on Vector Fields

By contrast, with scalar fields, vector fields are those where the mapped quantities have both magnitude and direction that are not independent of the locational coordinates used. Scalar fields all have an equivalent vector field and vice versa. For example, the scalar field of land elevation may be used to generate a vector field giving the maximum surface slope (a magnitude) and its aspect (a direction). This is a common operation in GIS. The results obtained may change when we use different map projections. Any vector field may be thought of as the slope and aspect field of a corresponding scalar field that ''generated'' it.

For the examples of scalar fields you suggested in the previous exercise, what is the meaning of the equivalent vector fields? If we have a scalar field of temperature, what does the vector field represent? How can the atmospheric pressure field be used to predict the vector field of winds?

Your local science library or bookshop will have many books in the mathematics section on scalar and vector fields. A useful one is McQuistan's *Scalar and Vector Fields: A Physical Interpretation* (1965). A good overview of vector fields in geographic information science can be found in Li and Hodgson (2004).

9.2. MODELING AND STORING FIELD DATA

As for other spatial object types, how a field is recorded and stored in a GIS can strongly affect the analysis that is possible. For fields, there are two steps in the recording and storage process: *sampling* the real surface and employing some form of *interpolation* to give a *continuous surface representation*. In this section, we briefly consider surface sampling but concentrate on five approaches to continuous surface representation: *digitized contours*, *mathematical functions*, *point systems*, *triangulated irregular networks* (TINs), and *digital elevation matrices* (DEMs). Each of these generates a digital description of the surface that is often called a *digital elevation model*. Details of simple interpolation techniques that may be used to produce such continuous representations from sample data are presented in Section 9.3. The end product of sampling and interpolation is a field that may be visualized and analyzed (or *processed*) to attach meaning to the underlying data. A selection of processing operations commonly used in GIS is discussed in the concluding sections of this chapter.

Step 1: Sampling the Real Surface

Whatever field description and processing methods are used, it is necessary to acquire suitable sample data. These often strongly affect how a field is modeled and stored. There are numerous possibilities. Sometimes we have a series of measured values of the field obtained by some method of direct survey. For example, we might have the rainfall recorded at sites where rain gauges are maintained or values of air temperature recorded at weather stations. For Earth surface elevation, recorded values are called *spot heights*. A more general term is *control point*. In terms of the general equation $z_i = f(\mathbf{s}_i)$, control points are a list of z values for selected locations in some pattern scattered over the region. Increasingly, field data are acquired from aerial and satellite remote sensing platforms, including very high spatial resolution Light Detection and Ranging (LIDAR) scanning usually providing z values over a regular grid of locations. In terms of the basic field equation, this is a solution in the form of a regular table of numbers. Field data may also be acquired by digitizing the contours on a map. Many mapping agencies produce grids of height data that appear to have been directly measured but have actually been produced from a digital version of the preexisting topographic maps. In terms of the equation

$$z_i = f(\mathbf{s}_i) = f(x_i, y_i) \tag{9.2}$$

digitized contours are previous solutions in the form of all the (x, y) values with various fixed z values—the contour heights. Since the contours may have been determined from a set of spot heights in the first place, such data should be treated with caution. Further processing can produce data in which the contour values are overrepresented.

Whatever the source, there are three important points to consider in relation to field data:

- The data constitute a *sample* of the underlying continuous field. Even if we wanted to, it is impractical to measure and record values everywhere across the surface.
- With the exception of a few relatively permanent fields such as the height of the Earth's surface, these data are all that we can ever have. Many fields or surfaces are constantly changing (think of most weather patterns), and only the values recorded at particular locations at particular times are available.
- Unless we go out into the real world and do the measurements ourselves, much of the time we have no control over where the sample data have been collected.

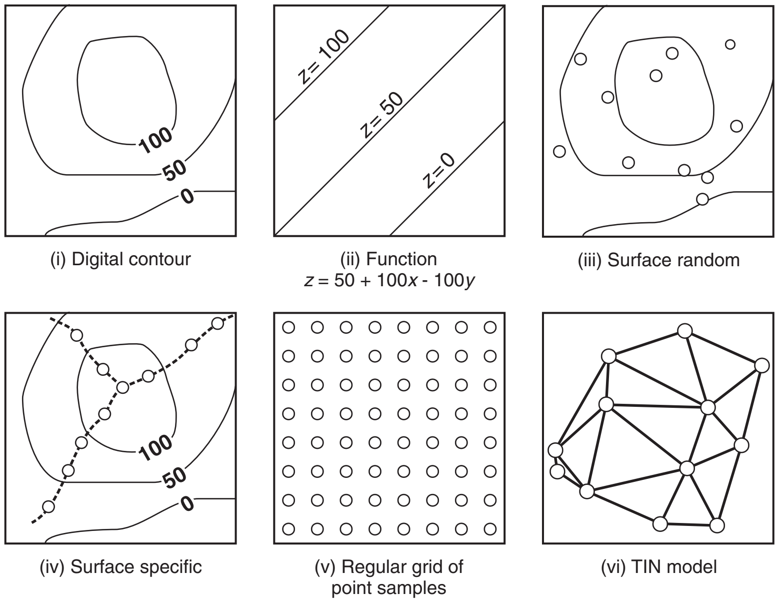
An indirect corollary of the last point, which has always been true in basic field surveying, is that the best control points are those where you have had an opportunity to influence the sampling design before collecting the height values to enhance the opportunities for further processing.

Step 2: Continuous Surface Description

As we have just seen, to represent faithfully any scalar field requires an effectively infinite number of points. What determines the number of points actually used is rarely the surface itself but instead our ability to record and store it. Even in the nearly-ideal situation of Earth surface elevation, which, in principle at least, could be measured everywhere, practical considerations dictate that any information system is unable to store all the data (indeed, until very recently, high-resolution measurements such as LIDAR put a considerable strain on computer resources anywhere they were collected). Generally, a surface of interest has been recorded at a limited number of control points and must be reconstructed to produce what we hope is a satisfactory representation of the truth. Often, we cannot be certain that the reconstructed surface is reasonable—for example, the air temperature on a particular day is no longer available, except in the form of the values recorded at weather stations. Today, *nobody* can possibly know the actual air temperature at a location where no record was made yesterday.

Reconstruction of the underlying continuous field of data from the limited evidence of the control points is called *interpolation* and is an example of the classic *missing data problem* in statistics. Whatever type of surface is involved and whatever control points are used, the objective is to produce a field of values to some satisfactory level of accuracy relative to the intended subsequent use of the data. It is therefore important to consider the possibilities for storing and representing a field before interpolation is undertaken, since the representation adopted may affect both the choice of interpolation technique (see Section 9.3) and the possibilities for subsequent analysis (see Section 9.4).

Several methods, illustrated in Figure 9.1, can be used to record field data. Each of these is considered in the sections that follow.



(i) Digital contour

(ii) Function
$z = 50 + 100x - 100y$

(iii) Surface random

(iv) Surface specific

(v) Regular grid of point samples

(vi) TIN model

Figure 9.1 Methods of storing fields.

Continuous Surface Description (1): Digitized Contours

An obvious way of recording and storing altitude is to digitize and store contour lines from a suitable map, as in Figure 9.1(i). Such data are readily acquired by digitizing the contour pattern of a printed map. Just as these data are easily acquired, they are also easy to display by "playing back" the stored coordinates. Some surface processing operations, such as calculation of the areas above or below specified heights, are easy to do with contour data. In production cartography concerned with topographic maps, where most plotting is of point or line information, this is the preferred method, but for GIS analysis it has severe limitations. First, the attainable accuracy depends on the scale of the original map together with both the spatial and vertical accuracy of the source map contours. Second, all information on surface detail between contours is lost. Third, the method oversamples steep slopes with many contours relative to gentle ones with only a few. Finally, many processing operations, such as finding the slope or even something as apparently simple as the elevation of an arbitrary location, are remarkably difficult to automate for contour maps.

Continuous Surface Description (2): Mathematical Functions

In some GIS applications, it is possible to use a functional expression such as

$$z_i = f(\mathbf{s}_i) = f(x_i, y_i) = -12x_i^3 + 10x_i^2y_i - 14x_iy_i^2 + 25y_i^3 + 50 \tag{9.3}$$

This gives the height of the field using an explicit mathematical expression involving the spatial coordinates. In principle, a single, compact mathematical expression is a very good way of recording and storing surface information, since it allows the height at any location to be determined. Figure 9.1(ii) shows a simple example. A more complex case—the surface described by the above equation—is shown in Figure 9.2. Note that x and y are likely to be expressed in kilometers, whereas z would be expressed in meters.

In this approach, the problem is to find a mathematical function, or series of functions, that interpolates or approximates the surface. By *interpolate*, we mean that the expression gives the exact value for every known control point, so that the surface it defines honors all the known data. In contrast, a function that *approximates* the surface may not be an exact fit at the control points and does not honor all the data. Sometimes—for example, in producing contour maps of the Earth's surface elevation—interpolation is required since the observed height information is known exactly. At other times, when control point data are subject to significant error or uncertainty, or where scientific interest focuses on trends in the surface values, approximation is more appropriate.

A mathematical representation of a surface has many advantages. First, it is a compact way of storing all the information. Second, the height at any location (x_i, y_i) can be found by substitution into the formula. Third, finding contour lines is straightforward, involving the solution of an equation for all coordinate values with the required z-value. Fourth, some processing operations, such as calculation of surface slope and curvature, are easily performed using the calculus to differentiate the function. The disadvantages of the approach lie in the often arbitrary choice of function used and a tendency for many functions to give values that are very unlikely or even impossible—for example, negative rainfall totals—in the spaces between data control points. It is also frequently almost impossible to derive a *parsimonious* function that honors all the data. A parsimonious function is one that does not use a large number of terms in x and y and their various powers.

In most GIS textbooks you will find little mention of this approach, but it is used more frequently than is realized. In some applications, complex functions are used to describe even relatively simple scalar fields. The best example is in operational meteorology, where atmospheric pressure patterns are often recorded this way. The method is also used in one statistical approach to analyzing spatially continuous data called *trend surface analysis* (see Section 10.2). Mathematical functions are also used in the method of *locally valid analytical surfaces*. From Figure 9.2, it is evident that not far beyond this small region, the equation will give extreme

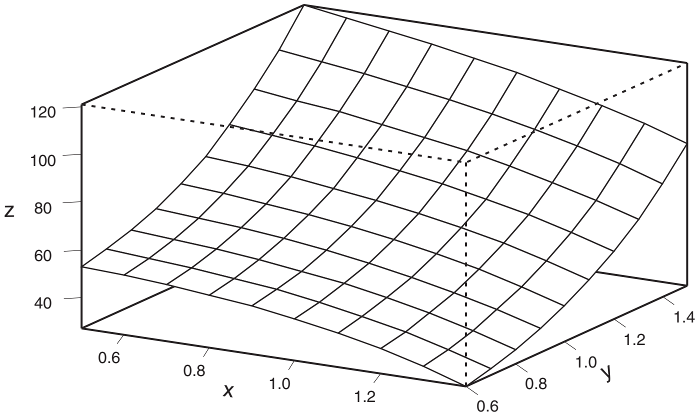


Figure 9.2 Surface from the equation presented in the text over a small range of (*x*, *y*) values. Note that the *z*-axis is exaggerated fivefold in this diagram.

values. For example, at $x = 0, y = 2$, the equation gives $z = 450$, making for extremely rugged terrain! This is typical of the difficulty of finding a function that fits the real surface over a wide area. For locally valid surfaces, this problem does not arise. The area covered by a field is divided into small subregions, over each of which the field behaves regularly. Each subregion is then described by its own mathematical function. The resulting collection of functions accurately and economically represents the entire surface. Effectively, this is also what is done when a TIN or DEM description of a surface is contoured.

Continuous Surface Description (3): Point Systems

Representing a surface by contours or a mathematical function produces a compact data file, but both representations are, in a sense, dishonest. What is stored is already an interpretation of the surface, one step removed from the original control point data. A third method avoids this problem by coding and storing the surface as a set of known control point values Under this general heading, there are three possible ways of locating the control points that can be called *surface random*, *surface specific*, and *grid sampling*.

1. In a *surface random* design, the control point locations are chosen without reference to the shape of the surface being sampled. The result, shown in Figure 9.1(iii), is an irregular scatter of control points that may, or (more likely) may not, capture significant features of the surface relief.
2. In *surface specific* sampling, points are located at places judged to be important in defining the surface such as peaks, pits, passes, and saddle points and along streams, ridges, and other breaks of slope. This is shown schematically in Figure 9.1(iv), where points along ridge lines and at a surface peak have been recorded. The advantage of this method is that surface-specific points provide information about the structural properties of the surface. Spot heights on most topographic maps are surface-specific sampling systems because they are usually located at significant points on the land surface such as hilltops and valley floors.
3. In *grid sampling*, we record field heights across a regular grid of (x, y) coordinates. This often appears in a GIS as a raster data layer, and if the field of interest is the height of the Earth's surface, it is called a *digital elevation matrix* (*DEM*) (see Figure 9.1v). The

advantages of grids are obvious. First, they give a uniform density of point data that is easily processed and facilitates the integration of other data held on a similar basis in a raster data structure. Second, spatial coordinates need not be stored explicitly for every control point, since they are implicit in each *z* value's grid location. To locate the grid relative to the real world, all that need be stored are the coordinates of at least one point, and the grid spacing and orientation. A third advantage is less obvious but very important. In a grid, we know not only each *z* value's position implicitly, we also know its spatial relationship to all other points in the data. This makes it easy to calculate and map other surface properties, such as gradient and aspect. Fourth, grid data can be readily processed using array data structures, available in most computer programming languages.

The disadvantages of grid data sets are the work involved in assembling them, the large arrays required, and the difficulty of choosing a single grid resolution appropriate across all of a large region. Because the number of points needed increases with the square of the linear resolution, changes in either the area covered or the resolution involved may be achieved only at great cost in extra data. For example, a standard 5 by 5 km "Profile" tile from the Ordnance Survey of Great Britain, with 10-m grid resolution, requires 250,000 values to be recorded, but halving the horizontal resolution to 5 m requires four times as many grid points (1,000,000). A tendency to oversample in areas of simple relief (such as a flat, dry salt lake bed) is also problematic. On the other hand, a large grid interval that avoids this problem might seriously undersample the surface in areas of high relief. In practice, the grid interval must be a compromise dependent on the objectives of the study. Cartographers encounter similar problems in choosing a single contour interval appropriate for large map series.

Continuous Surface Description (4): Triangulated Irregular Networks (TINs)

A common alternative to the DEM is a *triangulated irregular network* (TIN), illustrated in Figure 9.1(vi). TINs were originally developed in the 1970s as a way of contouring surface data, but they have subsequently been used to represent continuous surfaces based on a point sample. In a TIN, sample points are connected to form triangles, and the relief inside each triangle is represented as a plane or *facet*. In a vector GIS, TINs can be stored as polygons, each with three sides and with attributes of slope, aspect, and the heights of the three vertices. The TIN approach is attractive because of its simplicity and economy, since a TIN of 100 points will usually describe a surface as well as a DEM of several hundred, perhaps even several thousand, elements.

In creating a TIN, for best results it is important that samples are obtained for significant points such as peaks, pits, and passes, as well as along ridge and valley lines. Many GISs have facilities to do this, taking as their input a very dense DEM from which so-called *very important points* (VIPs) are automatically selected and used to build a TIN representation. The selected set of points may be triangulated in various ways, but typically the Delaunay triangulation is used. This uses proximity polygons as its basis, as described in Section 2.3.

9.3. SPATIAL INTERPOLATION

Spatial interpolation is the prediction of exact values of attributes at unsampled locations from measurements made at control points in the same area. In GIS, interpolation may be used to convert a sample of observations at control points into an alternative representation, typically either a contour map or a digital elevation model. Since we usually have no way of confirming the true values of the field away from the control points, interpolation is a type of spatial prediction. Figure 9.3 outlines the basic problem.

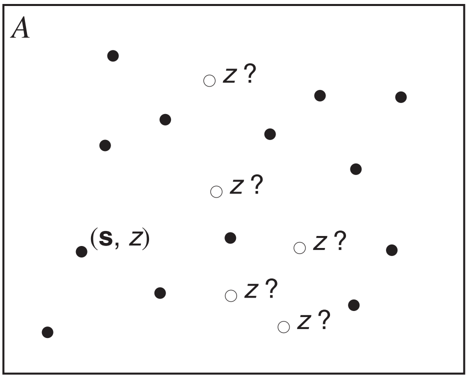


Figure 9.3 The interpolation problem. Control points are black circles, where we know the location, **s**, and the height, *z*, but we require the field height, *z*, anywhere in the region *A*—say, at the unfilled circle locations.

The First Law Again

Think for a moment about the first law of geography in relation to the possibility of spatial interpolation or prediction. Tobler's law tells us that ''everything is related to everything else, but near things are more related than distant things'' (Tobler, 1970), and, as we saw in Chapter 7, this appears in spatial analysis as spatial autocorrelation. Now consider a field of data that are not spatially autocorrelated, to which Tobler's law does not apply. Is spatial interpolation possible for such a field?

Hopefully, you answered a resounding ''no.'' The possibility of spatial interpolation depends on spatial autocorrelation being present. If it is not, then interpolation is not possible; we might just as well guess values based on the overall distribution of observed values, regardless of where they are relative to the locations we want to predict. An important concept to grasp here is that different interpolation methods are distinguished by the way that the concept *near* from the first law of geography is operationalized. This is always a key question in geography. We assume that space makes a difference; the question is, how? The interpolation methods described in the remainder of this chapter each answer this question in a different way. In choosing which method to use, you must consider how plausible the answers implied by each method are for the geographic problem at hand.

The best way to introduce interpolation is to attempt it by hand. The boxed exercise takes you through an example.

Spatial Interpolation by Hand and Eye

Figure 9.4 shows spot heights of the average January temperature (°F) in a part of Alberta, Canada. Your task is simple: Get a pencil (you will also need an eraser) and produce a continuous surface representation of these data by drawing contours of equal temperature (isotherms). While doing this, keep in mind three things:

1. Resist the temptation to join the dots. Remember that the data are unlikely to be exact and, even with a 0.1° resolution, each isotherm is likely to have substantial spatial width. In many applications, it is wildly optimistic to assume that the data are exact.

(*continues*)

(*box continued*)

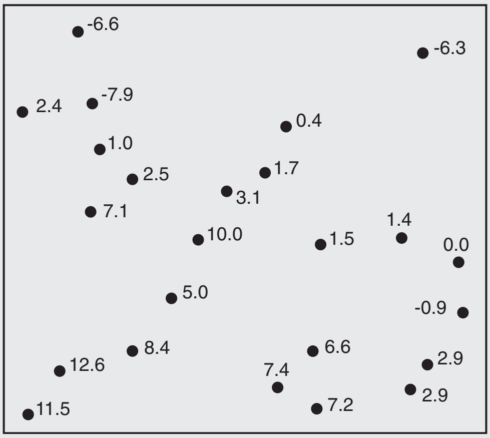


Figure 9.4 Average January temperature (°F) in a part of Alberta, Canada.

2. Experience suggests that it pays to start the process with a contour value in the middle of the data range and to work up and down from there.
3. Perhaps arguably, you should try to make the surface of average temperatures as smooth as you can, consistent with its honoring all the data. This means that there should be no inconsistencies where measured temperatures lie on the ''wrong'' side of relevant isotherms.

It's not as easy as it looks is it? Several points can be made based on this exercise. First, different people arrive at different solutions, making different predictions about the unknown values between the control points. Your solution may be a good one, but it is also only one of the many possible solutions. Without more information, there is no way be sure which solution is the best.

Second, because this is a map of average air temperature, we can be confident that assumptions of continuity and single value are reasonable, but what if the problem were to map the subsurface depth of a highly folded and faulted stratum where these assumptions did not hold?

Third, if you had access to additional information about the weather stations, such as their height above sea level, would this help? This illustrates the importance of prior knowledge in any interpolation. For example, if the surface were of elevation and not temperature, you would avoid valleys that rise and fall down their long profile, but you would be happy to create sharp V-turns in the contour lines, indicating drainage channels.

Finally, your confidence in the accuracy of each isotherm will not be consistent across the area. It will depend strongly on the number and distribution of control points, the chosen contour interval, and, more awkwardly, the unknown characteristics of the surface itself.

Given the difficulty of manual interpolation, it is natural to ask if it is possible to devise computer algorithms that interpolate in a consistent and repeatable way. In the remainder of this section, we introduce simple mathematical approaches to this problem. The next chapter describes more complex statistical methods of interpolation.

It is useful to think of the problem of predicting field values at unknown locations in the following way: If you had no information about the location of control points, what estimate would you make of the likely value of a new sample? Basic statistics tells us that the best estimate is simply the mean of the sample data points. In spatial terms, this would be equivalent to the situation represented in Figure 9.5. We make the assumption that all the unknown field heights have a single value equal to the mean, so that they form a single horizontal plane in the study area.

In Figure 9.5, higher values of the data set tend to be in the foreground and lower values in the background. Using a simple mean to predict values at unknown locations in this way, we are ignoring a clear spatial trend in the data. This results in a spatial pattern to our prediction errors, or *residuals*, with underprediction in the foreground and overprediction in the background. In other words, we know very well that the unknown values do not form a flat surface as shown. Instead, we expect them to exhibit some geographic structure, and making use of the spatial distribution of our

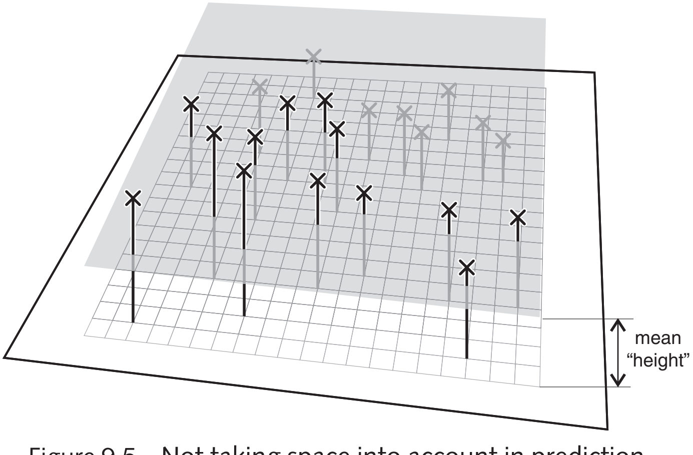


Figure 9.5 Not taking space into account in prediction.

samples and the first law of geography, we hope to be able to do better than this.

Automating Interpolation (1): Proximity Polygons

A simple improvement on using the mean is to assign to every unsampled point the value at its nearest control point using proximity polygons. In terms of the first law, of geography, this means that we are taking the idea of *near* to mean *nearest*. This operation is carried out by constructing proximity polygons for the control point locations and then assuming that each polygon has a uniform height value equal to the value at the control point. This technique was introduced almost a century ago by Thiessen (1911), who wanted to use rain gauge records to estimate the total rainfall across a region.

The proximity polygon approach has the virtue of simplicity, but as Figure 9.6 shows, it does not produce a continuous field of estimates. At the edges of each polygon, there are abrupt "jumps" to the values in adjacent polygons. In some situations, this may be the best we can do. Whether or not it is appropriate depends on the nature of the underlying phenomenon. If step changes are a reasonable assumption, then the approach will be fine. Also, remember that we may not have any way of knowing the accuracy of an interpolated surface, so this approach also has the virtue of making its assumptions immediately obvious to someone using the resulting spatial field. More smoothed fields may appear to have a spurious accuracy not justified by the observed data. Finally, if the data are not numerical but

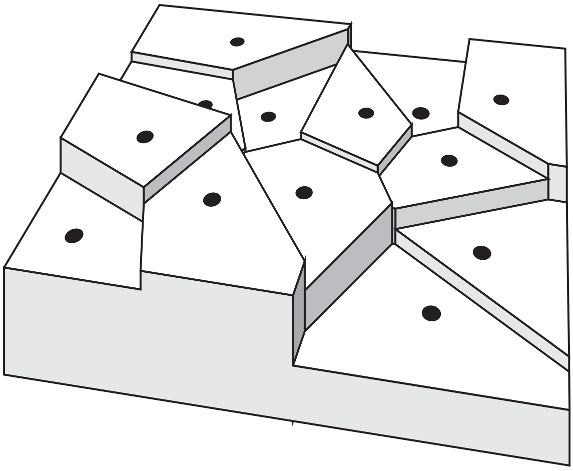


Figure 9.6 The ''blocky,'' discontinuous results of an interpolation using proximal polygons.

nominal—say, a soil, rock, or vegetation type—then the proximal polygon approach is often useful. However, bear in mind that processing nominal data in this way is not usually considered interpolation.

Automating Interpolation (2): The Local Spatial Average

Another way to approach interpolation is to calculate local spatial means of the sample data points. In effect, interpolation is a local statistic exactly like those discussed in Chapter 8. The idea here is that it is reasonable to assume that the first law of geography holds and to predict values at unsampled locations using the mean of values at nearby locations. The key question is, which locations are nearby? The proximity polygon approach has already suggested one answer: Use just the nearest location. Instead of using only the nearest control point, we can use only points within a fixed distance of the location where we wish to determine a value. In Figure 9.7, the effects and problems of this approach are highlighted.

The three maps show the results of using successive radii of 250, 500, and 750 m to determine local spatial means for the same set of spot heights. The obvious difficulty is that, because some locations are not within the chosen distance of *any* sample locations, it is not possible to estimate a full surface

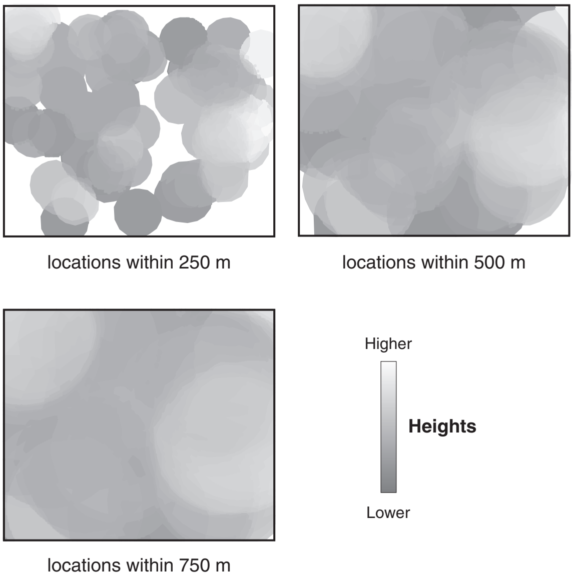


Figure 9.7 Interpolation using the mean of control points within 250, 500, and 750 m of locations to be estimated. Areas where no estimates are made are shown in white.

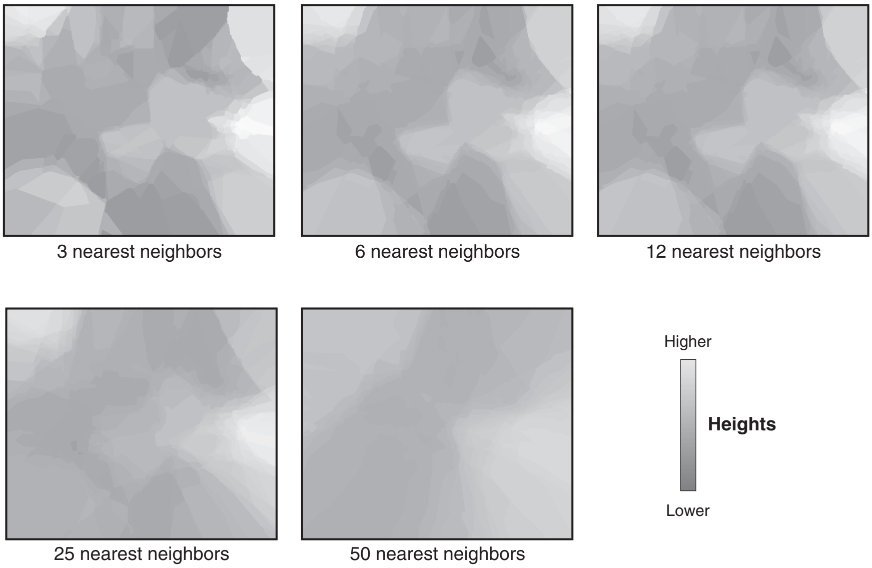


Figure 9.8 Nearest-neighbor interpolation for the data in Figure 9.7.

for the study region. A second problem is that, because points drop abruptly in and out of the calculation, the resulting field is not properly continuous. This is most obvious when the radius for including points in the local mean calculation is relatively small.

An alternative to a fixed radius is to use an arbitrary number of nearest-neighbor control points. For example, we might choose to use the six nearest control points to calculate the local mean at each unsampled location. Results for a series of different numbers of near neighbors are shown in Figure 9.8. This approach has the advantage that the effective radius for inclusion in the calculation of each local mean varies, depending on the local density of control points. In areas where control points are densely clustered, the radius is reduced; where control points are sparse, the radius is increased. An advantage of this method is apparent: All locations can have a value estimated, because all locations have three (or six, or however many) nearest neighbors. However, caution is required. For example, all the locations in the first panel of Figure 9.7, where there is no interpolated value, have no control point closer than 250 m away. This means that the interpolated surfaces in Figure 9.8 use control points *all* of which are farther away than 250 m to estimate values in those regions. This may not be very sensible in some cases. Another questionable aspect of both radius-limited and nearest-neighbor interpolation is that control points on only one side of the point to be estimated may be used if it happens that the nearest control points are all in more or less the same direction. In most software packages, it is possible to avoid this problem by requiring that some minimum number of control points in each direction away from interpolated locations be used.

Radius-limited and nearest neighbor interpolation share two characteristics. First, the chosen limit is arbitrary, whether it is a distance or a number of near neighbors. Second, as the sets of values on which estimates are based increase in size, discontinuous steps in the interpolated surface become smaller and a smoother appearance is produced. However, the appearance of smoothness is not real, because control points still drop in and out of the local mean calculations abruptly. One side-effect is that it is difficult to draw contours on the resulting interpolated surfaces. Furthermore, the larger the sets of control points we use for estimation, the more the interpolated surface becomes like the horizontal plane in Figure 9.4. With a little thought, those familiar with the concept should be able to see that this is a spatial version of the *central limit theorem* of classical statistics.

Automating Interpolation (3): the Inverse Distance Weighted Spatial Average

So far we have taken account of spatial proximity by using only control points judged to be "near" to calculate a local mean. A further refinement to interpolating unknown values is to use *inverse distance weighting* when determining the mean. This method was the basis of SYMAP, which, in the early 1960s pioneered the application of computers to processing spatial data. SYMAP consisted of a few hundred lines of FORTRAN code, but a simplified version was published by the geologist John Davis (1976), and it has been re-invented since then by others (see, for example Unwin, 1981, pp. 172–174). Rather than treating all included sample locations equally, nearer locations are given more prominence in calculating the local mean. The simple local mean calculation is

$$\hat{z}_j = \frac{1}{m}\sum_{i=1}^{m} z_i \tag{9.4}$$

where $\hat{z}_j$ is the estimated value at the jth location and $\sum z_i$ is the sum of m neighboring control points. It is implicit that each control point inside the critical radius is weighted 1 and all those outside are weighted 0. As in other spatial analysis settings, this idea can be expressed using a spatial weights matrix, **W**, so that

$$\hat{z}_j = \sum_{i=1}^{m} w_{ij} z_i \tag{9.5}$$

where each w_{ij} is a weight between 0 and 1 and is calculated as a function of the distance from $\mathbf{s}_j$ to the control point at $\mathbf{s}_i$. If the distance is d_{ij}, then an obvious function to use is

$$w_{ij} \propto \frac{1}{d_{ij}} \tag{9.6}$$

This sets the weight proportional to the inverse of the distance between the point to be interpolated and the control point. If we want the w_{ij} values to sum to 1 (as we should: why?), then we set each weight equal to

$$w_{ij} = \frac{1/d_{ij}}{\sum_{i=1}^{m} 1/d_{ij}} \tag{9.7}$$

Large values of d_{ij} where the control point is distant are thus given small weights, whereas control points at short distances are given large weights.

To see how this works, consider the situation shown in Figure 9.9. Here, we want to estimate the height of the field at the point shown as an open circle using the nearest four control points, which have z values of 104, 100, 96, and 88. Table 9.1 shows the calculations involved, for simple inverse distance weighting. This gives the required estimate from the weighted sum as 95.63. The simple average of these four heights would be (104 + 100 + 96 + 88)/4 = 97, so the inverse distance-weighted result of 95.63 is biased toward the nearer, in this case lower, values.

The mathematically inclined will have spotted a problem that must be handled by the software: What happens if an estimated location is *exactly* coincident with a control point? The distance d_{ij} is zero, and when we divide

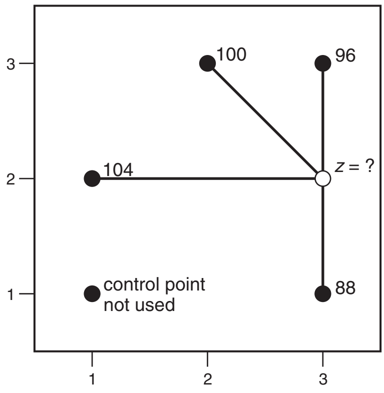


Figure 9.9 Inverse distance weighting in spatial interpolation.

Table 9.1 Illustrating Estimation Using Inverse Distance Weighting

Control point	*Height* z_i	x_i	y_i	*Distance* d_{ij}	*Inverse distance* $1/d_{ij}$	*Weight* w_{ij}	*Weighted value* $w_{ij}z_i$
1	104	1	2	2.000	0.50	0.1559	16.21
2	100	2	3	1.414	0.71	0.2205	22.05
3	96	3	3	1.000	1.00	0.3118	29.93
4	88	3	1	1.000	1.00	0.3118	27.44
Totals					3.21	1.0000	95.63

this into z, the result is undetermined. Because of this problem, inverse distance approaches test for the coincidence condition, and when it occurs, they use the control point value as the interpolated value. This is important because it guarantees that the method honors every data point. In the jargon, it is an *exact interpolator*. An equally important but less immediately obvious problem arising from the mathematics is that this method cannot predict values lower than the minimum or higher than the maximum in the data. This is a property of any averaging technique restricted to positive weights that sum to 1.

Inverse distance-weighted spatial averages are often used for interpolation in GIS. Given a set of control points, the first step is to lay a grid of points over the area. An interpolated value is then calculated for each point on the grid. The interpolated values on the grid may then be contoured to produce a surface representation. Contouring the interpolated grid is relatively simple. Even with this technique, there are at least three ways we could alter the procedure to change the final contour map:

1. *Specify a finer or a coarser grid* over which the interpolation is made. A very fine grid will add a lot of local detail; a coarse grid will produce a more generalized surface.
2. As for other local statistics, we may *alter the choice of neighboring control points* used. Whether we use near neighbors or a limited radius to choose them, as the number of control points increases, a smoother surface results.
3. We can *alter the distance weighting*. In the example, we used the actual distance, that is,

$$w_{ij} \propto \frac{1}{d_{ij}} \tag{9.8}$$

More generally, we can adjust the weight using an exponent k to arrive at the formula for the weights:

$$w_{ij} \propto \frac{1}{d_{ij}^k} \tag{9.9}$$

Higher values of k decrease the effect of more distant points and produce a "peakier" map, often with distinctive "bulls-eyes" around control points. Values less than 1 increase the effect of distant points and smooth the resulting map. We can also change the distance weighting function used. An alternative to inverse powers is the inverse negative exponential, given by

$$w_{ij} \propto e^{-kd_{ij}} \tag{9.10}$$

Whatever function is used, the calculations must still ensure that the weights at any interpolated point sum to 1.

Figure 9.10 shows two maps produced from the same data using $m = 12$ neighbors, but with the weights given by a simple inverse distance and inverse distance squared approaches. Although the general shape is similar, there are differences between the two maps. Many computer programs follow SYMAP's example and use $k = 2$ by default.

Relative to nonweighted schemes, one other point is worth making. As noted, the apparent continuity of the surfaces in Figure 9.8 is illusory, because the points included in each spatial average drop in and out of calculations abruptly. Inverse distance weighting changes this, so that the surface produced really does vary smoothly and continuously. This makes contouring of the interpolated surface from inverse distance weighting possible, as shown in Figure 9.10

It should be obvious that by changing any of the above aspects of the procedure, we can produce various reconstructions of a field from sample

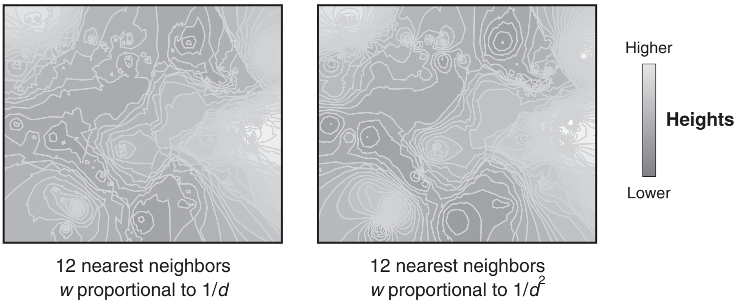


Figure 9.10 Different inverse distance weighted interpolation results.

data. Given the effectively infinite range of possible objective interpolation schemes, which is best? The answer is that there is no universally best scheme. It is up to the analyst to make sure that the chosen method is suited to the particular problem (see Rhind, 1971; Morrison, 1974; Braile, 1978). There are at least four ways that this issue can be addressed:

1. One simple way is to rely on the maps produced. Do they look reasonable? In a modern GIS, it is relatively easy to experiment with different settings and even techniques until a result that appears reasonable is obtained.
2. Alternatively, if a large number of control points are available, another approach is to perform interpolation using a subset of the data and to examine errors in the result relative to the unused control points. The preferred interpolation procedure is then the one that gives the smallest errors. This approach is known as *cross-validation*.
3. It is also possible to gain insight into the selection of the k parameter for the chosen weighting equation by running the interpolation at each control point location, but with the corresponding control point removed from the data. The interpolated values at each control point location are then used to estimate the overall error for the interpolation for that value of k. By repeatedly interpolating the control point data while varying k, a single best value can be efficiently determined (see Davis, 1976). This procedure is known as *leave one out cross-validation*.
4. Finally, in *kriging*, we use the control point data themselves to estimate the spatial structure in the underlying surface and use this information to determine appropriate spatial weights. We examine kriging more thoroughly in Chapter 10.

Automating Interpolation (4): Even More Options!

There are many other ways to interpolate a surface. Space and time don't allow us to go into these in detail, but mention should be made of three:

1. *Bicubic spline fitting* is a mathematical technique that finds contours that are the *smoothest* possible curves (in two dimensions) that can be fitted and still honor all the data. Logic suggests that this is a sensible, conservative approach. However, in regions where there are no nearby control points, this method can produce some very unlikely predictions.

2. *Multiquadric analysis* is a method developed by Hardy (1971) for application to topography. It is similar to the method of density estimation we applied to point data in that it centers a varying-sized circular cone on each of the n control points in the data. The cone sizes are themselves estimated so as to satisfy a series of linear equations that ensure that they honor all the data points exactly. The value of z at any point is then expressed as the sum of all the contributions from these quadric surfaces. We know of no GIS that implements this approach, but it has been widely used for the interpolation of rainfall data and is relatively easy to program.
3. One other interpolation technique deserves mention because it is widely used for terrain modeling and is especially relevant when the intention is to use the representation in computer visualizations. A TIN model for field data based on a set of sample point locations "tiles" the study region with triangles to the limit of the sample locations. In practice, the Delaunay triangulation (see Figure 2.5) is often used. This construction can be used to interpolate a z value for any location inside a triangle in the triangulation. We simply assume that each triangle is a flat facet and calculate z values based on this assumption. This is effectively an inverse distance-weighted approach based on the position of the unknown location in the surrounding triangle of points. The triangular structure makes it relatively easy to render images of a TIN mesh and also to generate reasonably realistic images of terrain.

All these methods are similar to inverse distance weighing in that they are *deterministic*. In every case, they assume that the data at the control points are exact and they use a deterministic, mathematical procedure to perform the interpolation. Given the data, method, and any required parameters, the results are uniquely determined. You may argue about the chosen parameters, but the results are verifiable and repeatable. An alternative is to once again use ideas about random processes and interpolate using statistical methods, as discussed in Chapter 10.

Spatial interpolation is straightforward using a computer to do the work, so you must ensure that the methods you use are appropriate for your specific problem. One problem that cannot be addressed adequately by any of the techniques discussed above is that sampled data points are not randomly distributed in space. You should consider the effects of this on the solutions obtained. For example, often sample points are denser in areas where people gathering the data felt it necessary. This might happen with mining surveys, for example, where more detailed investigations are carried out in promising areas. With climate data, sample points may be more densely clustered near centers of population. Unsampled locations in less densely sampled regions are inherently less well defined than those in more densely sampled regions. This is an important point to remember when reviewing the results of any interpolation procedure.

9.4. DERIVED MEASURES ON SURFACES

In most investigations, creating a continuous field by interpolation is a means to an end. We might, for example, want to know the field values for an ecological *gap analysis* where observations of the incidence of a plant or animal species are to be related to environmental factors like average January temperatures or mean annual rainfall. In addition to this direct use of interpolated values, it is sometimes necessary to derive measures that provide additional information about the *shape* of the field.

Although scalar fields are of interest in most branches of geography, physical geographers have tended to develop most of the available methods for summarizing and describing them. Geomorphologists interested in landforms have analyzed elevation fields, deriving useful information from properties such as average elevation, the frequency distribution of elevation values, and landform shape as described by slope and aspect. Similarly, climatologists have analyzed atmospheric pressure fields to derive the predicted geostrophic wind, while hydrologists have found the total basin precipitation from a field of precipitation depths. Often, similar methods have been developed and named independently, and there is an enormous variety of possible ways of describing surfaces. In the following account, we deal with a representative selection of descriptive and analytic measures of surfaces. Most of these have immediately obvious interpretations in the context of landscape (i.e., fields of elevation values), but many are applicable to other fields as well.

Relative Relief

Perhaps the simplest measure used is *relative relief*, which is the height range from the lowest to the highest point over some clearly specified area. A map of relative relief over a network of small grid squares gives a useful indication of the roughness of the surface and is yet another example of a local statistic. Prior to GIS and the widespread availability of accurate DEMs, relative relief was assessed by a variety of labor-intensive methods (see Clarke, 1966). In a DEM, relative relief is easy to compute. All that is required is to work across the grid, finding for each grid square the maximum and minimum height values within some defined neighborhood around that grid cell.

The Area/Height Relationship

Plots of the proportion of area at differing heights have been used frequently in geomorphology in attempts to detect the existence of flat planation surfaces (for reviews of the numerous available techniques, see Clark and Orrell, 1958; Dury, 1972). This can easily be derived from a histogram of height frequencies taken directly from a DEM. The ability to produce such frequency distributions often exists in GISs that handle raster data.

Slope and Gradient

A critical quantity when considering altitude is the slope of the ground surface. This obviously affects how easily we can walk up and down, and it is a key variable in the visualization of real landscapes. It is also central to an enormous range of ecological and geoscientific applications in GIS. Mathematically, the slope is the maximum rate of change of elevation at a point and is called the *gradient of the field*.

The left-hand side of Figure 9.11 is a contour map of a hill whose summit is at an elevation of 500 m. Suppose that we walk up to it from A, which is at 100 m. On the walk, we ascend through a vertical interval of 400 m and walk a plan view distance of 3 km. The tangent of the slope angle from A to B is thus

$$\tan\theta = \frac{\text{vertical interval}}{\text{horizontal distance}} = \frac{400}{3000} = 0.133 \qquad (9.11)$$

which is equivalent to an average *slope angle* of around 7.5°. We can specify a *slope* like this in any direction across the surface. Notice that this slope angle

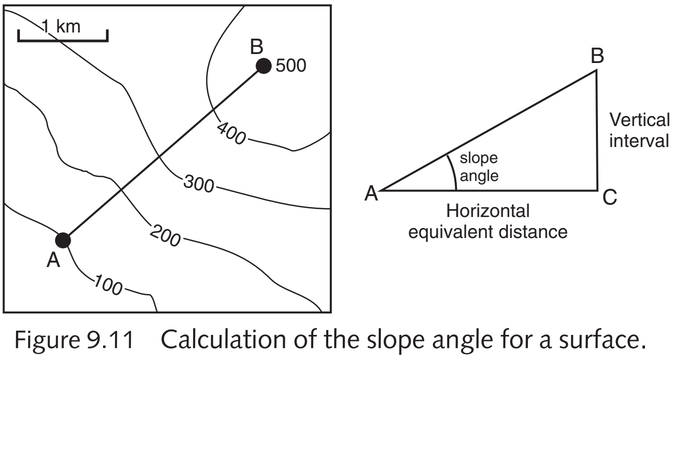


Figure 9.11 Calculation of the slope angle for a surface.

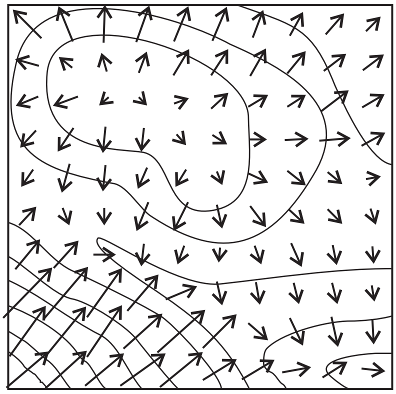


Figure 9.12 A vector field visualized using arrows pointing in the downslope direction of the gradient. Contour lines of the scalar field are also shown.

only applies along the direction AB and so is a *vector* quantity with both a magnitude (7.5°) and a direction (around 50° E of north). It should be apparent that slope can be measured in any direction. What is usually calculated in a slope map is the slope in the direction of the steepest slope through that point. This is the slope down which a dropped ball would roll, called the *fall line* by skiers. It is this slope that is termed the *gradient* of the field.

To display the vector field of gradient properly, one needs either two maps, one for each magnitude and direction, or a single map on which are drawn slope arrows with their heads facing in the correct direction and with lengths proportional to the slope magnitude. An example is shown in Figure 9.12.

Producing gradient maps from digital data is not as easy as it might appear. Most analysts working in GIS use either a TIN or a DEM as the starting point. Compared to the DEM case, it is easy to find the slope and aspect at a particular location using a TIN. We simply find the slope and aspect attributes of the containing triangle. For a DEM, the standard approach works across the grid point by point, calculating the gradient for each grid location in turn.

Figure 9.13 shows a typical grid point, P, surrounded by eight other points, each of which could usefully contribute information about the likely gradient at P. The different methods vary in the way they use this information. The simplest way is to assume that the slope across the four grid squares that meet at P is an inclined plane, as illustrated. The orientation of this plane can be specified by two slopes, one in the direction of the x-axis (θ_x), the other in the direction of y (θ_y). The slope in the x direction is estimated from the difference in the height values on either side of P as

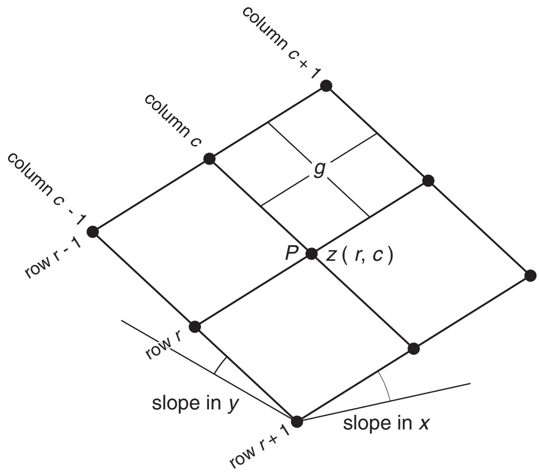


Figure 9.13 Estimating the gradient through a point in a DEM.

$$\tan\theta_x = \frac{z(r, c+1) - z(r, c-1)}{2g} \tag{9.12}$$

where g is the grid spacing in the same units as the z values. Similarly, the slope along the y direction is

$$\tan\theta_x = \frac{z(r+1, c) - z(r-1, c)}{2g} \tag{9.13}$$

These two slopes can then be resolved by Pythagoras's theorem to give the gradient:

$$\text{gradient at } P = \sqrt{\tan^2\theta_x + \tan^2\theta_y} \tag{9.14}$$

The direction, or *aspect*, of this gradient is found from

$$\tan\alpha = \frac{\tan\theta_x}{\tan\theta_y} \tag{9.15}$$

In using this method, one has to be confident that the original grid is sufficiently dense for the inclined plane assumption to be reasonable. It should be noted that use is made of information from only four neighboring points, and the central point, z (r, c) is ignored completely.

An alternative method using the concept of a locally valid analytical surface introduced in Section 9.2 has been suggested by Evans (1972). Briefly, this works its way across a grid of values performing a local operation that *approximates* the surface shape across each set of four grid squares using a quadratic polynomial surface fitted to all nine relevant points. This is fitted to observed heights by the method of least squares and is thus overdetermined, since only six points are needed to define the polynomial, whereas nine are available. The result is that each quadratic is not an exact fit to the nine points, so that there are potential difficulties if the lack of fit is large. Evans reports that discrepancies are not serious, at least for elevation data. All that remains is to find the gradient of the locally fitted quadratic function. This is accomplished by direct differentiation of the equation of the fitted surface.

In conclusion, two issues should be noted. The first is scale and the grid spacing of the DEM used. The gradient at a point is a mathematical limit as the distance over which it is measured goes to zero, but in practice, we evaluate the gradient over a distance equal to $2g$, twice the grid spacing. This means that any measure obtained is an estimate of the true gradient at the point, and will tend to smooth out steeper slopes and to miss detail in the surface relief. Second, many DEM data products are themselves interpolated from contours, and to save on computer memory, height values may be rounded to the nearest whole number of meters. In areas of low relative relief, the errors this introduces in both aspect and slope values can be significant.

Surface Specific Points and the Graph of a Surface

Whatever method is used to find the gradient of a field, it will sometimes give a value of exactly zero, indicating that the surface is locally flat. Examination of the formulae given in the previous section indicates that a zero gradient occurs if, and only if, both of the slopes along x and y are zero. This will be the case at the top of a hill or at the bottom of a pit. In the case of ridges and valley lines, the slope in at least one direction will be zero. Of course, if elevations are measured to high precision, the chance of two grid point values being exactly the same is very small. Usually, the precision of the data is such that rounding z values to the nearest convenient integer generates apparently equal values. It follows that points with zero gradient will occur on most surfaces. Having no gradient also means that the aspect is vertically upward and thus cannot easily be mapped. Such points are *surface specific* and are of six types:

1. Peaks higher than their immediate neighborhood
2. Pits lower than their immediate neighborhood
3. Saddles that lie at the self-crossing of figure-of-eight contours
4. Ridge lines
5. Flat valley bottoms or channels
6. Plains that are flat in every direction

Algorithms have been developed for the automatic detection of surface specific points. An interesting branch of surface analysis uses the distribution of pits, saddles, and summits, together with their connecting ridges (from a saddle to a peak) and channels (from a pit to a saddle), as a way of characterizing surface forms topologically. The lines connecting these surface specific points form a *surface network* whose properties can be analyzed using graph theory (Pfalz, 1976; Rana, 2004).

Catchments and Drainage Networks

Two major aspects of a drainage basin are its topographic form and the topologic structure of its drainage network. The manual quantification of these components is tedious and time-consuming. Automated determination is an ideal application of GIS technology, since watersheds comprise a method that completely partitions space and many environmental phenomena can be related to them. Furthermore, knowledge of drainage divides and drainage networks can be used to provide better estimates of slopes and aspects, since slopes should break at divides and at channels. Determination of drainage networks and the associated drainage divides is an important first step in the creation of an effective hydrologic information system.

A DEM contains sufficient information to determine general patterns of drainage and watersheds. The trick is to think of each grid height value as the center of a square cell and determine the direction of water flow out of this cell by inspection of the altitudes of the surrounding cells. Algorithms to determine flow direction based on this idea generally assume only four possible directions of flow (up, down, left, right—the *Rook's case*) or, occasionally, eight possible directions (the *Queen's case*). Each possible flow direction is numbered. A typical algorithm sweeps the entire DEM, labeling each cell by the assumed direction of water movement. Pits in the DEM are treated separately, usually by "flooding" them with virtual "water" until an outlet direction is found. To determine the drainage network, the set of flow directions is then connected with arrows. Since, in natural systems, small quantities of water generally flow overland, not in channels, we may also want to accumulate water flowing downstream through cells so that channels begin only when a threshold volume is reached.

Simulated drainage networks cannot capture all the detail of a real stream network. For example, real streams sometimes branch downstream, which cannot occur using the method described. Also, the number of streams joining at a junction, known as the *valency* of the junction, is almost always three in reality but may be as many as eight when an eight-direction algorithm is used. Junction angles are determined by the cell geometry in the simulation, but in reality they are a function of the terrain and of erosion processes. Finally, in areas of uniform slope, the technique generates large numbers of parallel streams, whereas in reality, streams tend to wander because of surface unevenness and the resulting junctions reduce the density of streams in such areas. As a result, the length of stream channel per unit of surface area, the *drainage density*, is often too high in simulations. Some of these limitations can be overcome using considerably more complex dynamic models based on TINs (see, for example, Tucker et al., 2001).

Similar logic can be used to determine the *watershed* of a point. This is an attribute of each point on the network and is given by the region upstream of the point that drains toward it. Using a grid of flow directions developed as above, it is easy to find the watershed of any cell. Simply begin at the specified cell and label all cells that drain to it, then all cells that drain to those, and so on, until the upstream limits of the basin are defined. The watershed is then the polygon formed by the labeled cells.

Viewsheds

Another surface operation is the calculation of a visibility region, or *viewshed*. Originally, programs were written to compute the line of sight from a point in a specified direction. As machines have become faster, it has been possible to extend these methods to estimate and map all the ground visible or potentially visible from a specified point. Viewsheds have application in the military; in locating unsightly constructions like pipelines, wind farms, and electricity lines; and in landscape architecture, where they are used in studies of landscape attractiveness. Communications companies also use viewsheds to locate transmitter and receiver towers.

As in finding gradients or determining surface specific points, calculating the viewshed of a point is an operation conducted locally on each grid point in a DEM. Some algorithms replicate the manual method of "drawing" a series of profiles radially out from the viewpoint, marking on each the hidden segments, and transferring these back to the base map. A simpler method finds the intervisibility to all other points making up the DEM. An imaginary profile is drawn from a viewpoint to every other grid point in turn (to some limit determined by the scale of the DEM, earth curvature, and so on), and successive heights along each profile where it crosses a grid line are listed and used to determine whether or not the point is visible (see Burrough and McDonnell, 1998, for an overview). Algorithms for finding viewsheds that use a TIN data model have also been described (DeFloriani and Magillo, 1994).

Surface Smoothing

Another operation carried out on surface data is *smoothing and generalization*. Typically, this is necessary when displaying a relief map at lower spatial resolution than that at which data were collected and stored. The standard approach to smoothing makes use of the idea of a moving average calculated across a field, where the height of every data point across a grid of values is replaced by the average of its near neighbors. Inevitably, this reduces the variance of the entire grid of values and results in a smoother map. The technique also has the potentially undesirable side effect of occasionally erasing significant hilltops and valley floors. More sophisticated algorithms avoid this problem.

9.5. MAP ALGEBRA

A framework that is often used for thinking about all of the surface analysis methods described above (and many more besides) is *map algebra*. It is most readily understood in the case of field data that are stored as a grid of values, but in principle, it is applicable to any type of field data. Map algebra was devised by Dana Tomlin and is presented in his 1990 book *Geographical Information Systems and Cartographic Modeling*, which you should consult for a much more detailed treatment than is given here (see also DeMers, 2001). Many GISs support map algebra, although this is often hidden behind other terminology such as *map calculator*.

The fundamental concepts of map algebra are exactly the same as those of mathematical algebra:

- *Values* are the things on which the algebra operates. Input data and output data (results) are presented as grids of values. Values can be categorical (nominal or ordinal) as well as numerical.
- *Operators* may be applied to values to transform them, or between two or more values to produce a new value. In mathematical algebra, the

(i)

7.6	8.1	8.0	8.6	8.2
7.8	8.1	8.8	8.8	8.6
8.2	8.6	8.9	8.0	8.9
7.8	8.0	8.3	8.1	8.2
8.6	8.2	8.3	9.0	8.6

8.4	8.6	7.9	7.7	7.1
8.8	8.2	8.1	7.7	7.1
9.1	8.3	8.4	7.6	7.9
8.8	9.1	8.9	8.2	7.6
9.0	8.8	8.9	8.8	8.3

(ii)

-8.4	-8.6	-7.9	-7.7	-7.1
-8.8	-8.2	-8.1	-7.7	-7.1
-9.1	-8.3	-8.4	-7.6	-7.9
-8.8	-9.1	-8.9	-8.2	-7.6
-9.0	-8.8	-8.9	-8.8	-8.3

(iii)

16.0	16.7	15.9	16.3	15.3
16.6	16.3	16.9	16.5	15.7
17.3	16.9	17.3	15.6	16.8
16.6	17.1	17.2	16.3	15.8
17.6	17.0	17.2	17.8	16.9

(iv)

8.4	8.6	8.0	8.6	8.2
8.8	8.2	8.8	8.8	8.6
9.1	8.6	8.9	8.0	8.9
8.8	9.1	8.9	8.2	8.2
9.0	8.8	8.9	9.0	8.6

(v)

8.1	8.8	8.8	8.8	8.8
8.6	8.9	8.9	8.9	8.9
8.6	8.9	8.9	8.9	8.9
8.6	8.9	9.0	9.0	9.0
8.6	8.6	9.0	9.0	9.0

Figure 9.14 Example small grids: (i) [left_grid] and [right_grid], (ii) –[right_grid], (iii) [left_grid] + [right_grid], (iv) result of a local maximum operation applied between [left_grid] and [right_grid], and (v) result of a focal maximum operation applied to [left_grid] using the shaded focal shape shown.

minus sign "–" is an operator that negates a single value when placed in front of it, as in –5. The plus sign "+" is also an operator, signifying the addition operation, which, when applied between two values, produces a new value: $1 + 2 = 3$.

- *Functions* are more complex but still well-defined operations that produce a new value from a set of input values. The input set may be a single value, as in $log_{10}(100) = 2$, or a set of values, as in *mean* $(\{1, 2, 3, 4\}) = 2.5$.

Now consider two small grids of values like those in Figure 9.14(i). If we want to apply an operation or function to these values, how should we proceed? In fact, we have a number of options, and map algebra clearly defines them, as described below. Note that we refer to these grids as [left_grid] and [right_grid] where necessary in the discussion.

Local Operations and Functions

A *local* operation or function in map algebra is applied to each individual cell value in isolation. For example, the local negation operation signified by the minus sign "–" and applied to the right-hand grid in Figure 9.14(i) results in the output grid, –[right_grid], in Figure 9.14(ii).

Applying a local operation between two grids involves applying the operation to values in corresponding positions in each grid and recording the result in the corresponding position in the output grid. The result of the + operation applied between the two grids in Figure 9.14(i) is shown in 9.14(iii). Another example is a local maximum operation between two (or more) grids, which assigns to each output location the maximum of the values at the corresponding location in the input grids. The result of this operation applied to the grids in Figure 9.14(i) is shown in 9.14(iv).

Focal Operations and Functions

We can also apply an operator or, more often, a function, *focally* to a grid. This means that the value at each location in the output grid is arrived at by combining values focused at the corresponding location in the input grid or grids. A simple example is *focal_max*, which would assign to each output location the maximum of the values among those at the location itself and its immediate neighbors in the input grid. The result of applying a focal maximum function to [left_grid] is shown in Figure 9.14(v).

Many functions can be applied focally in this way, such as *maximum*, *minimum*, *mean*, *median*, *standard deviation*, *range*, and so on. In addition to the function itself, the output grid will depend on how the focal neighborhood is defined in a particular case. In the above example, the focal neighborhood is the grid cell itself and its eight immediate neighbors. Some alternative neighborhood definitions are shown in Figure 9.15.

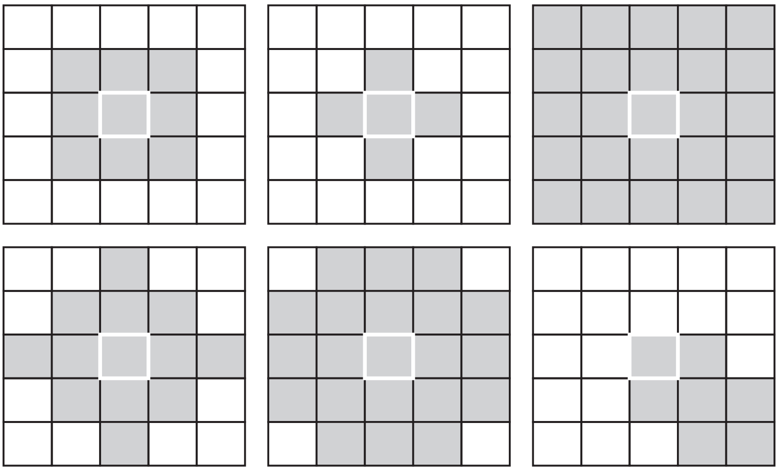


Figure 9.15 Some alternative possible definitions of the focal neighborhood relative to the central cell of this small grid.

A different choice of focal neighborhood will alter the output grid that results when a focal function is applied. Notice that there is no requirement that the neighborhood be symmetrical about the focal grid cell, as shown in the last example in Figure 9.15. A nonsymmetrical neighborhood might have application in understanding how air pollution spreads given a prevailing wind direction.

Zonal Operations and Functions

Zonal operations and functions are an extension of the focal concept. Rather than define operations with respect to each grid cell, a set of map zones are defined (for example, counties, census tracts, or regions of some specified land use) and operations or functions are applied with respect to these zones. Zonal operations are generally used to summarize the characteristics of the regions in question. The summary might be in terms of the mean, variance, or total amount of the phenomenon in question with respect to the surface.

Global Operations and Functions

Finally, some operations and functions are *global*, meaning that the values at each grid cell in an output grid may potentially depend on the values at all grid cell locations in the input grid(s). An operation that finds the cost (in time or money) of the shortest path from a specified location (say, a school) to every other location may have to take into account values at all locations in a grid to find the correct answer (the travel cost might, for example, be based on the land cover type and its slope).

9.6. CONCLUSIONS

Many important environmental phenomena, such as altitude, temperature, and soil pH, are interval or ratio scaled and form continuous, single-valued scalar fields. However, in almost all cases, our knowledge of the precise form of these surfaces is limited to a sparse and inadequate pattern of control points where measurements have been made. Usually, it is prohibitively expensive to make measurements at all the locations where values are required for use in subsequent studies, so we must interpolate from the sample data to reconstruct the complete surface.

Most GISs have capabilities to enable the creation of interpolated surfaces at the click of a mouse. The intention of this chapter has been to persuade you that this operation must be approached with some caution.

Depending on the approach used, very different results may be obtained. At the very least, find the system documentation and check how the system vendor decided to do the interpolation. Where the details are vague—for whatever reason—you should proceed with the utmost caution. Interpolation results reliant on the secret inner workings of a particular GIS are probably less useful than contours hand drawn by a human expert and are certainly less useful than the original control point data, which at least allow subsequent users to draw their own conclusions. Whenever you perform interpolation on field data control points, best practice is always to indicate the exact method used.

CHAPTER REVIEW

- *Scalar fields* are continuous, single-valued differentiable functions in which the attribute value is expressed as a function of location.
- They can be recorded and stored in a variety of ways, including DEMs, TINs, as explicit mathematical functions, or using digitized contours.
- *Spatial interpolation* refers to techniques used to predict the value of a scalar field at unknown locations from the values and locations of a set of survey locations or control points. Simple interpolation methods are based on local statistics.
- A *proximity polygon or nearest-neighbor approach* is appropriate for nominal data but results in a "blocky" or stepped estimate of the underlying field, which may not be plausible for interval or ratio data.
- The nearest-neighbor approach is extended by basing local estimates on *spatial averages* of more than one near neighbor. These may be chosen either on the basis of some limiting distance or on the basis of the *m* nearest neighbors. The resulting fields become progressively smoother as more neighbors are included. Eventually, when all control points in the study region are included, this is identical to the simple average. The interpolated surface does not necessarily honor the control points.
- The most popular approach is *inverse distance-weighted averages* that make use of an inverse power or negative exponential function to give more weight to nearby sample values in the calculation of spatial averages.
- Alternative interpolation techniques are based on *bicubic splines*, *multiquadrics*, and threading *contours across a TIN*.
- All these techniques are *deterministic* in the sense that once the method and any necessary controlling parameters are set, only one solution surface is possible.

- There are many ways that fields can be analyzed to assist in interpretation and understanding. These include calculation of the vector field given by the *gradient*, the identification of *watersheds* and *drainage networks*, *viewsheds* around a point, and *smoothed and generalized versions*.
- A useful framework for thinking about and extending surface analysis is *map algebra*.

REFERENCES

Braile, L. W. (1978) Comparison of four random-to-grid methods. *Computers and Geosciences*, 14: 341–349.

Burrough, P. A. and McDonnell, R. (1998) *Principles of Geographical Information Systems*, 2nd ed. (Oxford: Clarendon Press).

Clarke, J. I. (1966) Morphometry from maps. In: G. H. Dury (ed.), *Essays in Geomorphology* (London: Heinemann), pp. 235–274.

Clarke, J. I. and Orrell, K. (1958) An assessment of some morphometric methods. Durham, England: University of Durham Occasional Paper No. 2.

Davis, J. C. (2003) *Statistics and Data Analysis in Geology*, 3rd ed. (Hoboken, NJ: Wiley).

Davis, J.C. (1976) Contouring algorithms. In: *AUTOCARTO II, Proceedings of the International Symposium on Computer-Assisted Cartography* (Washington, DC: U.S. Bureau of the Census), pp. 352–359.

De Floriani, L. and Magillo, P. (1994) Visibility algorithms on triangulated terrain models. *International Journal of Geographical Information Systems*, 8 (1): 13–41.

DeMers, M. (2001) *GIS Modeling in Raster* (New York: Wiley).

Dury, G. H. (1972) *Map Interpretation*, 4th ed. (London: Pitman), pp. 167–177.

Evans, I. S. (1972) General geomorphometry, derivatives of altitude and descriptive statistics. In: R. J. Chorley, ed., *Spatial Analysis in Geomorphology* (London, England: Methuen), pp. 19–90.

Hardy, R. L. (1971) Multiquadric equations of topography and other irregular surfaces. *Journal of Geophysical Research*, 76 (8): 1905–1915.

Li, X. and Hodgson, M. E. (2004) Vector field data model and operations. *GIScience and Remote Sensing*, 41 (1): 1–24.

McQuistan, I. B.(1965) *Scalar and Vector Fields: A Physical Interpretation* (New York: Wiley).

Morrison, J. L. (1974) Observed statistical trends in various interpolation algorithms useful for first stage interpolation. *Canadian Cartographer*, 11 (2): 142–159.

Pfalz, J. L. (1976) Surface networks. *Geographical Analysis*, 8 (1): 77–93.

Rana, S., (ed.), (2004) *Topological Data Structures for Surfaces* (Chichester, England: Wiley).

Rhind, D. W. (1971) Automated contouring: an empirical evaluation of some differing techniques. *Cartographic Journal*, 8: 145–158.

Thiessen, A. H. (1911) Precipitation averages for large areas. *Monthly Weather Review*, 39 (7): 1082–1084.

Tobler, W. (1970), A. computer movie simulating urban growth in the Detroit region. *Proceedings of the I.G.U. Commission on Quantitative Methods. In Economic Geography*, 46, Supplement, June 1970

Tomlin, D. (1990) *Geographical Information Systems and Cartographic Modeling* (Englewood Cliffs, NJ: Prentice Hall).

Tucker, G. E., Lancaster, S. T., Gasparini, N. M., Bras, R. L., and Rybarczyk, S. M. (2001) An object-oriented framework for distributed hydrologic and geomorphic modeling using triangulated irregular networks. *Computers & Geosciences*, 27 (8): 959–973.

Unwin, D. J. (1981) *Introductory Spatial Analysis* (London: Methuen).

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 10

Knowing the Unknowable: The Statistics of Fields

CHAPTER OBJECTIVES

In this chapter, we:

- Describe the application of multivariate regression where the independent variables are spatial coordinates, that is, the method known as *trend surface analysis*
- Show how the *variogram cloud* and *semivariogram* can be used to describe the spatial structure of an observed field of data
- Describe interpolation by the method known as *kriging* in general terms, with reference to the discussion of least squares regression and the semivariogram
- Introduce variations on kriging that enable the same concepts to be applied to the analysis of other types of spatial data

After reading this chapter, you should be able to:

- Show how standard multiple linear regression can be developed using the spatial coordinates of some observations to give the geographic technique of trend surface analysis
- State the difference between trend surface analysis and *deterministic spatial interpolation* of the type undertaken in Chapter 9
- Implement a trend surface analysis using either the supplied function in a GIS, spreadsheet, or standard package program for statistical analysis

- Outline how the *semivariogram* that summarizes the spatial dependence in some geographic data can be used to develop a model for this variation and estimate its parameters
- Outline how a model for the semivariogram is used in optimum interpolation by kriging
- Describe some variations on this approach
- Make a rational choice when interpolating field data between inverse distance weighting, trend surface analysis, and geostatistical interpolation by kriging

10.1. INTRODUCTION

In Chapter 9, we examined some simple methods for spatial interpolation in which we reconstruct a field using the evidence supplied by some control points—the locations at which we know the height of the field. All of these methods make simplifying assumptions about the underlying spatially continuous phenomenon of interest, and all are deterministic in the sense that they use some specified deterministic mathematical function as their interpolator. Many geostatisticians argue that deterministic interpolators are unrealistic for two reasons:

- Because no environmental measurements can be made without error, almost all control point data have errors. Furthermore, measured values are often a "snapshot" of some changing pattern, which means that we ought to consider their time variability. From this viewpoint, it is ill-advised to try to honor all the observed data without recognizing the inherent variability.
- In choosing the parameters that control deterministic interpolators, we make use of very general domain knowledge about how we expect, say, rainfall totals or temperatures to vary spatially. Other than this, these methods assume that we know nothing about how the variable being interpolated behaves spatially. This is foolish, because we have—in the observed control point data—evidence of the spatial behavior, and any sensible interpolation technique should make use of this information.

In this chapter, we examine two approaches to the analysis of fields that are *statistical,* rather than mathematical, in nature. The first, *trend surface analysis*, is a variation on ordinary least squares regression, where specified functions are fitted to the locational coordinates (x, y) of the control point data to approximate trends in height, z, across the region of interest. Trend surface analysis is not widely used as a basis for interpolation of surfaces, although it is occasionally used as an exploratory method to give a rough idea of the spatial pattern in a set of observations. The second group of techniques, called *kriging*, attempts to make as much use as possible of the available control point data to develop an interpolator that models the underlying phenomenon in what is sometimes thought to be the *optimum* manner. This approach also makes simplifying assumptions about measurement variability, but attempts to include it *and* estimates of the data autocorrelation in the interpolation process.

In both cases, and unlike the methods discussed in Chapter 9, some measure of the error involved can be produced. Kriging is a sophisticated technique, widely used by earth scientists in mining and similar industries. There are many variants of the basic method, and we do not attempt to cover them all here. If you can understand the basic concepts on which kriging is based, you will be well equipped to deal with many of the more specialized variations on this approach.

10.2. REGRESSION ON SPATIAL COORDINATES: TREND SURFACE ANALYSIS

The methods examined in Chapter 9 are all exact interpolators that honor the data control points. In this section, we outline a technique called *trend surface analysis*, which, rather than honoring the data, deliberately generalizes the field into its major feature, or “trend.” In this context, the *trend* of a surface is a *global* property, any large-scale, systematic change that extends smoothly from one map edge to the other. Often, it might be appropriate to consider this as the first-order spatial pattern, as discussed in earlier chapters. Examples of such systematic trends might be the dome of atmospheric pollution over a city, the dome of population density over a city, or a north–south trend in mean annual temperatures.

The basics of trend surface analysis are very simple and are at heart a simple extension of multiple linear regression. We briefly discuss regression in Section 8.5 when we consider geographically weighted regression, but if necessary, you should consult any introduction to statistics for a basic account of this method.

Recall, first, that any *scalar field* can be represented by the equation

$$z_i = f(\mathbf{s}_i) = f(x_i, y_i) \tag{10.1}$$

which relates surface height (z) to each location, **s**, and its georeferenced pair of (x, y) coordinates. As it stands this is vague, since f denotes an unspecified function. Trend surface analysis specifies a precise and known mathematical form for this function and then fits it to the observed data using conventional least squares multiple linear regression. It is extremely unlikely that any simple function will exactly honor observed data, for two reasons. First, even where the underlying surface is simple, measurement errors will occur in the observed data. Second, it is unlikely that only one trend-producing process is in operation. It follows that there will be local departures from the trend, or *residuals*. Mathematically, we denote this as

$$z_i = f(\mathbf{s}_i) + \varepsilon_i = f(x_i, y_i) + \varepsilon_i \qquad (10.2)$$

That is, the surface height at the ith point is made up of the fitted trend surface component at that point plus a residual, or error, at that point.

The problem in trend surface analysis is to decide on a functional form for the trend part of the equation. There is an enormous range of candidate functions, but the simplest trend surface imaginable is an inclined plane, which can be specified as

$$z_i = \beta_0 + \beta_1 x_i + \beta_2 y_i + \varepsilon_i \qquad (10.3)$$

Mathematically, the trend is a linear polynomial, and the resulting surface is a *linear trend surface*. To calculate values for the trend part of this equation, we need to know the constant parameters β_0, β_1, and β_2 together with the coordinates of points of interest. These constants have a simple physical interpretation as follows. The first, β_0, represents the height of the plane surface at the map origin, where $x_{\mathrm{i}} = y_i = 0$. The second, β_1, is the surface slope in the x-direction, and the third, β_2, gives its slope in the y-direction.

This is illustrated in Figure 10.1. The linear trend surface is shown as a shaded plane passing through a series of data points, each shown as a circle. Some of the observed data points, in white, lie above the trend surface, while those below the surface are shaded gray. The trend surface is the one that best fits the observed control point data using the least squares criterion. It is thus exactly the same as a conventional regression model using as its two independent variables the locational coordinates.

At this point, the mathematical notation may become a little confusing. As mentioned in Section 8.5 (see Equation 8.10), the least squares solution to this problem is given by

$$\boldsymbol{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{z} \qquad (10.4)$$

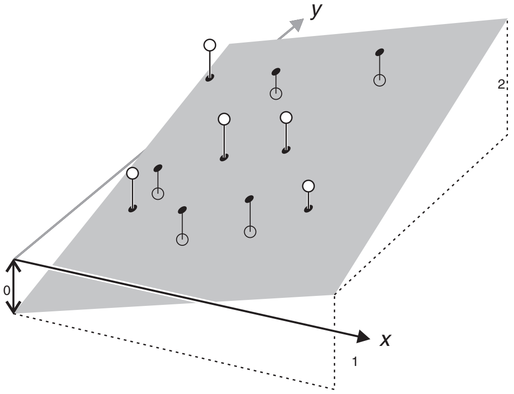


Figure 10.1 A simple linear trend surface.

where the augmented data matrix **X** is given by

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & y_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & y_n \end{bmatrix} \tag{10.5}$$

and **β** and **z** are vectors containing the estimated regression coefficients and the observed height values, respectively.

It is easiest to appreciate how this works by looking at an example. Table 10.1 shows the Alberta temperature data first considered in Section 9.3 and displayed in Figure 9.4. A series of temperatures have been observed across a surface to which a linear trend surface is to be fitted. The first step is to measure the locational coordinates (x, y). The second and third columns of Table 10.1 show these values, together with the temperatures z.

Table 10.1 Temperatures in Alberta in January (°F)

Control Point	x	y	z
1	1.8	0.8	11.5
2	5.7	7.1	12.6
3	1.2	45.3	2.4
4	8.4	57.1	−6.6
5	10.2	46.7	−7.9
6	11.4	40.0	1.0
7	15.9	35.4	2.5
8	10.0	30.9	7.1
9	15.7	10.0	8.4

(*Continues*)

Table 10.1 (Continued)

Control Point	*x*	*y*	*z*
10	21.1	17.5	5.0
11	24.5	26.4	10.0
12	28.5	33.6	3.1
13	33.5	36.5	1.7
14	36.4	42.9	0.4
15	35.0	4.7	7.4
16	40.6	1.6	7.2
17	39.9	10.0	6.6
18	41.2	25.7	1.5
19	53.2	4.4	2.9
20	55.3	8.3	2.9
21	60.0	15.6	−0.9
22	59.1	23.2	0.0
23	51.8	26.8	1.4
24	54.7	54.0	−6.3

The boxed section below shows the required calculations, but for accuracy and to reduce labor, this type of calculation would almost always be done using standard computer software or a built-in function in a GIS.

Calculating the Best-Fit Linear Surface

We proceed exactly as with multiple linear regression, described above, with the coordinates as two of the independent variables. The augmented matrix of the data, **X**, is thus

$$\mathbf{X} = \begin{bmatrix} 1 & 1.8 & 0.8 \\ \vdots & \vdots & \vdots \\ 1 & 54.7 & 54 \end{bmatrix}$$

To save space, we have not written out the full matrix, which has 24 rows and 3 columns. Its transpose, $\mathbf{X}^\mathrm{T}$, thus has 3 rows and 24 columns:

$$\mathbf{X}^\mathrm{T} = \begin{bmatrix} 1 & \cdots & 1 \\ 1.8 & \cdots & 54.7 \\ 0.8 & \cdots & 54 \end{bmatrix}$$

Simple but tedious multiplication gives us

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} 1 & \cdots & 1 \\ 1.8 & \cdots & 54.7 \\ 0.8 & \cdots & 54 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1.8 & 0.8 \\ \vdots & \vdots & \vdots \\ 1 & 54.7 & 54 \end{bmatrix} = \begin{bmatrix} 24 & 715.1 & 604.5 \\ 715.1 & 30065.23 & 16324.6 \\ 604.5 & 16324.6 & 22046.47 \end{bmatrix}$$

Again, to save space, we have not written out each matrix in full. Note how a 3 x 24 matrix post multiplied by a 24 by 3 matrix gives a symmetric 3 by 3 result (see the Appendix). The next step is to invert this matrix. Take our word for it that this is

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 0.290278546 & -0.004319108 & -0.00476111 \\ -0.004319108 & 0.00011989 & 2.9653 \times 10^{-5} \\ -0.00476111 & 2.9653 \times 10^{-5} & 0.000153948 \end{bmatrix}$$

Although we have done the work of finding this inverse for you (it's not so hard with a computer!), notice that we retain as many digits in the working as possible. With this in mind, we have shown the smallest number in this inverse, which is 0.000029653, in exponent/mantissa form as 2.9653×10^{-5}. Many years ago, one of us (Unwin, 1975a) noted that the $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ matrix that arises in polynomial trend surface analysis when dealing with more complex functions than the linear is often what numerical analysts call ''ill-conditioned'', making blind reliance on a computer, with its fixed and finite numerical precision, sometimes hazardous. Inversion of such matrices can be very sensitive to even small changes in the element values. Retention of as much precision as possible at each stage of the calculation is therefore advisable.

To determine $\boldsymbol{\beta}$, we also need $\mathbf{X}^{\mathrm{T}}\mathbf{z}$, which is

$$\mathbf{X}^{\mathrm{T}}\mathbf{z} = \begin{bmatrix} 1 & \cdots & 1 \\ 1.8 & \cdots & 54.7 \\ 0.8 & \cdots & 54 \end{bmatrix} \cdot \begin{bmatrix} 11.5 \\ \vdots \\ -6.3 \end{bmatrix} = \begin{bmatrix} 73.9 \\ 1588.79 \\ 299.31 \end{bmatrix}$$

Note that here multiplying a 3 by 24 matrix by a 24 by 1 column vector produces a 3 by 1 column vector. The final least squares solution is given by the product of the two intermediate matrices:

$$\begin{aligned} \boldsymbol{\beta} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{z} \\ &= \begin{bmatrix} 0.290278546 & -0.004319108 & -0.00476111 \\ -0.004319108 & 0.00011989 & 2.9653 \times 10^{-5} \\ -0.00476111 & 2.9653 \times 10^{-5} & 0.000153948 \end{bmatrix} \begin{bmatrix} 73.9 \\ 1588.79 \\ 299.31 \end{bmatrix} \\ &= \begin{bmatrix} 13.16438146 \\ -0.119826593 \\ -0.258655349 \end{bmatrix} \end{aligned}$$

(*Continues*)

(*box continued*)

Again, to guard against numerical errors, we have retained as many digits as possible. Our best-fit, linear trend surface for these data is thus the inclined plane described by the equation

$$\hat{z}_i = 13.16 - 0.1198x_i - 0.2587y_i$$

in which $\hat{z}_i$ is the estimated temperature at location s_i with coordinates (x_i, y_i). A contour map of this surface is shown in Figure 10.2. It is apparent that the surface does not honor the data, although the overall trend is reasonable, with temperatures falling from southwest to northeast, as we might expect in the Northern Hemisphere.

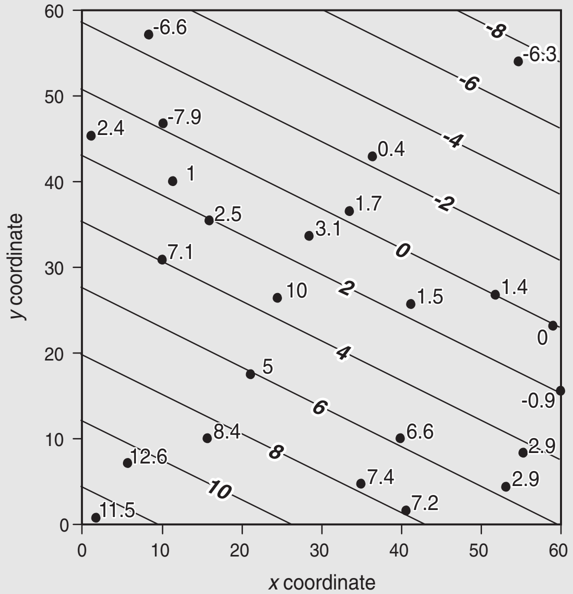


Figure 10.2 Contours of the least squares linear trend surface fitted to the data of Table 10.1 and Figure 9.4.

In some studies, it is the form of this trend that is of major interest, but in other studies, interest may also center on the distribution of the local residuals. From the previous equations, it is obvious that these can be calculated as

$$\varepsilon_i = z_i - (\beta_0 + \beta_1 x_i + \beta_2 y_i) \tag{10.6}$$

That is, the residual at each point is given by the difference between the observed surface height at that point and the value predicted by the fitted surface. Maps of residuals are a useful way of exploring the data to suggest local factors that are not included in the trend surface.

Finally, it is customary to derive an index of how well the surface fits the observed data. This is provided by comparing the sum of squared residual values for the fitted surface to the sum of squared differences from the simple mean for the observed z values. This is better known as the *coefficient of determination* used in standard regression analysis and given by the square of the coefficient of multiple correlation, R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2} = 1 - \frac{\text{SSE}}{\text{SS}_z} \tag{10.7}$$

where SSE stands for “sum of squared errors” and SS_z stands for “sum of squared differences from the mean.” This index is conventionally used in regression analysis in general and indicates how much of an improvement the fitted trend surface is compared to simply using the data mean to predict unknown values. If the residuals are large, then SSE will be close to SS_z and R^2 will be close to 0. If the residuals are near 0, then R^2 will be close to 1. In the boxed example, R^2 is 0.732, indicating a reasonably good fit between the trend surface and the observed data.

Whether or not this fit is statistically significant can be tested using an F-ratio statistic

$$F = \frac{R^2/\text{df}_{\text{surface}}}{(1 - R^2)/\text{df}_{\text{radius}}} \tag{10.8}$$

where $\text{df}_{\text{surface}}$ is the degrees of freedom associated with the fitted surface, equal to the number of constants used, less 1 for the base term β_0, and $\text{df}_{\text{residuals}}$ is the degrees of freedom associated with the residuals, found from the total degrees of freedom ($n - 1$), less those already assigned, that is, $\text{df}_{\text{surface}}$. In the example, $\text{df}_{\text{surface}} = 3 - 1 = 2$ and $\text{df}_{\text{residuals}} = 10 - 1 - 2 = 7$, so that

$$F = \frac{0.732/2}{0.268/21} = \frac{0.3658}{0.01279} = 28.608 \tag{10.9}$$

This F-ratio indicates that the surface is statistically significant at the 99% confidence level, and we can assert that the trend is a real effect and is not due to chance sampling from a population surface with no trend of the specified linear form.

If this test had revealed no significant trend in the data, several explanations might be adduced. One possibility is that there really is no trend of any sort across the surface. Another is that there is a trend in the underlying surface but our sample size, n, is too small to detect it. A third possibility is that we have fitted the wrong sort of function. No matter how we change the values of the $\boldsymbol{\beta}$ parameters in our linear trend equation, the result is always a simple inclined plane. Where this does not provide a significant fit, or where geographic theory might lead us to expect a different shape, then other, more complex, surfaces may be fitted. Exactly the same technique is used, but the calculations rapidly become extremely lengthy. Suppose, for example, that we wish to fit a dome or trough-like trend across the study area. The appropriate function to use is a quadratic polynomial, giving a surface:

$$z_i = f(x_i, y_i) = \beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 x_i y_i + \beta_4 x_i^2 + \beta_5 y_i^2 + \varepsilon_i \tag{10.10}$$

This is still a basic trend model, but there are now six parameters, β_0 to β_5, to be estimated, and you should be able to see that $\mathbf{X}$ is now the six-column matrix

$$\begin{bmatrix} 1 & x_1 & y_1 & x_1y_1 & x_1^2 & y_1^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_i & y_i & x_iy_i & x_i^2 & y_i^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & y_n & x_ny_n & x_n^2 & y_n^2 \end{bmatrix} \tag{10.11}$$

so that the term $(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ will be a 6 by 6 matrix and the inversion will be considerably more complicated, and certainly not something to attempt by hand. Computer calculation of the $\boldsymbol{\beta}$ parameters is not particularly difficult. The addition of further terms produces more complex cubic, quartic, quintic, and so on surfaces, but in practice, these are seldom used, because difficulties arise in using many correlated independent variables. There is also a danger of *overfitting* the trend surface when the aim of the procedure is to generalize surface trends in the first place. Other types of surfaces, including oscillatory ones, may also be fitted (for reviews, see Davis, 2002, or Unwin, 1975b).

In the bad old days of user-unfriendly mainframe machines, many geoscientists spent time and effort writing long, complex programs to calculate and map polynomial trend surfaces. Nowadays, GISs such as *ArcGIS* and *IDRISI* have the capability built in. In fact, as long as you can move the output into a mapping program, it is easy to fit trend surfaces using any software that offers basic statistical analysis, such as *R*, *SPSS*, or *MINITAB*, and—provided that you have its data analysis routines loaded—even in *Microsoft Excel*.

Whatever the merits of trend surface analysis, it should be obvious that it is a relatively "dumb" technique:

- There will generally be no compelling reason to assume that the phenomenon of interest varies in such a simple way with the spatial coordinates, or even with some combination of the coordinates squared, cubed, and so on.
- Almost always, in practice, there will be spatial autocorrelation in the residuals. This will indicate that our model is misspecified, which implies that we can't reliably use the fit to make statistically significant interpretations of the results.
- Although the control point data are used to fit a chosen model for the trend by least squares multiple regression, other than simple visualization of the pattern they appear to display, the data are not used to help select this model.

In short, although it has definite merit as an exploratory technique and is much used by mathematically inclined geologists (see Davis, 2002, pp. 397–416 for a detailed review), the theoretical underpinnings of trend surface analysis are weak. Moreover, as we noted in Section 9.2 and illustrated in Figure 9.2, it provides one approach to continuous surface description of the sort used by locally valid analytical surface techniques and in surface gradient estimation.

It should be clear, however, that instead of specifying in advance the general type of surface shape, it would be useful to have some way of using the evidence of the observed control point data to inform our work. It turns out that our old friend *spatial autocorrelation* is the key to this approach, but instead of examining this in the context of area objects, we need to develop an approach suitable for a continuously varying field.

10.3. THE SQUARE ROOT DIFFERENCES CLOUD AND THE (SEMI-)VARIOGRAM

A natural way to characterize the spatial autocorrelation across a surface that we have sampled at a set of n control points is to plot the differences in height values for pairs of control points against their difference in distance. A plot that does precisely this is the *variogram cloud*. First, examine the data in Figure 10.3. These are control points (spot heights) gathered across a 310 × 310 foot survey area (see Davis, 2002, Figures 5.66 and 5.67 for

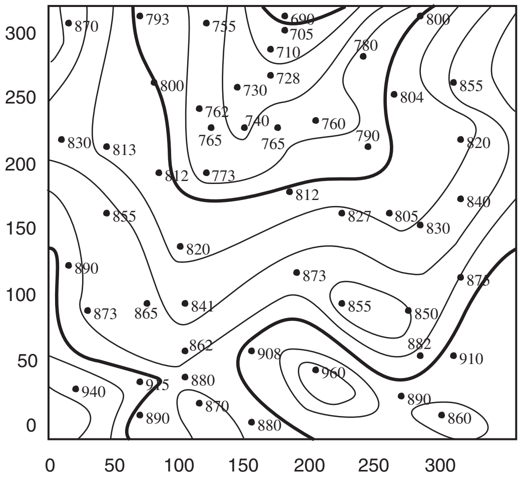


Figure 10.3 Spot heights and a contour pattern. Note that this contour pattern is for indication only and was done by hand.

Source: (Data from 2002)

details). A hand-drawn contour scheme has been added to give you some sense of an overall structure. There is a general upward slope from north to south, with some more confusing things happening in the southern part of the mapped area.

From these data, for each possible pair of points (there are 52 spot heights and therefore 1326 pairs of points), we plot the square root of the difference in their heights against the distance between them. This gives us the *square root differences cloud* (Cressie, 1993, p. 41) of points shown in Figure 10.4.

What can we tell from this figure? The reason for using the term *cloud* is obvious, but what it shows is that there is a tendency for larger differences in height to be observed the farther apart are two control point locations. This is a very "*messy*" trend, even including spot heights separated by as much as 300 feet that have no height difference.

Referring back to Figure 10.3, we can see that most of the upward trend in heights is from north to south. In fact, we can also make a plot of pairs of spot heights whose separation is almost north–south in orientation. In other words, we only plot pairs of points whose separation is close to exactly north–south in orientation. If we were to restrict the pairs used to *precisely* this directional separation, then we would probably have no pairs of points to plot. Instead, we allow separations at north–south plus or minus 5°. Similarly, we can plot pairs that lie almost exactly east–west of one another. Both of these sets of pairs are plotted in a single cloud in Figure 10.6. Pairs of points almost on an N–S axis are indicated by open circles, and pairs almost on an E–W axis are indicated by filled circles.

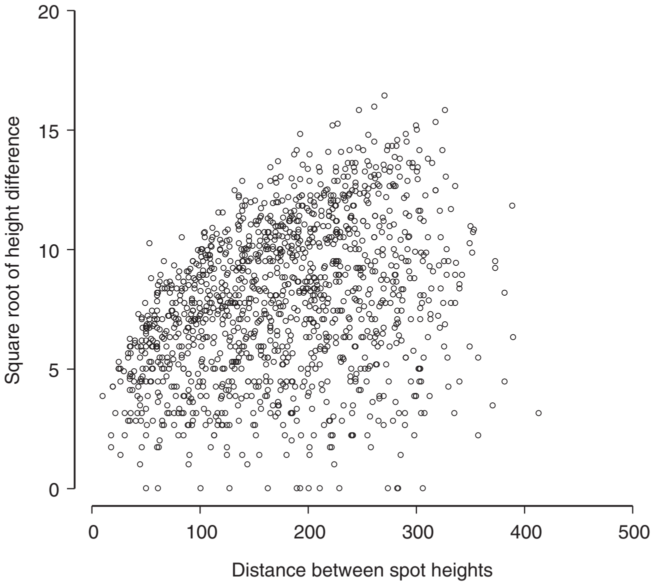


Figure 10.4 The square root differences cloud for the spot height data in Figure 10.3.

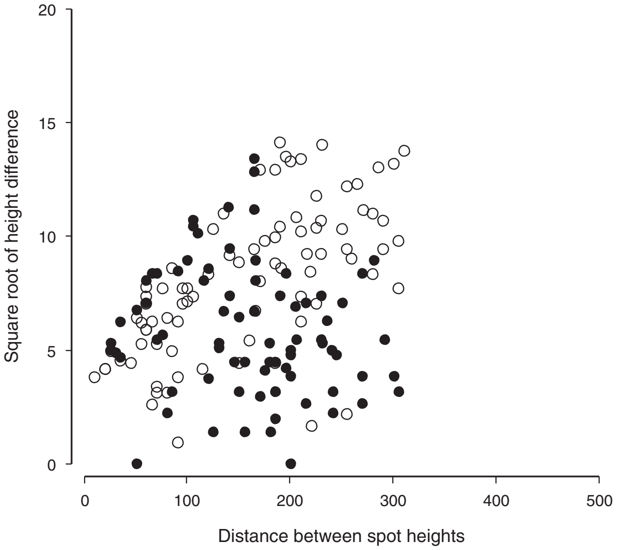


Figure 10.5 Square root difference clouds for N–S-oriented pairs in Figure 10.4 (open circles) and for E–W-oriented pairs (filled circles).

There are several things to note about this diagram:

- Far fewer points are plotted. This is because pairs at NS±5° or at EW±5° are much less numerous. In fact, we would expect only 10/180 = 1/18th as many points in each group as in the full plot of Figure 10.4.
- The distance range of the plot is shorter because of the allowed orientations and the shape of the study area. Since the study area is about 300 ft in both the N–S and E–W directions, only pairs of points at these separations are available. Note that this is another example of an edge effect in spatial analysis.
- Although there is considerable overlap in the two clouds, it is evident that *in general*, there are greater differences between N–S separated pairs of spot heights than between E–W separated pairs. This is consistent with the overall trends in the data indicated by the contours in Figure 10.3.
- This difference is indicative of *anisotropy* in this data set, that is, there are directional effects in the spatial variation of the data.

Cloud plots can be useful exploratory tools, but they are often difficult to interpret, largely because there are so many points. A more condensed summary is provided by subdividing the distance axis into intervals, called *lags*, and constructing a summary of the points in each distance interval. For each lag, we calculate a measure of central tendency (mean or median) and summarize the variation around this mean using *box plots*, one for each lag. Figure 10.6 clearly shows a rising trend in the difference between (square root) control point heights at greater distances, with edge effects becoming clear in the dropoff in height differences beyond lags 6 and 7. These lags correspond to separations of around 300 ft. Lags 8, 9, and 10 are made up of differences between spot heights in diagonally opposite corners of the study region, and inspection of Figure 10.3 shows that these heights are typically similar to one another. A less restricted study region would probably not show this effect at these distances. What this all shows is, of course, that height differences tend to increase as the separation distance increases, and the farther apart two control points are, the greater is the likely difference in their surface heights. This is consistent with what we expect of earth surface relief and indeed for most surface entities, which show strong positive spatial autocorrelation at short distances, with a generally rapid fall in dependence as the separation increases.

The square root difference cloud is one of a family of possible plots that can be used to characterize a surface. Its cousin is the *semivariogram cloud*, a plot that is of immense practical and theoretical importance. In Figures 10.3–10.6 the square root of the height difference was used, but Figure 10.7

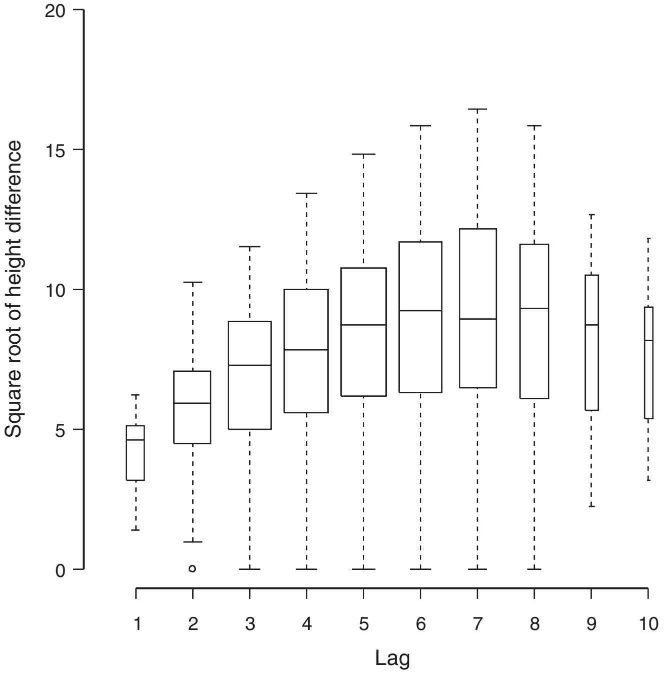


Figure 10.6 A series of box plots for distance intervals summarizing the data in Figure 10.4.

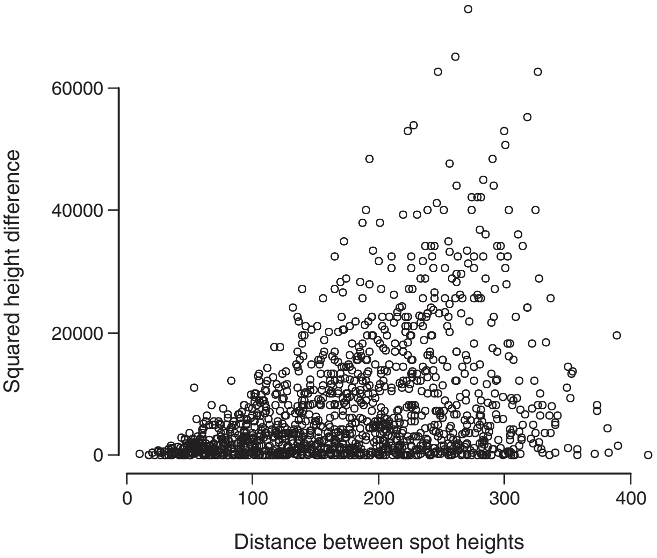


Figure 10.7 The semivariogram cloud for the data in Figures 10.3–10.6.

plots the *squared differences in height*, or *semivariances*, for these same data to give an example of a *semivariogram cloud*.

Thought for Sophisticates

Why, then, did we start by introducing the square root difference? This is primarily to help visualization of the overall shape of the plot and follows the advice given in Cressie (1993, pp. 41–42). It is clear that taking the square of the height differences tends to exaggerate extreme values, leading to badly skewed distributions in each bin that make it hard to gauge from the box plots whether or not a large value is a result of this skew or of some atypical observation. Ideally, what we would like in each lag is a nicely balanced box plot indicative of a symmetric distribution of values around their mean or median. The square root transformation helps achieve this.

But our description can be even more concise. We can summarize this semivariogram cloud, again using box plots in each of a series of lags across the entire distance range. This is shown in Figure 10.8 and provides an *estimate* of a *continuous* function called the *experimental semivariogram*. Where the context allows, you will often seen this term shortened to *variogram*. Losing the prefix *semi* is hardly important, but it pays to remember the

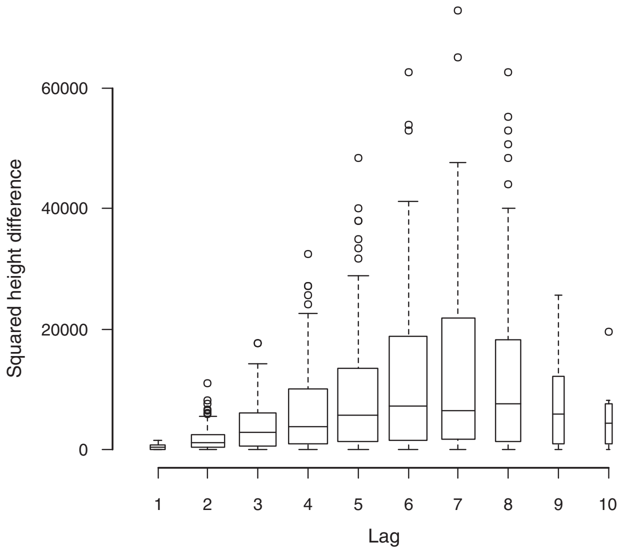


Figure 10.8 A series of box plots for distance intervals summarizing the semivariogram cloud of Figure 10.7.

word *experimental* and to keep in mind that it refers to a construct that is a property, not of the observed control points, but of the entire surface that they sample.

The equation for this estimation is

$$2\hat{\gamma}(d) = \frac{1}{n(d)} \sum_{d_{ii}=d} \left(z_i - z_j\right)^2 \qquad (10.12)$$

The right-hand side of this equation is straightforward, consisting of the sum of the squares of all the pairs of height control point values at a given distance d, divided by their number $n(d)$. In other words, it is simply their mean. The left-hand side of the equation uses the standard notation in which γ is the conventional symbol for the semivariogram. The "hat" tells us that we are dealing with an estimate at distance (d), and the 2 arises in the original development of this idea by the French geostatistician Georges Matheron (1963). This equation also makes it clear that there is a similarity between this measure and Geary's C measure of spatial autocorrelation (see Section 7.6). In essence, the (semi)variogram is an application of exactly the same idea to control point data, with the additional provision that we wish to estimate its value at a series of distances.

It should be noted that the estimation procedure implied by the above equation is not straightforward. In particular, for a given distance d, more likely than not, there will be *no* pair of observations at precisely that separation. Therefore, as for the variogram cloud box plots, we make estimates for distance bins (or *lags*) rather than continuously at all distances. Thus, the above equation should really be rewritten as

$$2\hat{\gamma}(d) = \frac{1}{n(d \pm \Delta/2)} \sum_{d \pm \Delta/2} \left(z_i - z_j\right)^2 \qquad (10.13)$$

indicating that the estimate is made over pairs of observations whose separations lie in the range $d - \Delta/2$ to $d + \Delta/2$. It is also important to note that the form of the equations shown here depends only on the distance between observations, effectively assuming that the underlying phenomenon is *isotropic*, with no directional effects of the type detected in the field shown in Figures 10.3 and 10.5.

10.4. A STATISTICAL APPROACH TO INTERPOLATION: KRIGING

In Chapter 9 we reviewed some simple mathematical methods of interpolation, particularly the much-used method of inverse distance weighting (IDW), where the height of any continuous surface, z_i, at location $\mathbf{s}_{\mathrm{i}}$ is estimated as a distance-weighted sum of the sample values in some surrounding neighborhood. The Achilles' heel of this approach is the arbitrariness in the choice of distance weighting function used and in the definition of the neighborhood. Although the choice of method may be based on expert knowledge, both are determined without any reference to the characteristics of the data being interpolated. By contrast, in trend surface analysis, in Section 10.2 we specify the general form of a function (usually a polynomial) and determine its exact form by using all the control point data to find the best fit according to a particular criterion (least squared error). In a sense, trend surface analysis lets the data "speak for themselves," whereas IDW interpolation forces a set structure onto them.

It would make sense to combine the two approaches in some way, at least conceptually, by using a distance weighting approach, but at the same time letting the sample data speak for themselves in the best way possible to inform the choice of function, weights, and neighborhood. *Kriging* is a *statistical* interpolation method that is *optimal* in the sense that it makes best use of what can be inferred about the spatial structure in the surface to be interpolated from an analysis of the control point data. It was developed in France in the 1960s by Georges Matheron as part of his *theory of regionalized variables*, which in turn was a development of methods used in the South African mining industry by Dani Krige. The basic theory is nowadays usually called *geostatistics*, and has been much developed from the original ideas and theories by numerous spatial statisticians. This section explains a little of how kriging works.

A Pronunciation Problem

How do you pronounce *kriging*? Many people pronounce it with the *i* as an ''ee'' sound, and a hard *g* (as in *golf*), but, in view of its derivation from a South African family name, perhaps it should sound more like *kric-king*.

The basis of interpolation by kriging is the distance weighting technique outlined in Chapter 9. Recall that for every location, $\boldsymbol{s}_i$, we estimated an interpolated value as a weighted sum of contributions from n neighboring data control points. The neighborhood over which this was done was set arbitrarily by changing the number of included points, while the rate of decay of influence with distance was changed by arbitrarily varying an inverse distance function. In essence, all that kriging does is to use the control point data as a sample to find optimum values for the weights of the data values included in the interpolation of each unknown location. Although often referred to by the one name, kriging, in fact there are several types of kriging that are alike in that they draw on regionalized variable theory, but that make different assumptions about the properties of the field being interpolated. Because it is the one most often used, we will illustrate the technique by examining *ordinary kriging*.

In order to interpolate in this way, three steps are involved:

1. Producing a description of the spatial variation in the sample control point data
2. Summarizing this spatial variation by a regular mathematical function
3. Using this model to determine interpolation weights

Step 1: Describing the Spatial Variation

We have already done this! Our estimate of the (semi)variogram provides all that we need to know about the spatial variation in the field. Step 2 is a lot trickier.

Step 2: Summarizing the Spatial Variation by a Regular Mathematical Function

Once we have approximated the (semi)variances by mean values at a series of lags, the next step is to summarize the experimental variogram using a mathematical function. The experimental variogram is an *estimate*, based on the known sample of control points of a *continuous* function that describes the way the variance of the height of the field changes with distance. Often, for reasons of mathematical convenience, the semivariogram is *fitted* to match a *particular* functional form that has appropriate mathematical properties. The task of finding the underlying function is illustrated in Figure 10.9.

No matter how well it might fit the (semi)variogram data, we cannot use any function that comes to hand for this purpose. There are a number of properties that candidate functions must have to be what is called *authorized*. Of these, the most important is that, because it is essentially modeling variances at distances and any variance must be a positive number, the function cannot give negative values. Also, by definition, the (semi)variance at the origin where $d = 0$ should be zero. In practice, it often happens that any line through the experimental values intercepts the (semi)variance axis

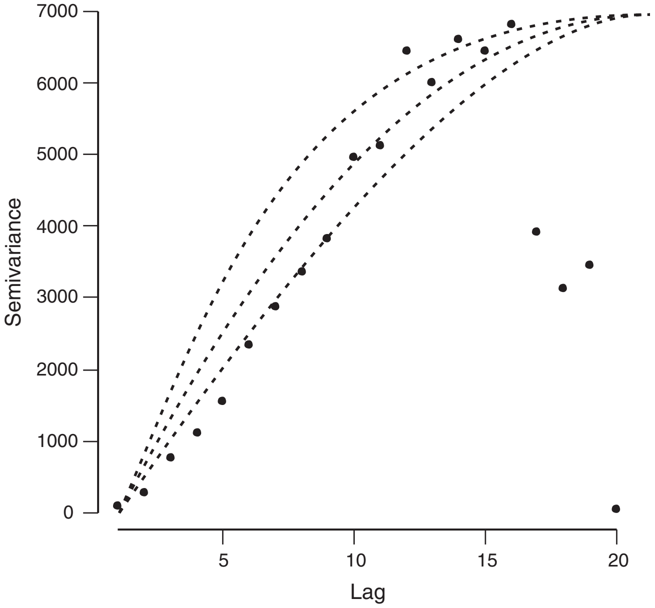


Figure 10.9 Estimating a continuous function that models the semivariogram. The data points are the empirical or ''experimental'' estimates in each of 20 distance bands. Dashed lines indicate possible smooth mathematical functions that might be fitted to these data. Note that the low values at lags 17 through 20 have been ignored.

at some positive value. This implies some discontinuity, and in the gold mining industry, where kriging was first developed and used, it was natural to identify the phenomenon with small lumps of gold dispersed throughout the rock body and call it the *nugget effect*. More generally, the nugget variance can often be ascribed to errors of measurement and to spatial variation that lies below the shortest sampling interval in the data.

Figure 10.10 shows a selection of possible functions that might be fitted. Figure 10.10(i) shows an example of an *unbounded* model where the semivariance, $\gamma(d)$, increases without limit. This may at first seem unlikely, but there are circumstances in which it can be appropriate. In the example the rate of increase is linear, with a distance exponent set at unity, but this could equally be according to some power of the distance. Note also that for generality we have included a *nugget* effect, with the line intersecting the semivariance axis at some positive value usually denoted as $\gamma(d) = c_0$. Panel (ii) shows a similar model, but in this case the increase in semivariance reaches a maximum value, at which $\gamma(d) = c_0 + c_1$, where the value $c_0 + c_1$ is called the *sill*. The distance, d, at which this happens is called the *range*, denoted a. The range is effectively the extent of any neighborhood we need to consider around each location in the field. It is the

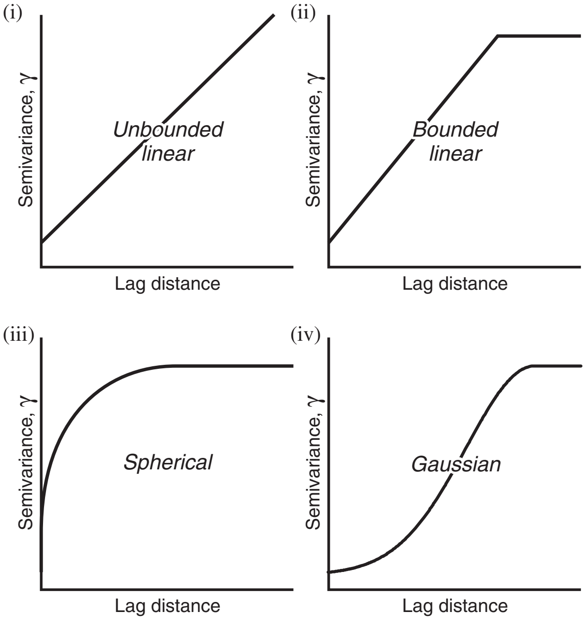


Figure 10.10 Some examples of the distance decay in dependence for four possible semivariance models.

distance at which the semivariogram levels off and beyond which the semivariance is constant. Beyond the range, pairs of points might just as well be selected at any separation. If the range in a data set is (say) 250 m, this means that it would be impossible to tell if a pair of observations was taken from locations 250 m or 25 km apart. There is no particular spatial structure in the data beyond the range. For obvious reasons, this is called a *bounded linear* model. There are many possible authorized models and hybrids with features of more than one model that could be fitted to any real-world experimental data, and fitting itself can be a tricky operation, so in practice, many workers use a model that is sufficiently robust to allow for many possible semivariogram shapes. Panel (iii) shows one such shape that is implemented virtually as the default in many GISs. This is the *spherical* model. Finally, panel (iv) shows a so-called *Gaussian model*, in effect using half of the familiar normal curve to describe the change in semivariance with distance. Notice that in contrast to the spherical model, this model has the property of leveling off as it approaches $d = 0$.

These models can all be described mathematically. For example, the spherical model starts at a nonzero variance ($\gamma_0 = c_0$) for the nugget and rises as in an elliptical arc to a maximum value, the sill, at some distance, the range, a. The value at the sill should be equal to the variance, σ^2, of the function. This model has the mathematical form

$$\gamma(d) = c_0 + c_1\left(\frac{3d}{2a} - \frac{1}{2}\left(\frac{d}{a}\right)^3\right) \tag{10.14}$$

for variation up to the range, a, and then

$$\gamma(d) = c_0 + c_1$$

beyond it.

An Example

How good is a spherical model for the experimental semivariogram shown in Figure 10.9? One possible fitted spherical model is shown in Figure 10.11.

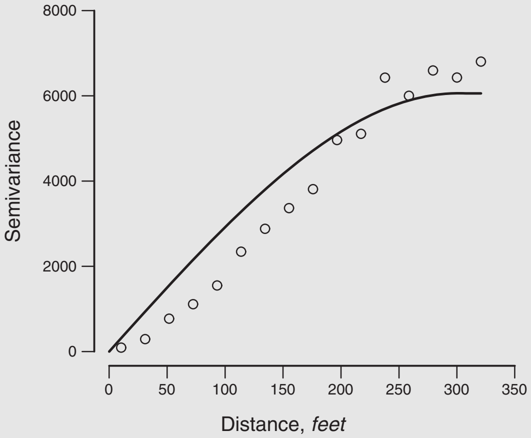


Figure 10.11 A spherical model fitted to the data of Figure 10.9.

The fit here is by no means perfect, and in fact, for the shape of the data, perhaps a Gaussian model would be a better starting point.

Clearly, a fitted variogram model can only be an approximation to the spatial variation in a real data set. In spite of its limitations, it remains a powerful summary of the overall properties of a spatial data set. A good idea of its summary power is provided by Figure 10.12. On the left-hand side of the diagram is a series of surface profiles with steadily increasing local

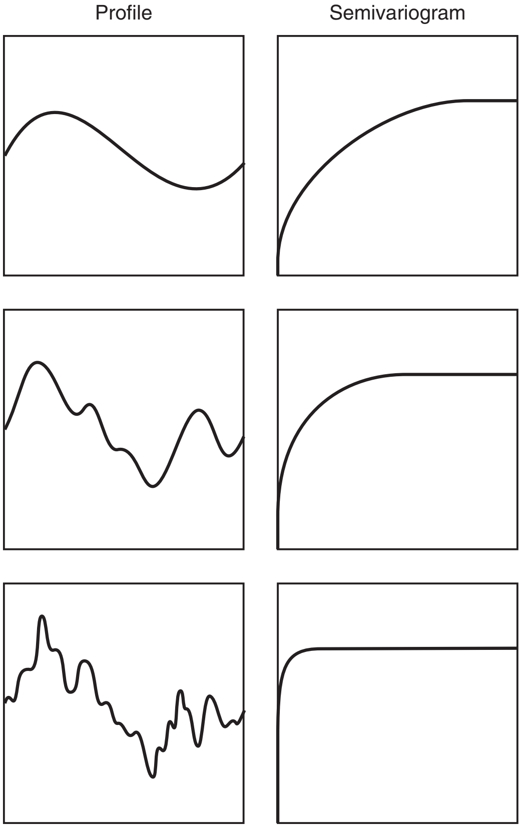


Figure 10.12 Typical spatial profiles and their associated semivariograms. All plots are on the same scales.

variation in the attribute values. On the right-hand side are the corresponding semivariogram models that we might expect. As local variation in the surface increases, the range decreases and the nugget value increases. Because the overall variation in the data values is similar in all cases, the sill is similar in all three semivariograms. The most distinct effect in these plots is the way that the semivariogram range captures the degree to which variation in the data is spatially localized.

According to two authorities, "choosing [semivariogram] models and fitting them to data remain among the most controversial topics in geostatistics" (Webster and Oliver, 2007, p. 127). The difficulties arise because:

- The reliability of calculated semivariances varies with the number of point pairs used in their estimation. Unfortunately, this often means that estimates are more reliable in the middle distance range rather than at short distances or long ones (why?), and the least reliable estimates are thus the most important for reliable estimation of the nugget, range, and sill.

- Spatial variation may be anisotropic, favoring change in a particular direction. In fact, based on the findings in Section 10.3 and Figure 10.5, we should really consider using an anisotropic model for these data. It is possible to fit a semivariogram that is a function of the separation *vector*, rather than simple distance, although this complicates matters considerably.
- Everything so far assumes that there is no systematic spatial change in the mean surface height, a phenomenon known as *drift*. When drift is present, the estimated semivariances will not simply be due to random variation but will be contaminated by a systematic amount. We explore this further in the next boxed section.
- The experimental semivariogram can fluctuate greatly from point to point. Sometimes, for example, there is no steady increase in the variance with distance, as is implied by most models.
- Many of the preferred functional forms for the variogram are nonlinear and cannot be estimated easily using standard regression software.

Modeling the Semivariogram: A Cautionary Tale

Figure 10.13 shows the variogram cloud for the Alberta temperature data given in Table 10.1.

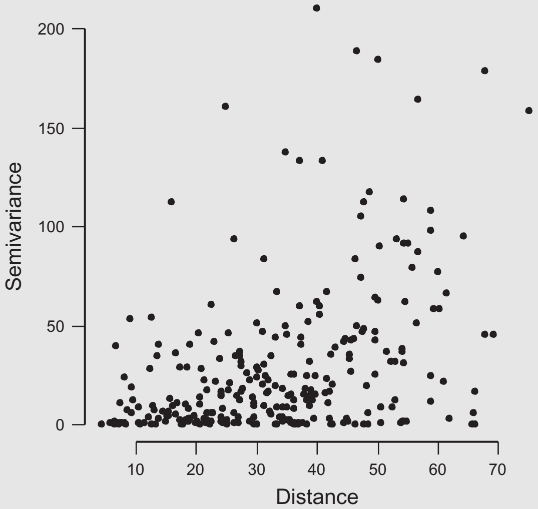


Figure 10.13 The variogram cloud for the Alberta temperature data in Table 10.1.

Since there are 24 data points, the plot contains 276 points and, as before, we can summarize it by using a series of lagged distance bands and plotting the mean value in each band. Figure 10.14 shows the result.

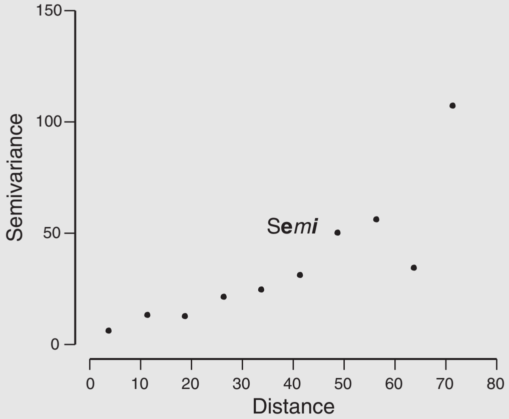


Figure 10.14 A semivariogram for the Alberta temperature data.

This plot is fairly typical of the kind of result that real-world field data generate. Would it be appropriate to model it using a spherical model? If not, why not?

In fact, we already know why we should not attempt to use this experimental semivariogram at all. As we saw in Section 10.2, there is a strong trend in the mean value of these data that can be well described by a linear trend surface.

In other words, there is evidence of drift in the mean height of the field, which means that the assumption made in ordinary kriging that the unobserved mean of the field is constant is very unlikely to be true. A concave upward form of the experimental semivariogram is often indicative of drift. The correct way to interpolate these data would be to subtract this drift in the mean and then create and model the experimental semivariogram of the residuals from the trend surface. This is done in *universal kriging.*

How best to estimate the experimental semivariogram and how to choose an appropriate mathematical function to model it is to some extent a “black art” that calls for careful analysis informed by good knowledge of the variable being analyzed. It is definitely not something that you should leave to be decided as a default in a simple-minded computer program, although this is precisely what some GIS systems attempt to do. Many experienced workers fit models by eye, others use standard but complex numerical approaches, and still others proceed by trial and error.

Step 3: Using the Model to Determine Interpolation Weights by Ordinary Kriging

Now that we have seen how the spatial structure of a data set can be described using the semivariogram function, how can this information be used to improve the estimation of continuous data from sampled locations? It is important to remember from the outset that kriging is just another form of interpolation by a weighted sum of local values in which we aim to find the best combination of weights for each unsampled location, based on its spatial relationship to the control points and on the relationships between the control points as summarized in the semivariogram. In addition to the assumptions that a surface has a constant but unknown mean with no underlying trend and that it is isotropic, reliance on the model we have fitted to the experimental semivariogram also means that we assume the following:

- The semivariogram is a simple mathematical function with some clearly defined properties.
- The same semivariogram applies over the entire area, and all other variation is assumed to be a function of distance. This is effectively an assumption about stationarity in the field, but instead of assuming that the variance is everywhere the same, we assume that it depends solely on the distance. For reasons that you may appreciate from Chapters 4, 5, and 6 on point patterns, this is sometimes called *second-order stationarity*, and the hypothesis about the variation that it contains is called the *intrinsic hypothesis*.

We wish to estimate a value for every unsampled location $\mathbf{s}$ using a weighted sum of the z values from surrounding control points, that is,

$$\hat{z}_{\mathbf{s}} = w_1 z_1 + w_2 z_2 + \ldots + w_n z_n = \sum_{i=1}^{n} w_i z_i = \mathbf{w}^{\mathrm{T}}\mathbf{z} \tag{10.15}$$

where $w_1 \ \ldots \ w_n$ is a set of weights applied to sampled values in order to arrive at the estimated value. To show how simple kriging computes the weights, we will use the very simple "map" shown in Figure 10.15. Although much of what follows is normally hidden from view inside a computer program, we think it is instructive to work through the detail of the calculation.

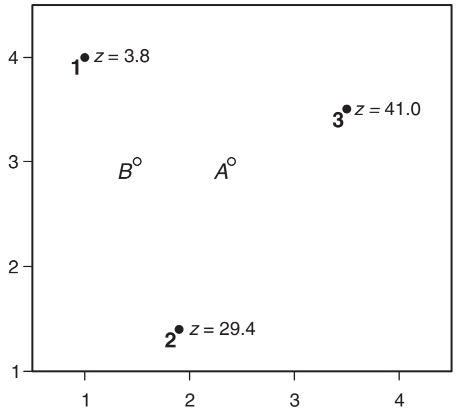


Figure 10.15 Data used for the ordinary kriging example. The open circles are the two locations whose values are be estimated using a weighted sum of just three surrounding control points shown along with their *z* values as filled circles, numbered 1 to 3.

Interpolation by Hand?

At this point, you might like to see if you can thread isolines at *z* = 30 and *z* = 40 height units through these data. What value might be expected at the locations to be estimated?

It can be shown (see, for example, Webster and Oliver, 2007, p. 152) that estimation error is minimized if the following system of linear equations is solved for the vector of unknown weights, **w**, and a quantity we introduce called a *Lagrangian multiplier*, denoted λ;

$$\begin{array}{ccccccccc} w_1\gamma(d_{11}) & + & w_2\gamma(d_{12}) & +\ldots+ & w_n\gamma(d_{1n}) & + & \lambda & = & \gamma(d_{1p}) \\ \vdots & & \vdots & & \vdots & & \vdots & = & \vdots \\ w_1\gamma(d_{n1}) & + & w_2\gamma(d_{n2}) & +\ldots+ & w_n\gamma(d_{nn}) & + & \lambda & = & \gamma(d_{np}) \\ w_1 & + & w_2 & +\ldots+ & w_n & + & 0 & = & 1 \end{array} \tag{10.16}$$

where n is the number of data points used, each of the terms $\gamma(d)$ is the semivariance for the distance between the relevant pairs of points, and the last equation is a constraint such that the weights sum to 1. This is necessary to ensure that the kriging estimates do not have any systematic bias. This equation is much easier to represent in matrix form as the system

$$\begin{bmatrix} \gamma(d_{11}) & \gamma(d_{12}) & \cdots & \gamma(d_{1n}) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(d_{n1}) & \gamma(d_{n2}) & \cdots & \gamma(d_{nn}) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \times \begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \lambda \end{bmatrix} = \begin{bmatrix} \gamma(d_{1p}) \\ \vdots \\ \gamma(d_{np}) \\ 1 \end{bmatrix} \tag{10.17}$$

which gives a standard system of linear equations

$$\mathbf{A} \cdot \mathbf{w} = \mathbf{b} \tag{10.18}$$

that can be solved in the usual way by premultiplying both sides by the inverse, $\mathbf{A}^{-1}$, to get the required weights:

$$\mathbf{w} = \mathbf{A}^{-1} \cdot \mathbf{b} \tag{10.19}$$

There are some similarities here to least squares regression. However, instead of observed data values, in this case the entries in the matrices are based on the calculated values of a fitted semivariance function according to the distances between the data control points. Given a semivariance function for the surface we are interpolating, all required values can be determined and the set of weights calculated.

We will work through the estimation for location A in Figure 10.15 using just the three indicated control points. This is a very simple, almost trivial, example, but it will give some indication of what is involved.

In our example with $n = 3$ control points there are four simultaneous equations, three for each of the data points to be used and one to constrain the weights to sum to unity. To solve this system of equations for an unknown point, we must assemble the matrices **A** and **b,** invert **A** to find $\mathbf{A}^{-1}$, and then solve for the weights and the Lagrangian in the vector **w**. The key quantities in both **A** and **b** are the semivariances given by our chosen model for the semivariogram, calculated at the distances between the relevant control points. Normally, the semivariogram would be estimated from the data as in Section 10.3, but for simplicity's sake, we will use a very simple unbounded linear model in which the semivariance increases by 60 variance units for every unit increase in distance and there is no nugget, that is:

$$\gamma(d) = 0 + 60(d) \tag{10.20}$$

The data matrix for the three control points and two locations to be estimated is

s	x	y	z
1	1.0	4.0	3.8
2	1.9	1.4	29.4
3	3.5	3.5	41.0
A	2.4	3.0	?
B	1.5	3.0	?

(10.21)

The matrix **A** can thus be assembled starting with the matrix of distances between the control points:

$$\mathbf{D} = \begin{bmatrix} 0 & & \\ 2.75 & 0 & \\ 2.55 & 2.64 & 0 \end{bmatrix} \tag{10.22}$$

Since this matrix and many of those that follow are symmetrical, we have only listed the lower triangle, and for display only, we have rounded our results. This means that you may see some discrepancies in the results if you work through this example using the values shown.

For matrix **A**, each element is replaced by the semivariance at the appropriate distance calculated from our hypothetical model and then augmented by a row and a column. For example, the distance between sample points 1 and 2 is 2.75 coordinate units, so

$$\gamma\left(d_{1,2} = 2.75\right) = 0 + 60(2.75) = 165 \tag{10.23}$$

This yields

$$\mathbf{A} = \begin{bmatrix} 0 & & & 1 \\ 165.08 & 0 & & 1 \\ 152.97 & 158.40 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \tag{10.24}$$

Similarly, the column vector **b** is assembled using the distances from the unknown point A whose value is to be estimated, to the three control points:

$$\mathbf{d} = \begin{bmatrix} 1.72 \\ 1.68 \\ 1.21 \end{bmatrix} \tag{10.25}$$

With the same model for the semivariogram and augmented by the extra row, this gives us values for the matrix **b**:

$$\mathbf{b} = \begin{bmatrix} 103.23 \\ 100.58 \\ 72.50 \\ 1 \end{bmatrix} \tag{10.26}$$

The required inverse of **A** is (once again, take our word for it, and once again, showing just the lower triangle of a symmetric matrix) is

$$\mathbf{A}^{-1} = \begin{bmatrix} -0.004 & & & \\ 0.002 & -0.004 & & \\ 0.002 & 0.002 & -0.004 & \\ 0.335 & 0.345 & 0.320 & -105.931 \end{bmatrix} \tag{10.27}$$

Finally, premultiplying **b** by this inverse gives the weights vector as

$$\mathbf{w} = \begin{bmatrix} 0.2603 \\ 0.2912 \\ 0.4485 \\ -13.445 \end{bmatrix} \tag{10.28}$$

Note that the three weights sum to 1.0, as they should (ignore the last value in **w**—it is the Lagrangian). Just as in IDW interpolation, it is the nearest points that have the largest weight, with the nearest point (3) having the most influence and the most distant one (1) having the least. We must now calculate the weighted sum of the control point values from

$$\begin{aligned} \hat{z}_s &= \sum_{i=1}^{n} w_i z_i \\ &= 0.2603(3.8) + 0.2912(29.4) + 0.4485(41.0) = 27.94 \end{aligned} \tag{10.29}$$

Notice that this is for a *single* unknown point on the field. For every other point where we want to make an estimate, we have to go through the final steps of the process again—calculating a new **b** matrix each time by measuring the distances, computing the semivariances, and computing the required sum.

Interpolating Another Point

It is easy to illustrate this. If we wish to estimate the value at location B using the same three control points, all that we need to redo are the last three steps starting from the augmented vector of distances between this new point and the three controls:

$$\mathbf{b} = \begin{bmatrix} 60 \times 1.118 \\ 60 \times 1.649 \\ 60 \times 2.062 \\ 1 \end{bmatrix} = \begin{bmatrix} 67.08 \\ 98.96 \\ 123.69 \\ 1 \end{bmatrix}$$

$$\mathbf{w} = \mathbf{A}^{-1} \cdot \begin{bmatrix} 67.08 \\ 98.96 \\ 123.69 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.5244 \\ 0.3359 \\ 0.1397 \\ -9.739 \end{bmatrix}$$

Since no new control point has entered the neighborhood, the inverse of **A** remains the same as before. The final estimate is

$$\begin{aligned} \hat{z}_s &= \sum_{i=1}^{n} w_i z_i \\ &= 0.5244(3.8) + 0.3359(29.4) + 0.1397(41.0) = 17.60 \end{aligned}$$

Notice how this result reflects the fact that B is closer to point 1 than A, so that it, rather than point 3, gets the higher weight.

Perhaps you can see that one way to compute estimates for a whole map would be to create a huge **A** matrix with the semivariances for all possible distance pairs, solve for the inverse $\mathbf{A}^{-1}$, and then use this at every one of the locations to be interpolated. This isn't done for two reasons. First, such very large matrices are often difficult to invert. Second, what we find is that distant points have weights so close to zero that they contribute negligibly to the final summations. Making use of this fact, many systems compute a "rolling" estimate using just control points that are close to the location to be estimated and recalculating the changing **A** matrices as they go along.

This is only a small example, but you should be able to see several important points:

- Kriging is computationally intensive. There will usually be many more samples, so that inversion of a considerably larger matrix than the 4 × 4 example described here is required.
- You need a suitably programmed computer. Although some GIS systems offer semivariogram estimation, modeling, and kriging, most serious workers in the field continue to use specialist software

such as GSLIB (Deutsch and Journel, 1992), Variowin (Pannatier, 1995), and GS+ (see www.geostatistics.com). Webster and Oliver (2007) give a relatively accessible account.

- As with most things statistical, it helps if you have a lot of data (large *n*). The more control points you have, the more distance pairs there are to enable the semivariogram to be estimated and modeled.

Experimenting with Kriging

It is useful to experiment with the various options available in ordinary kriging, as set by the model that we assume for the experimental semivariogram and the spatial configuration of the control point data. A simple computer-assisted learning tool to do this is E(Z)-Kriging (Walvoort, 2004).

All kriging results depend on the model fitted to the estimated semivariogram from the sample data and the validity of any assumptions made about height variation of the field. As we have seen, estimation of a semivariogram function is not simple and includes some more or less arbitrary decisions (how many distance bands? what distance intervals? what basic model to fit? what values to take for the sill and the nugget?). Different choices lead to different interpolated fields, and it is often the case that we will have no reason to favor one over another. Fortunately, “even a fairly crudely determined set of weights can give excellent results when applied to data” (Chilès and Delfiner, 1999, p. 175; see also Cressie, 1991, pp. 289–299). This observation leads to an important question about kriging. Simply put, noting that we are likely to get a reasonable single location estimate from most distance decay weighting functions that sum to unity and decline with distance, what’s so special about kriging that justifies the additional computational and modeling effort?

Some idea of the value of the approach may be obtained from Figure 10.16, which compares two interpolated surfaces produced for Davis’s (2002) set of elevations that we first presented in Figure 10.3. The surface on the left was produced using inverse distance-weighted interpolation with an inverse power law distance decay with $k = 2$. This surface shows some of the characteristic unrealistic “bulls-eye” effects that one soon comes to associate with IDW interpolation. The surface on the right side of the figure was produced by ordinary kriging, with a circular variogram fitted to the data. Although the estimates produced at any particular location do not differ much between these two maps, it is clear that the kriged surface appears much more realistic and that the method has some ability to adjust the

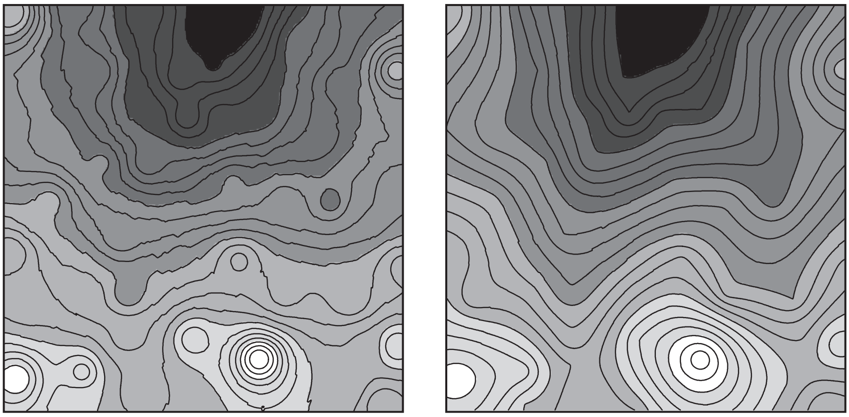


Figure 10.16 Interpolated surfaces for the data shown in Figure 10.3 (data from Davis, 2002). The left-hand surface is an IDW interpolation, while that on the right was produced by ordinary kriging.

spatial structure of estimates to reflect local variations in the surface structure. The additional subtlety of this interpolation was achieved without any of the potential improvements that might be produced by including an underlying trend surface (as in universal kriging) or by incorporating anisotropy, as can be done via semivariogram modeling.

From a theoretical perspective, there are two reasons to prefer kriging to simpler methods. First, if the correct model is used, the methods used in kriging have an advantage over other interpolation procedures in that the estimated values have *minimum error* associated with them. This is why the method is sometimes called *optimum interpolation.* Second, this error is quantifiable. For every interpolated point an *estimation variance* can be calculated, which depends solely on the semivariogram model, the spatial pattern of the points, and the calculated weights. The estimation variance is given by the weighted sum of the semivariances of the distances from the control points to the location of the estimate. Returning to our estimates in Figure 10.15, the estimation variance for location A is simply the sum of the semivariances of the distances, each weighted by the appropriate kriging weight:

$$\sigma_P^2 = \sum_{i=1}^{i=n+1} w_i b_i \tag{10.30}$$

Thus, for location A, the estimation variance is

$$\sigma_A^2 = 0.2603(103.23) + 0.2912(100.58) + 0.4485(72.50) - 13.445(1) = 75.23 \tag{10.31}$$

So, to estimate a 95% confidence interval around the original estimate, we need to take the square root to get the standard error and then set limits at 1.96 times this above and below the estimate. Since $\sqrt{75.23} = 8.67$ and $1.96 \times 8.67 = 17$, we discover that at this level the estimate could be anywhere between 11 and 45. Given the starting data, this is hardly surprising. However, the ability to compute an estimation variance for every location means that not only does kriging produce estimates that are in some sense optimal, it also enables valuable maps to be drawn of the likely error in these estimates. These can be used, for example, to decide where the most benefit might be obtained by measuring additional locations in the field.

Other Members of the Kriging Family

There are other forms of kriging that ask slightly different questions about the surface under study, make different assumptions about it, or recognize differences in the nature of the data used. The method we have outlined and illustrated in detail is called o*rdinary kriging*. The success of the underlying theory of regionalized variables, especially in practical geologic exploration and prospecting, has led to a whole family of extensions. Texts such as Cressie (1993), Chilès and Delfiner (1999), Goovearts (1997), Isaaks and Srivastava (1989), and Webster and Oliver (2007) cover this ground, and there is at least one Web site devoted to geostatistics at www.ai-geostats.org. All that we can do here is to point you in the right direction by outlining what each member of the family does and how it might be used:

- *Simple kriging* assumes that the mean height of the field is known and doesn't vary spatially in a systematic way due to *drift*. Any variation in height depends only on distance, not on direction, so the field is assumed to be *isotropic.* The assumption about the mean may appear restrictive, but there are situations, notably when it can be assumed to be zero (for example, using z scores or residuals from a regression), that arise from time to time. In many circumstances this method differs little from ordinary kriging.
- *Universal kriging* retains the isotropic assumption but corrects for regional drift in the mean. At still more advanced levels, methods have been developed to allow for kriging of anisotropic fields. In practice, universal kriging models drift by some form of *trend surface* and then compute the semivariogram using residual values from that surface.
- *Block kriging* is an approach that has a great deal of utility in mining engineering. Instead of estimating the value for specific point locations, it predicts the average value of the spatially continuous variable over a defined area or block. In mining, where blocks of

rock are to be extracted, this is a useful approach. The modifications required to ordinary kriging aren't as great as might be expected. We simply form the vector **b** using the average semivariances between the block concerned and the control point locations and proceed in the usual way.

- *Indicator kriging* is a nonlinear and nonparametric approach used where the field variable is a binary 0/1 *indicator* of the presence or absence of some phenomenon. This enables many types of data, including some nonnumeric qualitative attributes, to be accommodated.
- *Disjunctive kriging* is related to indicator kriging, but instead of predicting a field of binary assignments, it estimates the *probability* of the field exceeding some specified threshold value. This is precisely the sort of information that might be required in spatial decision making applications of GIS and is an approach that perhaps should have more prominence than it has achieved to date.
- Finally, although it does not complete the kriging family roll call, we have *co-kriging*, which extends the analysis to two or more variables considered at the same time. This is most useful where two variables are known to be spatially associated—for example, assays of the content of two different minerals in rock samples. The information contained in the associated variable is used to enable better estimations (as measured by a decreased estimation variance) of the other variable.

10.5. CONCLUSIONS

This chapter has covered a lot of difficult and detailed material. When studying the statistical approach to field data, it pays to take time and to ensure that each stage of development of the techniques is familiar before moving forward into the unknown. We hope that, after reading this chapter, you will have achieved at least some of the learning objectives with which we started.

In any real analysis, the major decision you face is *which* of the techniques presented in this chapter and the previous one should be used. If you have data that contain a lot of error and are only interested in making rough generalizations about the shape of the field, then trend surface analysis seems appropriate. If, on the other hand, you want to create contour maps that honor all your data but do not need estimates of interpolation error, then some form of IDW will almost certainly be adequate. However, if you want the same properties but also some indication of the possible errors, then kriging is the approach to take. However, as the previous few paragraphs indicate, if you use it, you must clearly appreciate the properties of the surfaces that you are studying.

Perhaps the most important general point that emerges is that, just as we can apply statistical models and logic to point and area data, we can apply them to spatially continuous fields to improve substantially the estimates we make of unknown values of the field, relative to "guesstimation" derived from a simple mean of all the data points. The fundamental reason for this capability is that phenomena do not usually vary randomly across space, but tend to exhibit characteristic spatial structures or autocorrelation effects. Whether we represent the autocorrelation of our data with a rule of thumb like IDW or attempt the more involved process of estimating and fitting a variogram, ultimately we are hoping to take advantage of this fundamental fact.

CHAPTER REVIEW

- The approaches to surface analysis discussed in Chapter 9 are all deterministic and do not involve any statistical theory, whereas both of the approaches described in this chapter appeal to and make use of statistical theory.
- Standard multiple linear regression can be applied to spatial data using the approach known as *trend surface analysis*, in which the independent variables are the spatial coordinates of the observations. This is a useful exploratory technique.
- It is useful to use a *square root variogram cloud,* which plots the square root of height differences against separation distance for all pairs of control point observations. At some cost in introducing an artifact arising from the bin classification used, this can be summarized by a series of measures in which these numbers are assigned to bins at increasing *lag distances.*
- A second plot that shows the spatial structure of a set of observations is the *experimental semivariogram*. The semivariogram function $\gamma(d)$ is based on the squared (rather than square root) differences between pairs of observations at distance d apart.
- Semivariograms are estimated starting from a *variogram cloud* that records squared differences between observed values and their separation distance as a scatterplot.
- Semivariograms are summarized using standard continuous mathematical functions that model the variance as a function of the separation distance. Often this will be nonzero even at the origin, giving a *nugget* variance, and will increase to a limit, called the *sill*, at some distance referred to as the *range* beyond which there is no spatial dependence.

- The *spherical model* is often used and is capable of describing many experimental semivariograms.
- So-called *ordinary kriging* uses the modeled semivariogram to determine appropriate weights, based on observed sample values, for an estimate of unknown values of a continuous surface. These estimates are statistically optimal, but the method is computationally intensive.
- There is a whole family of methods based on kriging (simple, universal, block, indicator, disjunctive, and co-kriging)
- If you are serious about spatial analysis using continuous field data, then you must have a working knowledge of the field of *geostatistics.* This is a difficult and technical subject, but it might at least make you think twice before uncritically using the functions in your favorite GIS because they seem to work. There's a good chance that they don't!

REFERENCES

Chilès, J-P. and Delfiner, P. (1999) *Geostatistics: Modeling Spatial Uncertainty* (New York: Wiley).

Cressie, N. (1993) *Statistics for Spatial Data* (New York: Wiley).

Davis, J. C. (2002) *Statistics and Data Analysis in Geology (*Hoboken, NJ: Wiley).

Deutsch, C. V. and Journel, A. G. (1992) *GSLIB Geostatistical Software Library and User's Guide* (New York: Oxford University Press).

Goovaerts, P. (1997) *Geostatistics for Natural Resource Evaluation* (Oxford and New York: Oxford University Press).

Isaaks, E. H. and Srivastava, R. M. (1989) *An Introduction to Applied Geostatistics* (New York: Oxford University Press).

Matheron, G. (1963) Principles of geostatistics. *Economic Geology*, 58: 1246–1266.

Pannatier, Y. (1995) *Variowin: Software for Spatial Analysis in 2D* (New York: Springer-Verlag).

Unwin, D. J. (1975a) Numerical error in a familiar technique: a case study of polynomial trend surface analysis. *Geographical Analysis*, 7: 197–203.

Unwin, D. J. (1975b). An introduction to trend surface analysis. *Concepts and Techniques in Modern Geography*, 5, 40 pages (Norwich, England: Geo Books). Available at http://www.qmrg.org.uk/catmog.

Walvoort, D. J. J. (2004) *E(Z)-Kriging: Exploring the World of Ordinary Kriging* (Wageningen: Wageningen University & Research Centre). Available at http://www.ai-geostats.org/index.php?id=114.

Webster, R. and Oliver, M. (2007). *Geostatistics for Environmental Scientists*, 2nd ed. (Chichester, England: Wiley).

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 11

Putting Maps Together—Map Overlay

CHAPTER OBJECTIVES

In this chapter, we:

- Point out that *polygon overlay,* the most popular map overlay method is but one of at least 10 possible ways by which geographic objects might be overlaid
- Illustrate the basics of *sieve mapping* using Boolean yes/no logic
- Underline the importance of ensuring that the data used are fit for the intended purpose, including greater than usual concern for ensuring that the inputs are correctly *coregistered* onto the same coordinate system
- Examine some of the typical issues that arise in *Boolean overlay*
- Develop a general theory of map overlay based on the idea of *favorability functions* and outline some possible approaches to their calibration

On reading this chapter, you should be able to:

- Understand and formalize the GIS operation called *map overlay* using *Boolean* logic
- Understand why *coregistration* of any maps used is critical to the success of any map overlay operation
- Give examples of studies that have used this approach
- Outline how overlay is implemented in vector and raster GIS
- Appreciate how sensitive overlay is to error in the input data modeling strategy adopted and the algorithm used
- List reasons why such a simple approach as Boolean overlay can be unsatisfactory
- Outline and illustrate alternative approaches to the same problem

- Describe how these methods find utility in the techniques of multicriteria decision making and provide some examples from geographic information analysis

11.1. INTRODUCTION

In this chapter we examine the very popular geographic analytical method known as *map overlay*, which is shorthand for methods of combining information from different map layers. The techniques introduced so far have almost all been ones that are applied to single maps made up of points, areas, or fields. Yet, one of the most important features of any GIS is its ability to combine spatial data sets (or maps produced from them) to create new maps that incorporate information from a variety of sources. Generically, this process has been given the name *map overlay* and is just the GIS version of an old technique known as *sieve mapping*, which was used by land use planners to identify areas suitable or unsuitable for some activity. In this approach, the entire study area is considered potentially suitable; then areas are disqualified on the basis of a series of criteria until all that remains are the areas still considered suitable. Before the advent of GIS systems and digital spatial information, planners used a light table and a series of transparent map overlays, one for each criterion, on which areas deemed unsuitable were blacked out. When all these binary maps were stacked on top of each other, light would shine only through those areas deemed suitable—hence left unshaded—on all the overlaid maps. Although this technique was first formalized by a landscape planner (McHarg, 1969), the idea is as old as the hills, and in retrospect, it has been used in many analyses.

In a GIS environment many types of map overlay are possible, and it is possible to argue that overlay in one form or another is involved in much geographic information analysis. As Table 11.1 shows, there are at least 10 general ways in which we can combine the different types of geographic objects in an overlay.

Table 11.1 Possible Types of Map Overlay: A Geometric View

	Points	*Lines*	*Areas*	*Fields*
Points	Point/point			
Lines	Line/point	Line/line		
Areas	Area/line	Area/line	Area/area	
Fields	Field/point	Field/line	Field/area	Field/field

Thought Exercise

Either by finding case studies in the literature or by thinking it through from first principles, how many of these 10 overlays can you illustrate? You should be able to think of useful example of every one.

Each of these types of operation presents different algorithmic and analytical problems, some with their own distinct names, but the most common type of analysis is a map overlay where we take one map of planar enforced area objects and determine its intersection with another. For obvious reasons, this "areas-over-areas" process is called *polygon overlay* and, quite apart from its use in geographic information analysis, it finds application in numerous GIS functions. For example, polygon overlay is necessary in *areal interpolation*, where we have the populations for one set of areas and want to use these to estimate the populations of a different set of overlapping areas. Apportioning population from the first set of polygons to polygons in the second set according to the areas of overlap between the two can solve this problem. This is a simple, albeit rough, way to estimate the population of areas from statistics for another set of areas. In ecological *gap analysis*, use is often made of similar techniques to estimate the areas where specified plant, animal, or bird species are likely to be found. Each input map describes the environmental conditions favored by the species, and these maps are overlaid in some way to identify those areas seen to be favorable on all the criteria (see Franklin, 1995). The basic low-level GIS operations called *windowing* and *buffering* both also involve overlay of polygons.

Two Overlay Examples

It is useful to examine briefly two simple case studies demonstrating polygon overlay. Keep these in mind as we progress through this chapter.

Landslides in Gansu, China

Catastrophic mass movements are a major environmental hazard on the loess (or wind-blown dust) plateau of China, causing loss of life and severe damage to infrastructure and farmland. Although the causes of any specific landslide are dynamic and transient, related to changes in water content or

(*continues*)

(*box continued*)

the incidence of earthquakes, the longer-term stability of loess landforms is largely determined by static factors related to landform slope shapes and geology. An obvious GIS approach to this problem, taken by Wang and Unwin (1992), is to identify landscape factors thought to be significant in the occurrence of landslides and to create binary maps, one for each factor, where areas coded ''1'' are thought to be susceptible to landslides due to that factor and those coded ''0'' are thought to be safe. All the maps can then be combined by an overlay operation to identify areas that are susceptible (or not) based on all factors. In fact, these authors use just three input maps:

- Slope steepness, estimated from a digital elevation matrix. Slopes greater than 30° were thought to be at most risk.
- Slope aspect, derived from the same source, with slopes with a northern aspect at more risk.
- Rock type, derived from a geologic map, with slopes developed on the unstable loess materials at most risk.

Overlay of these three map layers was used to produce a map of the areas thought to be at most risk. The analysis was conducted in a raster GIS environment. Numerous subsequent studies have used GISs to address the same problem and are summarized in Lee and Choi (2004).

Nuclear Waste Dumps

Openshaw et al. (1989) present a similar example in the social sciences. They use an overlay strategy to identify areas in the United Kingdom suitable for disposal of hazardous nuclear waste. The criteria they used came from the industry itself:

- Areas with few people (fewer than 490/km^2)
- Areas with good railroad access (less than 3 km from a line)
- Areas not already designated as conservation areas

Their analysis was conducted in a vector GIS environment, but the result again was a map showing where, according to the three input criteria, nuclear waste might be dumped. The useful surprise was that, even in a country as densely populated as the United Kingdom, the number of sites that meet these criteria is much higher than had been assumed. This does not, of course, mean that all the identified areas would be even remotely reasonable as dumping grounds, but the analysis narrows down the solution space. Indeed, the demonstration that a very large number of sites were identified might even be taken as evidence that the criteria used were themselves at fault!

11.2. BOOLEAN MAP OVERLAY AND SIEVE MAPPING

Although the areas of application are very different, the two studies sketched out above use essentially the same analytical strategy involving:

- Map overlay of sets of areas on top of each other.
- Successive disqualification of areas on the basis of each criterion until those remaining are found to be susceptible on all criteria. Because of the yes-no nature of this approach, it is called *Boolean* after the mathematician who developed *binary* (true/false) logic.

Together these two parts of the analysis produce a *Boolean overlay*.

Figure 11.1 illustrates the process. Here we have two categorical maps to be overlayed. Map A has the rock types "limestone" and "granite," and Map B has the land uses "arable" and "woodland." Overlay produces Map A & B, with four possible *unique conditions* given by the combinations granite/arable, granite/woodland, limestone/woodland, and limestone/arable. If the intention is to find those areas with the unique condition limestone/woodland, then this overlay would be a Boolean *sieve mapping* selection, and the result would be the heavily outlined area shown.

In Boolean overlay in a raster environment it is worth noting that, although the operation is based on logic, it can also be performed using simple multiplication. Table 11.2 shows how this works by examining every possible combination of the unique conditions for the overlay of Figure 11.1 and coding "limestone" = 1, "granite" = 0, on Map A and "woodland" = 1,

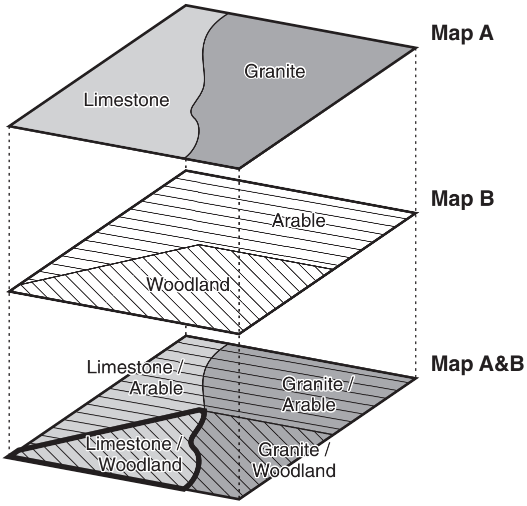


Figure 11.1 Schematic illustration of map overlay.

Table 11.2 Boolean Overlay as Multiplication*

Unique Condition	*Map A*	*Map B*	*Map A&B*
Granite/arable	0	0	0
Granite/woodland	0	1	0
Limestone/arable	1	0	0
Limestone/woodland	1	1	1

*This is the logical AND operation on the layers in Figure 11.1, where limstone/woodland is the suitable combination of interest.

"arable" = 0 on Map B. It can be seen that only the unique condition coded "1" on both input maps ends up coded "1" on the output map. In short, the yes/no selection is the result of a multiplication of all the 0/1 values on the input criterion maps. We will return to this observation in Section 11.3.

Getting the Data

Decisions on what to include in an overlay analysis are almost always made with one eye on what data are available. In an ideal world, we would have digital data to the same standards of accuracy and precision for each desired input layer, and these data would all be georeferenced to the same coordinate system. In practice, this is rare, and inputs are often scanned or digitized maps, originally compiled to widely varying standards of accuracy, with different locational precision and georeferenced to different coordinate systems. It cannot be emphasized too strongly that if insufficient care is taken, using such data can be a recipe for disaster. If the outputs from overlay analysis are to be of reasonable quality, it is essential that the input information is consistent in its accuracy and precision. There are at least four traps awaiting the unwary:

1. Failure to realize that with digital map data, accuracy and precision are themselves often a function of the map scale. This amounts to a requirement that data sources are all at more or less the same map scale or from recording devices that have similar accuracy and precision. Further, the input scales should be consistent with the scale required for the result. Overlaying, say, data from a 1:10,000 map of woodland areas on a geologic map at 1:250,000 does not result in locational precision in the output map equivalent to that of the 1:10,000 data, yet one often sees results of complex overlays that seem to forget this simple truth.
2. In map overlay, the digitized data that purport to relate to the same object, such as the boundary of an area, may come from different

sources. This means that the data contain more than one digital approximation to the same objects, and these may be different. The well-known result in map overlay in vector-based systems is the creation of "sliver" polygons where the lines do not coincide. Similar situations may also occur with differently resampled raster data.

3. The false belief that everything on a map is accurately located. Maps are drawn with a view to the communication of information and were never intended to be a source of data for a GIS. The data they display have been generalized, often resulting in displacement of the outlines of objects relative to their true positions to avoid incomprehensible clutter. Alternatively many geographic entities are represented by symbols, rather than by their true outline on the ground. Perhaps the most obvious example of both practices is that roads are shown on small-scale maps with lines whose widths are much greater than their real widths on the ground.
4. Finally, even more care is required when the input maps are themselves the results of data manipulations, such as an interpolated field variable like a digital elevation matrix or, worse, some derivative from it, such as a slope map.

All of this is not to say that mixing data sources in these ways is completely nonsensical; indeed, the spatial integration of diverse data sets is a major motivation for the use of GIS. However, it is important to point out that the results of overlay analysis must be interpreted with these issues in mind and with a suitably questioning attitude.

Getting Data into the Same Coordinate System

A look at Figure 11.1 should convince you that map overlay is possible only if all the input data have coordinates that are registered accurately to the same locational coordinate system. We might, for example, have some data from a GPS georeferenced to the WGS84 system that we wish to combine with data georeferenced to a State Plane Coordinate System, to latitude/longitude, or to another projection-based coordinate system such as the Universal Transverse Mercator. For an overlay to make sense, the inputs must all refer to the same parts of the Earth's surface, making it necessary to *coregister* them all to the same system.

Figure 11.2 shows what is involved. Here we have the same object, a river, on two maps, A and B, and the objective is to bring the data on Map B into the system used on Map A. To do this, a *grid-on-grid transformation* is necessary to do three things:

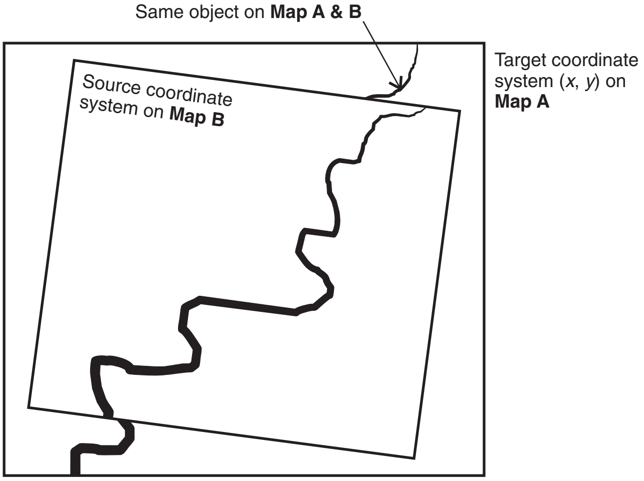

Same object on **Map A & B**
Target coordinate system (*x*, *y*) on **Map A**
Source coordinate system on **Map B**

Figure 11.2 The coregistration problem in map overlay. For an overlay analysis to be accurate, it is essential that all the overlaid maps be registered to the same coordinate system.

- Move the origin of the coordinates used in Map B to the same point in Map A. This is called a *translation* of the origin.
- Change the scale on both x- and y-axes. The locational coordinates in B might be in units of 0.1 mm away from the origin, whereas in A they might be on some nonmetric scale. This is a *scaling* of the axes.
- Because, as illustrated, Map B's coordinates may well be in a system where x and y are not parallel to the same axes in Map A, coordinates may need to be rotated to correct for this. This is a *rotation* of the axes.

The coregistration problem arises frequently when transferring data from a semiautomatic digitizer or scanner to a GIS system, and transforming one set of coordinates into another is often required when integrating data from several sources into a GIS. For this reason, this operation is part of the standard GIS toolkit. The method often uses "tick marks" on each of the map frames to guide the system in creating the correct coregistration using an *affine transformation* that will translate the map origin and scale and rotate the axes to bring the two map coordinate systems into alignment. The mathematical details of this operation can be found in, among others sources, Maling (1973) and Harvey (2008). The necessary information can be acquired in one of at least three ways:

1. From knowledge of the source and target systems to develop the appropriate transformation matrix, an example being a transfer from known latitude/longitude coordinates to a projection-based

system such as the Universal Transverse Mercator. The affine transformation between known coordinate systems is mathematically defined for many common map projections.

2. By recording at least three known points on one of the maps—for example, the southwest, southeast, and northeast corners—and their equivalent values in the target coordinates and using these to define the required transformation. These points are often referred to as *tick points*.
3. An approach commonly used in operational GIS uses ordinary multiple least squares regression relating the (x', y') coordinates in a target coordinate system to those in the original system. Given the availability of GPS, the original coordinates (x, y), of a set of so-called *ground control points* (GCPs) might well be observed directly in the field or read off from a map at as high a precision and accuracy as feasible at that map scale. The same locations are then recorded in the target system, and multiple regression is used to estimate the best-fitting transformation constants. An affine transformation can be reduced to two equations in which the two sets of coordinates are related as follows:

$$\begin{aligned} x' &= t_x + r_{11}x + r_{12}y \\ y' &= t_y + r_{21}x + r_{22}y \end{aligned} \tag{11.1}$$

Here the intercept values t_x and t_y are associated with the translation between the two coordinate systems, while any rotation and scaling components are described by the set of values r_{11}, r_{12}, r_{21}, and r_{22}. These equations are identical to the equations of multiple linear regression and express the transformed coordinates x' and y' as linear functions of the original coordinates x and y. Thus, we can rewrite them using the standard notation as

$$\begin{aligned} x' &= \alpha_0 + \alpha_1 x + \alpha_2 y \\ y' &= \beta_0 + \beta_1 x + \beta_2 y \end{aligned} \tag{11.2}$$

You should be able to see that the vectors of regression constants α and β are estimates of the required parameters. This approach is used in almost every GIS and allows use of more points than the simple tick point approach. Mather (1995), Unwin and Mather (1998), and Morad et al., (1996) demonstrate that the quality of the estimated transformation depends on the number of ground control points used (the more the better) and their spatial distribution (even coverage is preferred). The locational accuracy and precision of the ground control points are also important, so well-defined locations, such as large-angle road intersections or the corners of fields, should be favored, and coordinates should be recorded as precisely as possible.

An advantage of the regression approach is that it can be extended to cover transformations that *warp* one set of coordinates into another by a nonlinear transformation, as may be required when source maps are based on unknown projections. Nonlinear transformations may be estimated by including additional higher-order terms in x^2, y^2, xy, and so on, in the regression, exactly as in trend surface analysis (Section 10.2). For example:

$$\begin{aligned} x' &= \alpha_0 + \alpha_1 x + \alpha_2 y + \alpha_3 x^2 + \alpha_4 y^2 + \alpha_5 xy \\ y' &= \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy \end{aligned} \tag{11.3}$$

Should this give an inadequate transformation, even higher-order terms can be added, and this procedure is implemented in most proprietary GIS software. Note here that increasingly better statistical fits to the data, as might be reported by the GIS, do not necessarily imply better coregistrations.

Overlaying the Maps

Once we have correctly coregistered the input maps, it is possible to overlay them. In a raster GIS environment, this is straight forward. All that need be done is to work over the entire map, pixel by pixel, testing whether or not the various criteria have been met. For the moment, note that when sieve mapping in a raster environment, simple arithmetic multiplication of the 0/1 unsuitable/suitable values suffices to produce values in the output map.

The same operation can be performed in a vector environment, although it is a trickier business. In a vector environment, the software must:

- Create a new map of polygon objects by finding all the intersection areas of the original sets of polygons
- Create attributes for each new polygon by concatenating attributes of polygons in the original maps whose intersection formed it
- Reestablish the topological relationships between the new polygons and ensure that the map remains planar enforced
- Identify which new polygons have the desired set of attributes

You will appreciate that the geometry involved in potentially thousands of polygon intersection operations is far from trivial, so a fast, efficient, and accurate polygon overlay algorithm is a *sine qua non* for any vector GIS.

It Makes Me Cross

Since it involves computing whether or not every line segment in the data intersects with every other possible segment, even the first step in a vector overlay is tricky. In one of the most famous papers on GIS ever written, David Douglas (1974) describes how he offered $100 to some hot-shot student programmers to produce a routine that would do this with typical GIS data. Nobody won the prize because, although the basic mathematics of an algorithm is easy, there are numerous special cases for which any software has to cater. Douglas's paper was titled ''It Makes Me Cross.'' This illustrates a well-known computer programmer's dictum that 99% of the code is written to handle the exceptions. Computational geometry (see de Berg et al., 1997) produces many such exceptional cases.

Logical Problems in Boolean Overlay

Despite its popularity in GIS analysis, sieve mapping using Boolean overlay presents a number of difficulties and can almost always be improved upon, often using exactly the same data. Problems arise mainly from simplistic assumptions about the data and the implied relationships between the attributes. The consequences of these assumptions for error in the final maps deserve more detailed attention. For example:

1. *It is assumed that the relationships really are Boolean:* This assumption is usually not only scientifically absurd, it frequently also involves throwing away a great deal of metric information. In the two simple case studies with which we began this chapter, there is nothing particularly important about using a 30° value to represent slopes above which landslides are deemed to be possible or a population density of less than 490/km^2 to represent suitable areas for nuclear waste disposal. It is ridiculous to score slopes of 29° as "without risk" and those at 31° as "at risk." The two-valued (yes/no) nature of the logic in sieve mapping produces abrupt spatial discontinuities that do not adequately reflect the continuous nature of at least some of the controlling factors.
2. *It is assumed that any interval- or ratio-scaled attributes are known without significant measurement error:* In the simple limestone/ granite, woodland/arable example this issue doesn't arise, but if the observations had been on an interval- or ratio-scaled variable,

then error of some magnitude is certain to be present. There is a particular problem if the data used are derived in some way. An example would be slope angles derived by estimation from a digital elevation matrix that, in turn, was estimated from either spot heights or contour data. Both the interpolation process used and the estimation of slope introduce error, and whether or not this error is significant when included an overlay operation is usually an open question.

3. *It is assumed that any categorical attribute data are known exactly:* Examples might be in an overlay using categories such as rock or soil type that are products of either a classification (for example, satellite-derived land use) or an interpretative mapping (as in a geologic or soil survey). In both of these cases, it is likely that the category to which each land parcel is assigned is a generalization and that there are also inclusions of land with other properties.
4. *It is assumed that the boundaries of the discrete objects represented in the data are certain and are recorded without any error:* In our example, we assume that the boundaries of the attributes shown are exact. Yet, either as a result of real uncertainties in locating gradual transitions in interpreted mapping or due to errors introduced in digitizing a caricature of these boundaries from paper maps, this is almost never the case. The boundaries of the mapped units may themselves be highly uncertain. The archetypal example of a map made up of "fuzzy" objects with indeterminate boundaries is a soils map (Burrough, 1993; Burrough and Frank, 1996). If you are using a raster data structure, then it is important to note that this automatically introduces similar boundary errors into the data.

Accounts of these problems of error and generalization in GIS, and some of the strategies that can be used to minimize them or at least understand their impact on derived results, are found in Veregin (1989), Heuvelink and Burrough (1993), and Unwin (1995, 1996).

11.3. A GENERAL MODEL FOR ALTERNATIVES TO BOOLEAN OVERLAY

Fortunately, there are many methods other than simple Boolean overlay that we can use in map overlay. Many authors have recognized two basic approaches, characterized as either *knowledge* or *data driven*. In the *knowledge-driven* approach, we use the ideas and experience of experts in the field to determine which criteria to use. In the *data-driven* approach, we use any available data to suggest which criteria should be used. In practice, most map overlay studies use some combination of the two, and the distinction is not always clear-cut. First, we will examine the knowledge-driven approach.

In this section, we develop a more general model for the map overlay process based on the idea of a *favorability function*. Although the formal detail is our own, the idea itself is based on work by the geologist Graeme Bonham-Carter (1995). In this interpretation, a Boolean overlay evaluates the favorability/suitability of parcels of land for some activity or process, such as landslides or nuclear waste dumping. Moreover, as Table 11.2 shows, what in mapping terms is a Boolean overlay operation can also be regarded as the evaluation of a simple mathematical function at every location on the input map. This function can be written

$$F(\mathbf{s}) = \prod_{M=1}^{m} X_M(\mathbf{s}) \tag{11.4}$$

where $F(\mathbf{s})$ is the *favorability* evaluated as a 0/1 binary at each location $\mathbf{s}$ in the study area, and $X_M(\mathbf{s})$ is the value at $\mathbf{s}$ in input map M coded "1" to indicate if the cell is "favorable" and 0 if it is not according to the factor recorded on map M. The Greek capital letter pi (Π) indicates that these input values should be multiplied together—Π indicating that the product is required in the same way that the capital sigma (Σ) indicates the sum of a series of terms. It may be easiest to think of the set of locations $\mathbf{s}$ as pixels in a grid, but this is not necessary, although, as we have seen, the practical implementation is more involved for areal unit maps. In its use of the single symbols F and X to indicate whole maps, this is also an example of Tomlin's map algebra in action (Tomlin, 1990; see also Section 9.5). Notice too that in its selection of the input maps and any critical thresholds, this is also a wholly knowledge-based approach to map overlay.

This is a very limited approach to the problem, and we can improve it in several ways:

- By evaluating the favorability, F, on a more graduated scale of measurement such as an ordinal (low/medium/high risk) or even a ratio scale. An appropriate continuous scale might be a spatial probability scale, on which each pixel is given a value in the range 0 (absolutely unfavorable) to 1 (totally favorable).
- By coding each of the criteria used (the map's *X_M* values) on some ordinal or ratio scale.
- By weighting the criteria used to reflect knowledge or data about their relative importance. In a Boolean overlay all the inputs have the same

weight, but we often have some ideas about the relative importance of the criteria. In our favorability function, this is equivalent to inserting into the equation a weight, w_m, for each criterion.

- By using some mathematical function other than multiplication—for example, by *adding* the scores.

All of these extensions to basic overlay have been tried, and sometimes they have been developed into very sophisticated tools for making locational decisions based on the favorability of sites under various assumed criteria and weights. Some of the approaches have been used frequently enough to have acquired their own names, but we can generalize this discussion by arguing that Boolean overlay is a special case of a general *favorability function*

$$F = f(w_1X_1, w_1X_1 \ldots, w_mX_m) \tag{11.5}$$

In this equation, F is the output favorability, f represents "some function of," and each of the input maps, X_1 to X_m, is weighted by an appropriate weight, w_1 to w_m. Note that we have dropped the (**s**) notation here, but it is understood that evaluation of the function occurs at each location in the output map using values at the same location in the input maps. In the following sections, we discuss a number of alternatives to simple Boolean overlay that may be considered specific examples of this generalized function.

11.4. INDEXED OVERLAY AND WEIGHTED LINEAR COMBINATION

The simplest alternative is to reduce each map layer thought to be important to a single metric and then add up the scores to produce an overall index. This approach has been called an *indexed overlay*. In this approach, we add parcel values together to give a favorability "score" for each:

$$F = \sum_m X_m \tag{11.6}$$

Although we have changed the functional form from multiplication to addition, each input X_m remains as a binary map. The overall effect of this change is to transform F into an ordinally scaled variable, with $m + 1$ degrees of favorability from 0 (no risk/unfavorable) to m (high risk/very favorable).

In its basic form, this is another example of a knowledge-based approach. For example, in another study of the favorability of slopes for landslides, Gupta and Joshi (1990) used a variant of this method in the Ramganga catchment of the Lower Himalaya but assigned ordinal scales to each input map X. Based on the knowledge gained from an analysis of data on past landslides, the individual criteria used were assigned to classes on an ordinal scale of risk (low $= 0$, medium $= 1$, high $= 2$), and these classes were summed to give an overall risk measure. Three input criterion maps were used (*lithology*, *land use*, and *distance from major tectonic features*), so their final favorability F was an ordinal scale with a range from 0 to 6.

An advantage of this approach is that attaching an additional weight, w_m, to each of the input criteria can easily modify it, so that the favorability becomes

$$F = \sum w_m X_m \tag{11.7}$$

It is conventional to *normalize* such a summation by dividing by the sum of the individual weights to give a final favorability

$$F = \frac{\sum_m w_m X_m}{\sum_m w_m} \tag{11.8}$$

This is the classic approach known in the literature as *weighted linear combination* (Malczewski, 2000). Numerous methods have been used to determine the weights. In ecological *gap analysis*, they are often computed in a data-driven approach by comparing the observed incidence of the species on the particular habitat criterion to the numbers expected if that species had no special habitat preference. In *multicriteria evaluation* (MCE), they are often derived from expert knowledge and opinions but by a formal procedure.

Return to the two examples of map overlay in Section 11.1. The authors of these works overlaid their maps essentially using sieving in which each layer acts as a Boolean yes/no *constraint* on the result. Deterministic overlay of this sort is of limited use, and the authors of both works make this point strongly. Especially in a controversial case such as nuclear waste disposal, it is necessary to be able to handle *multiple conflicting criteria* and probably also *multiple conflicting objectives*.

In such cases, not all stakeholders in the process would agree on the objectives of the exercise, the necessary input layers, or the thresholds to be applied to establish any constraints. Similarly, if some form of weights is to be used, there would be disagreement on any or all of the weighting applied to each factor in comparison to the others, the scores assigned to the individual pixels or land parcels, and, in a mathematically literate example, perhaps also the nature of the arithmetic used in the combination. In studies where GIS is used to inform a decision, these arguments could be critical and this kind of operation is central to the use of *spatial decision support systems* (SDSS; Jankowski and Nyerges, 2001). Furthermore, if the objective is to allow public participation in the decision-making process, it is clear that mapping the outputs under different scenarios created by different assumptions is a major advance (Elwood, 2006; Sieber, 2006). The classic early paper on MCE is by Carver (1991), with more recent summaries by Eastman (1999) and Malczewski (1999). The next box outlines a simple example.

Using GIS to Suggest Sites for Wind Farms

Given the push to develop sources of renewable energy, in a number of recent studies GIS overlay was used to suggest potential sites for wind farms. Typically, such studies use overlay as their basic approach and use constraints/factors developed from some external source, such as published planning guidance.

An early study by Sparkes and Kidner (1996) used 19 binary constraints on wind farm in Wales, sieving out all those areas that were:

- Within 3 km of an airport
- Within 1 km of a National Park
- Within 1 km of a National Trust property
- Within 3 km of a military danger zone
- Within 1 km of a scenic area
- Within 1 km of a Forest Park
- Within 2 km of a built-up area
- Within 5 km of a city centroid
- Within 2.5 km of an urban centroid
- Within 1.5 km of a town centroid
- Within 1 km of a small town or village centroid
- Within 750 m of a small village, hamlet, or isolated settlement
- Within 250 m of a lake, marsh, or reservoir
- Within 300 m of a motorway, A road, or B road
- Within 250 m of a railway
- Within 200 m of a river or canal
- Within 250 m of a radio or TV mast
- Within 1 km of a picturesque or scenic feature
- Below 100 m in elevation

Anti–wind farm protest groups would, of course, dispute almost all of these constraints. How reasonable do they seem to you?

In a more recent but essentially similar study of the Baltic Sea region of Denmark, Hansen (2005) used four absolute binary constraints related to protected land and faunal habitats together with 23 weighted factors such as proximity to the coast, lakes, power lines, and so on. The data layers were selected after interviews with planning agencies and were integrated together using an approach based on fuzzy set theory (see Section 1.3), in which the study area was classified by the summation of the weighted factors. The resulting map shows that for northern Jutland at least, diminishing returns have set in and there is very little space for more turbines.

If you think you could devise a set of criteria and weights, Hydro Tasmania provides a simple interactive tool, ''Where would you build a wind farm?'' at http//www.hydro.com.au/education/discovery/GIS/windfarm.htm.

The point to understand is that there is no ''correct'' answer to this question, only a solution that is correct given the constraints, factors, and weights you decide to use. All that the geographic information processing system can do is make it easier to see the consequences of any particular set of inputs and thus perhaps reduce the size of the solution space.

It should be clear that the ability to devise weights both for map layers and for criteria within a specific layer that in some sense capture the general opinion would be extremely useful. One method that has been implemented in GIS (Eastman et al., 1995) makes use of results from Saaty's analytical hierarchy process (Saaty, 1977), in which $n \times n$ matrices of pairwise comparisons between n factors are summarized as a best-fit vector of weights by their first (the principal) eigenvector. There are also other approaches.

11.5. WEIGHTS OF EVIDENCE

Sometimes it is unnecessary to use external knowledge to inform the choice of weights. The *weights of evidence* method uses a knowledge-based approach to decide on the map layers to be included but then uses a data-driven approach to determine appropriate weights. Its basis is to use the available data to compute a *weight of evidence* and then use this to estimate the favorability, F, as a probability in the range 0 to 1. The key concept is a theorem due to the Reverend Thomas Bayes, known as *Bayes' Theorem*.

Suppose we have two *events* that are independent of each other—the classic example is flipping two unbiased coins. What is the probability of each possible pair of outcomes? These probabilities are called *joint probabilities*, denoted

$$P(A\,\&\,B) \tag{11.9}$$

If the events are truly independent, it is obvious that

$$P(A\,\&\,B) = P(A) \cdot P(B) \tag{11.10}$$

So, for two heads as our events, we have

$$P(H\,\&\,H) = P(H) \cdot P(H) = 0.5 \times 0.5 = 0.25 \tag{11.11}$$

To understand the weights of evidence approach, we must introduce a different probability linking two events. This is the *conditional probability* of an event *A given that the other event, B, is known to have occurred*. It is denoted

$$P(A : B) \tag{11.12}$$

and referred to as the probability of *A given B*. This will usually not be the same as the joint probability of *A and B* because the fact that *B* has already occurred provides additional evidence either to increase or reduce the chance of *A* occurring. In statistics, the theorem is used to guide how additional evidence should lead us to adjust our expectations. For example, consider the question of the probability of rain tomorrow (A), given that we know that it has rained today (B). Clearly, the fact that it is raining today is evidence we can use in our assessment of the hypothesis that it will rain tomorrow, and in most climates, meteorological persistence means that if it rains today, it is more rather than less likely to rain tomorrow.

Bayes' Theorem allows us to find $P(A{:}B)$. The basic building block that we need to prove the theorem is the obvious proposition that

$$P(A\,\&\,B) = P(A : B)P(B) \tag{11.13}$$

This may be obvious, but it is by no means self-evident: Study it carefully. In words, this equation states that the joint probability of two events is, indeed *must* be, the conditional probability of the first, given that the second has already occurred, $P(A{:}B)$, multiplied by the simple probability of the second event $P(B)$. Note that the multiplication on the right-hand side of this equation is justified only if we are prepared to assume that *A*:*B* and *B* are independent of each other. Exactly the same reasoning allows us to form the symmetrical expression

$$P(B\,\&\,A) = P(B : A)P(A) \tag{11.14}$$

Now, it must be the case that

$$P(A\,\&\,B) = P(B\,\&\,A) \tag{11.15}$$

so that

$$P(A : B)P(B) = P(B : A)P(A) \tag{11.16}$$

which leads to the theorem in the form in which it is usually stated:

$$P(A : B) = P(A)\frac{P(B : A)}{P(B)} \tag{11.17}$$

The term $P(A)$ is the probability of event *A* occurring, and the ratio $P(B{:}A)/P(B)$ is termed the *weight of evidence*. If this ratio is more than 1, it shows that the occurrence of *B* increases the probability of *A*; if the ratio is less than, 1 it reduces it.

In spatial work, it is usual to estimate the required probabilities using the proportions of the areas involved. Thus, for $P(B)$ we take the area over which the criterion *B* occurs as a proportion of the total area, and for $P(B{:}A)$ we take the proportion of the area of *A* that is also *B*. The required conditional probability $P(A{:}B)$ can then be calculated. For example, consider a 10,000-km^2 region in which 100 landslide events have been recorded over the previous 10 years. The probability of a landslide event per square kilometer is then 1 in 100, or 0.01—this is the baseline probability *P*(landslide). Now, say that of those 100 events, 75 occurred in regions whose slope was greater than 30°, but that only 1000 km^2 of the region has such slopes. The probability of a landslide, given that the slope is greater than 30°, is then 0.075. This is consistent with Equation (11.17), because we have

$$\begin{gathered} P(\text{landslide} : \text{slope} > 30^\circ) = P(\text{landslide})\frac{P(\text{slope} > 30^\circ : \text{landslide})}{P(\text{slope} > 30^\circ)} \\ 0.075 = 0.01 \times \frac{0.75}{0.1} \end{gathered} \tag{11.18}$$

It is easy to see that the weight of evidence associated with land slopes over 30° is 7.5, considerably greater than 1. If we assume independence between the slope factor and other factors for which we also have maps and can do similar calculations, then overlay can be based on the weights of evidence values to produce maps of the posterior probability of occurrence of landslides given the presence or absence of those factors at each location in the study region. This approach to map overlay has been much used in exploration geology and is illustrated and described in more detail in Bonham-Carter (1991) and Aspinall (1992). Lee and Choi (2004) provide an extended and clear example of the application of this approach to mapping landslide susceptibility in Korea.

11.6. MODEL-DRIVEN OVERLAY USING REGRESSION

A third alternative to simple Boolean overlay is to use regression techniques to calibrate a model linking the favorability to each of the criteria thought to be involved. This also combines data-driven and knowledge-based approaches, but essentially it relates back to the weighted linear combination version of the favorability function

$$F = \sum_m w_m X_m \tag{11.19}$$

and implements it as a standard multiple regression by adding an intercept constant, w_0, and an error term ε:

$$F = w_0 + \sum_m w_m X_m + \varepsilon \tag{11.20}$$

This model can be calibrated using real data to estimate values of w_0 through w_m that best fit the observed data according to the least squares criterion of goodness of fit. In the map overlay context, this requires that we have a sample of outcomes where we can associate some measure of favorability with combinations of values of the criterion variables, X_1 through X_m. In our landslide example, Jibson and Keefer (1989) provide an illustration of this approach in the context of predicting where landslides might occur. They use a sample of landslide incidents together with a series of factors related to each slide. Ordinary least squares regression enables them to say how important each factor is, and from this they derive weights for use in the production of a map of landslide favorability.

However, the ordinary least squares regression approach cannot easily be used in most map overlay exercises for three reasons:

1. The favorability, F, is rarely measured as a continuous, ratio-scaled variable, as the model demands. Instead, it is usually the binary (0/1) presence or absence of the phenomenon under study.
2. In many map overlay studies, the environmental factors involved are best considered as categorical assignments, such as the geology or soil type, rather than as continuous ratio-scaled numbers.
3. Any regression analysis makes assumptions about the error term that are very unlikely to be upheld in any practical application. In particular, our old friend spatial autocorrelation ensures that the regression residuals are unlikely to be independent.

Technically, some of these difficulties can be circumvented using *categorical data analysis*, where the dependent variable is restated as the "odds" (equivalent to the probability) of an occurrence and this is regressed on a set of probabilities of membership of each of the criteria (see Wrigley, 1985, for an accessible introduction to the method). It follows from the multiplication law of probability that these terms must be multiplied together, and this is achieved by formulating the model in terms of the logarithms of the odds. The result is a *log-linear* model. Estimation of such models is not simple; to date, two methods have been adopted.

Wang and Unwin (1992) used a categorical model to estimate the probability of a landslide for each of the unique conditions given by their overlay, calibrating a model of the form

$$P(\text{Landslide}) = f(\text{slope aspect, rock type, slope angle}) \tag{11.21}$$

where all the criterion variables on the right-hand side of the equation consisted of coded categories. Using *logistic regression*, the input layers may consist of both categorical and numeric data. Modeling variation across a region of the likelihood of deforestation, given the proximity of roads and other human land-use activity, is a more recent application of this method (see Apan and Peterson, 1998; Mertens and Lambin, 2000; Serneels and Lambin, 2001).

Finally, it is worth noting that many researchers using some of the techniques we have mentioned, especially model-driven approaches, might not characterize their work as overlay analysis at all. They are more likely to think of such work as spatial regression modeling of some sort, with sample data sets consisting of pixels across the study region. Nevertheless, eventually, a new map is produced from a set of input maps, so within the broad framework considered in this chapter, overlay analysis seems a reasonable description of what they are doing. This perspective draws attention to the issues of spatial accuracy we have discussed, which are easily overlooked in such work and which can dramatically affect the reliability of results. We would also expect you to be wondering by now where the autocorrelation problem has disappeared to in such work: surely the input data are not independent random samples? The answer is that autocorrelation has invariably *not* gone away, but it *is* routinely ignored.

More complex techniques that address the problem are available—*spatial autoregression* (Anselin, 1988) and *geographically weighted regression* (Fotheringham et al., 2000, 2002; see Section 8.5).

11.7. CONCLUSIONS

In this chapter, we have moved some way from the relatively well-defined analytical strategies used when analysis is confined to a single map or its digital equivalent and when the objective is to show that visually apparent map patterns really are worthy of further attention. Typically in overlay analysis, the objectives are less clear and the preferred analytical strategy is less clearly defined. Very often, too, the quality of the data used may be suspect. It follows that, when examining the results of an overlay analysis, it is sensible to pay close attention to the compatibility of the data used, their coregistration to the same coordinate system, and the way in which the favorability function in the output map was computed.

The issue here is not some absolute standard of accuracy and precision, but whether or not the ends justify the means. The ends in question are very often policy related: what should be done to abate the identified risks or to prepare for a change in areas where it is estimated to be likely? Indeed, estimating the probability of change—whatever its nature—*across space*, as we do with GIS, means that the output from such analysis is often important in determining the likely scale and scope of a problem. This means that overlay analysis can have a very significant impact on spatial decision making.

It is tempting to conclude that a sufficiently smart analyst could produce whatever output map suits the circumstances (and the requirements of whoever is paying for the analysis). The uncertainties we have mentioned and the range of options available to the analyst in approaching overlay certainly provide the flexibility required to arrive at any desired conclusion. Technically, the only way to address the uncertainty that this raises is to perform sensitivity analyses where the variability in the possible output maps is examined. Very often, it turns out that even the results obtained with poor data and basic Boolean methods provide the guidance required for appropriate responses, assuming, of course, that the attendant uncertainties are kept in mind at all times.

CHAPTER REVIEW

- Map overlay is a popular analytical strategy in GIS. Although we can think of at least 10 basic overlay forms, area on area overlay is by far the most common.
- Any overlay analysis involves four steps, all of which can be problematic: determining the inputs, getting compatible data, coregistering them on the same coordinate system, and performing the overlay itself.
- Coregistration is achieved by means of a translation of the origin, followed by rotation and scaling of the axes in an affine transformation. Typically in a GIS, this process involves regression using tick points on both sources.
- Usually, polygon overlay is used in a *Boolean* yes/no analysis that emulates a well-known technique from landscape planning called *sieve mapping*.
- Boolean overlay makes many frequently unjustified assumptions about the data and the relationship being modeled.
- Overlay can be classified as data or knowledge driven.
- Boolean overlay can often be replaced by alternatives that are more satisfactory, such as *indexed overlays, weighted linear combinations, weights of evidence,* and *model-based methods* using regression.
- In *spatial decision support systems,* there are formal ways in which weights can be established.
- These can all be seen as ways of calibrating an underlying *favorability function.*
- Finally, despite all the reservations outlined in the chapter, overlay analysis works in the sense that its results are often good enough, provided that the uncertainties we have discussed are kept in mind.

REFERENCES

Anselin, L. (1988) *Spatial Econometrics: Methods and Models* (Dordrecht, The Netherlands: Kluwer).

Apan, A. A. and Peterson, J. A. (1998) Probing tropical deforestation: the use of GIS and statistical analysis of georeferenced data. *Applied Geography*, 18(2): 137–152.

Aspinall, R. (1992) An inductive modelling procedure based on Bayes' theorem for analysis of pattern in spatial data. *International Journal of Geographical Information Systems*, 6(2): 105–121.

Bonham-Carter, G. F. (1991) Integration of geoscientific data using GIS. In: M. F. Goodchild, D. W. Rhind, and D. J. Maguire, eds., *Geographical Information Systems: Principles and Applications*, Vol. 2 (London: Longman), pp. 171–184.

Bonham-Carter, G. F. (1995) *Geographic Information Systems for Geosciences* (Oxford: Pergamon).

Burrough, P. (1993) Soil variability: a late 20th century view. *Soils & Fertilisers*, 56: 529–562.

Burrough, P. and Frank, A. U., eds. (1996) *Geographical Objects with Uncertain Boundaries* (London: Taylor & Francis).

Carver, S. J. (1991) Integrating multi-criteria evaluation with GIS. *International Journal of Geographical Information Systems*, 5(3): 321–339.

de Berg, M., van Kreveld, M., Overmars, M., and Schwarzkopf, O. 1997. *Computational Geometry: Algorithms and Applications* (Berlin and New York: Springer).

Douglas, D. (1974) It makes me so CROSS. Harvard University Laboratory for Computer Graphics and Spatial Analysis, Internal memorandum. Reprinted in: Peuquet, D. J., and D. F. Marble (1990) *Introductory Resources in Geographic Information Systems* (London, England: Taylor and Francis), pp. 303–307.

Eastman, J. R. (1999) Multi-criteria evaluation and GIS. In: P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, eds., *Geographical Information Systems, Volume: 1 Principles and Technical Issues* (Chichester, England: Wiley), pp. 493–502.

Eastman, J. R., Jin, W., Kyem, P. A., and Toledano, J. (1995) Raster procedures for for multi-criteria/multi-objective decisions. *Photogrammetric Engineering and Remote Sensing*, 61: 539–547.

Elwood, S. (2006) Critical issues in participatory GIS: deconstruction, reconstruction and new research directions. *Transactions in GIS*, 10: 693–708.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2000) *Quantitative Geography: Perspectives on Spatial Data Analysis* (London: Sage).

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002) *Geographically Weighted Regression* (Chichester, England: Wiley).

Franklin, J. (1995) Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*, 19(4): 474–499.

Gupta, R. P. and Joshi, B. C. (1990) Landslide hazard using the GIS approach—a case study from the Ramganga Catchment, Himalayas. *Engineering Geology*, 28: 119–145.

Hansen, H. S. (2005) GIS-based multi-criteria analysis of wind farm development. In: Hauska, H. and Tueite, H., eds: *ScanGIS 2005—Proceedings of the 10th Scandinavian Research Conference on Geographical Information Science* (Stockholm: Swedish Department of Planning and Environment), pp. 75–87 (available at also www.scangis.org/scangis2005/papers/hansen.pdf).

Harvey, F. (2008) *A Primer of GIS: Fundamental Geographic and Cartographic Concepts* (New York: Guilford Press).

Heuvelink, B. M. and Burrough, P. A. (1993) Error propagation in cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Systems*, 7(3): 231–246.

Jankowski, P. and Nyerges, T. (2001) *Geographic Information Systems for Group Decision Making* (London: Taylor & Francis).

Jibson, Randall W., and D. K. Keefer (1989) Statistical analysis of factors affecting landslide distribution in the new Madrid seismic zone, Tennessee and Kentucky. *Engineering Geology*, 27: 509–542.

Lee, S., and Choi, J. (2004) Landslide susceptibility mapping using GIS and the weight-of-evidence model. *International Journal of Geographical Information Science*, 18(8): 789–814.

Maling, D. H. (1973) *Coordinate Systems and Map Projections* (London: George Philip).

Malczewski, J. (1999) *GIS and Multicriteria Decision Analysis* (New York: Wiley).

Malczewski, J. (2000) On the use of weighted linear combination method in GIS: common and best practice approaches. *Transactions in GIS*, 4(1): 5–22.

Mather, P. M. (1995) Map-image registration using least-squares polynomials. *International Journal of Geographical Information Systems*, 9(5): 543–545.

McHarg, I. (1969) *Design with Nature* (New York: Natural History Press).

Mertens, B. and Lambin, E. F. (2000) Land-cover-change trajectories in Cameroon. *Annals of the Association of American Geographers*, 90(3): 467–494.

Morad, M., Chalmers, A. I., and O'Regan, P. R. (1996) The role of root-mean-square error in geo-transformation of images in GIS. *International Journal of Geographical Information Systems*, 10(3): 347–353.

Openshaw, S., Carver, S., and Fernie, F. (1989) *Britain's Nuclear Waste* (London: Pion).

Saaty, T. L. (1977) A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15: 234–281.

Serneels, S. and Lambin, E. F. (2001) Proximate causes of land-use change in Narok District, Kenya: a spatial statistical model. *Agriculture, Ecosystems and Environment*, 85: 65–81.

Sieber, R. (2006) PPGIS: a literature review and framework. *Annals of the Association of American Geographers*, 96: 491–507.

Sparkes, A. and Kidner, D. (1996) A GIS for the environmental impact assessment of wind Farms. Available at http://proceedings.esri.com/library/userconf/europroc96/PAPERS/PN26/PN26F.HTM.

Tomlin, D. (1990) *Geographic Information Systems and Cartographic Modeling* (Englewood Cliffs, NJ: Prentice Hall).

Unwin, D. J. (1995) Geographical information systems and the problem of error and uncertainty. *Progress in Human Geography*, 19(4): 549–558.

Unwin, D. J. (1996) Integration through overlay analysis. In: M. Fischer, H. J. Scholten and D. Unwin, eds., *Spatial Analytical Perspectives in GIS* (London: Taylor & Francis), pp. 129–138.

Unwin, D. J. and Mather, P. M. (1998) Selecting and using ground control points in image rectification and registration. *Geographical Systems*, 5(3): 239–260.

Veregin, H. (1989) Error modelling for the map overlay operation. In: Goodchild, M. F. and Gopal, S., eds., *Accuracy of Spatial Databases* (London: Taylor & Francis), pp. 3–18.

Wang, S. Q. and Unwin, D. J. (1992) Modelling landslide distribution on loess soils in China: an investigation. *International Journal of Geographical Information Systems*, 6(5): 391–405.

Wrigley, N. (1985) *Categorical Data Analysis for Geographers and Environmental Scientists* (Harlow, England: Longman).

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 12

New Approaches to Spatial Analysis

CHAPTER OBJECTIVES

This final chapter deals with methods for the analysis of geographic information that rely heavily on the existence of computer power, a process that has been called *geocomputation*. It differs from earlier chapters in two respects. First, because we cover a lot of new ground in overview, you will find its style different, with many pointers to further reading. If you want to be up-to-date with the methods we discuss, we advise you to follow up these references to the research literature. Second, most of these methods have been developed relatively recently. At the time of writing, we do not know if any of them will become part of the mainstream geographic information analyst's toolkit. It follows that our treatment is provisional and, to an extent, partial. However, because these methods originate in changes in the wider scientific enterprise, rather than in the latest technological fads, we are confident in presenting them as representative of new approaches to geographic information analysis.

Bearing these comments in mind, our aims in this chapter are to:

- Discuss recent changes in the GIS environment, both technical and theoretical
- Describe the developing field of *geocomputation*
- Describe recent developments in *spatial modeling* and the linking of such models to existing GIS

After reading this chapter, you should be able to:

- Describe the impact on the GIS environment of increases in both quantities of data and computer processing power

- Briefly outline the implications of *complexity* for the application of statistical ideas in geography
- Describe emerging geographic analysis techniques in *geocomputation* derived from *artificial intelligence*, *expert systems*, *artificial neural networks*, *genetic algorithms*, and *software agents*
- Describe *cellular automaton* and *agent-based* models and how they may be applied to geographic problems, and outline possible ways of coupling spatial models to GIS
- Describe the implications for spatial geographic information analysis of developments in *networked computing* with reference to *computational complexity*
- Discuss how online *virtual earth applications* and *user-generated map content* may affect geographic information analysis

12.1. THE CHANGING TECHNOLOGICAL ENVIRONMENT

Imagine a world where computers are rare, expensive, enormous, and accessible only to a small number of experts. This was the world in which many of the techniques we have introduced in this book were first developed. If you have consulted the references at the end of each chapter, you will have unearthed research from as far back as the 1950s and some published even earlier. Even some of the more advanced techniques we have discussed are well into "middle age." Kriging is a child of the 1960s (Matheron, 1963), with venerable parents (Youden and Mehlich, 1937, cited in Webster and Oliver, 2007), while Ripleys' K function first saw the light of day in 1976 (Ripley 1976). By contrast, the *International Journal of Geographical Information Systems* first appeared in 1987. Spatial analysis was going strong well before GIS was a gleam in the collective geographic eye. In short, contemporary GIS systems are used in a world that is very different from the one in which classical spatial analysis was invented.

Of course, both of the methods mentioned above—kriging and the K function—would be all but impossible without computers, and most of the methods we have discussed have seen continuous development before, during, and since the advent of cheap, powerful computing on the desktop. Electronic computers themselves are well over half a century old, but it is hard to exaggerate just how rapidly the computational environment has changed. In the late 1960s in the United States, the earliest scientific calculator cost around $5000—about $30,000 at today's prices. So, 30 years ago, for the price of 10 modern (very) powerful desktop personal computers (PCs), you could buy a machine weighing 18 kg that could do basic arithmetic, some trigonometry, and not much else. By contrast, today's laptop PCs are as capable as the room-sized mainframe computers of 30 years ago, at a fraction of the cost.

A Personal Note

The first computer David O'Sullivan used regularly was an Apple II Plus, the state of the art in desktop computers in 1980. That machine had a 1-megahertz 8-bit processor, 64 kilobytes of random access memory (RAM) and no hard drive. His most recently purchased computer, 2008 vintage, used by his children, has twin 2.6-gigahertz 32-bit processors (around 20,000 times more processing power than the Apple II) and 2 gigabytes of RAM plus 512 megabytes more on the video card (providing about 40,000 times as much RAM as the Apple II). This computer also has 500 gigabytes of hard disk space (which already seems limited), whereas the 1980s computer had only 140-kilobyte floppy disks. The laptop computer this chapter is being written on (one of four 'proper' computers in a household of four people) is similarly specified, and, of course, there are mobile phones, digital cameras, and music players in the house, each with as much processing power (if not more) than a PC from 30 or so years ago.

This history and anecdote are interesting, but what do they have to do with spatial analysis? One of the arguments of this chapter is that changes in computing have completely altered how spatial analysis is and should be conducted. This is not to diminish the importance of all the classical material and concepts that you have plowed through to get this far, but it is to suggest that the development of the computing environment in which we work affects both the questions that *are* asked and *can* be asked and the approaches that *are* and *can* be taken to answer them. This claim is debatable, but the debate is important and likely to have far-reaching effects for anyone engaged in GIS and spatial analysis.

Two interrelated changes are often asserted to have occurred. First, computer power is more plentiful and cheaper; second, data are more plentiful, as well as easier and cheaper to acquire. These changes are interrelated to the degree that most of the more numerous data are produced with the aid of the more plentiful computing resources; conversely, much of the great increase in computing power is dedicated to analyzing the growing amount of data. Although in broad outline these claims are self-evidently true, we should pause to consider them a little more closely:

- Cheap, powerful computing is more widely available than ever. This is obviously true, but it is worth pointing out that an enormous proportion of current computing power either remains unused or is used for tasks that make small demands on the available resources (word processing, buying books online, answering trivia questions on Wikipedia and so on).
- At first, the claim that data are cheaper and more plentiful than ever is also hard to refute. Certainly, large generic data sets such as (government-gathered census data and detailed, remote-sensed imagery) are more readily available to researchers in more convenient forms than previously could have been imagined. We use the term "generic" advisedly: such data are often not gathered with specific questions in mind. They are not gathered to assist researchers in answering a specific question or to test a specific hypothesis. High-quality data, properly controlled for confounding variables and so forth, are as expensive as ever to obtain in the natural sciences, and in the social sciences they are perhaps as *unattainable* as ever (see Sayer, 1992).

So, although we would agree that the computational environment of GIS has changed considerably, it is important to be clear about exactly what has changed. None of the changes that have occurred have fundamentally altered the basic concepts discussed in the previous chapters. This can occasionally be a difficult truth to hold on to amid all the hype that surrounds contemporary technological developments. It also warns against the simplistic idea that so much data and computing power are now available that we are in a position to answer all geographic questions. First, many of the questions are difficult and are likely to remain resistant to even the most computationally sophisticated methods. David Harel's book on computational complexity, *Computers Ltd: What They Really Can't Do* (Harel, 2000), lists a large number of interesting, essentially spatial problems that simply cannot and will never be solved exactly using any digital computer. Second, perhaps even *because* the data are cheap, many readily available data sets simply may not allow us to answer many very interesting questions. It may be more realistic to suggest that the data merely allow us to ask more questions! Answering those questions may actually require us to gather yet more data to answer them. Having sounded this note of caution, we return to the theme of the most recent technological developments in Section 12.4.

12.2. THE CHANGING SCIENTIFIC ENVIRONMENT

Meanwhile, it's not just technologies that have changed. Until relatively recently, the scientific worldview was *linear*. If the world really was linear, equations like $Y = a + bX$ would always describe the relationships between things very well. More importantly, in a linear world, the effect of X on Y would always be *independent* of all the other factors that might affect Y. In fact, we know very well that this view of things is rarely tenable. Simple $Y = a + bX$ expressions rarely describe the relationships between factors very well. Most relationships are nonlinear, meaning that a small increase in X could cause a small increase in Y, or a big increase in Y, or even a decrease in Y, *depending on everything else*—in other words, all the other factors that affect Y. In practice, even common day-to-day observable events are highly interdependent and interrelated. Today's air temperature is dependent on numerous factors: yesterday's air, ground and sea temperatures, wind directions and speeds, precipitation, humidity, air pressure, and so on. And all of these factors, in turn, are related to one another in complex ways. The *complexity* we are describing should be familiar to you. For example, when the U.S. Federal Reserve lowers interest rates by 0.25%, 25 different experts can offer 25 different opinions on how the markets will react—only for the markets to react in a 26th way that none of the experts anticipated. Indeed, writing in the wake of a year of turmoil in world financial markets, the previous sentence would seem to dramatically understate the unpredictability! In spite of the ubiquity of complex realms like these, where science, for all its sophistication, has relatively little useful to say, the linear view of the world has persisted. This is partly because the technologies that have been the product of that view have been so effective.

Complexity is a technical term for an emerging scientific, *nonlinear* view of the world (see Waldrop, 1992). The study of *complex systems* has its origins in thermodynamics (Prigogine and Stengers, 1984) and biology (Kauffman, 1993), two areas where large systems with many interacting elements are common. Some of these ideas have begun to make their way into physical geography and biogeography (Harrison, 1999; Malanson, 1999; Phillips, 1999), and also into human geography and the social sciences (Allen, 1997; Byrne, 1998; Portugali, 2000; Manson, 2001; O'Sullivan, 2004). Perhaps the key insight provided by the complexity perspective is that when we work with nonlinear systems, there is a limit to our power of prediction *even if we completely understand the mechanisms involved*. This is why the weather forecast is still wrong so often and why economic forecasts are almost always wrong.

Critically, most of the mathematics required for this nonlinear worldview is beyond the reach of analytic techniques, and computers are therefore essential to the development of theories about complex systems and complexity. This is similar to the problem that statistical distributions can be impossible to derive analytically but may be readily simulated by Monte Carlo methods. It may even be that the reason that science remained blind to the evident complexity of the real world until recently was not solely the technological and explanatory success of the conventional view, but the unavailability of any conceptual or practical tools with which to pursue alternative views. In the same way that Galileo's telescope enabled the exploration of a new astronomy, the modern computer is enabling exploration of the "new" world of complex systems. There is nothing unique or revolutionary about this development: the tools and concepts of any research program have always been interrelated in this way. Indeed, many of these developments were foreseen many years ago in a prophetic article by Warren Weaver (1948).

The major topics in this chapter may all be seen as manifestations in geography and GIS of these broader changes in the tools (computers) and ideas (the world is complex and is not reducible to simple linear mathematical descriptions) of science more generally:

- Increases in computing resources have led to attempts to develop automated "intelligent" tools for the exploration of the greatly increased arrays of data that may contain interesting patterns indicative of previously undiscovered relationships and processes. We have already discussed the Geographical Analysis Machine (GAM), an early example of this trend, in Section 6.7. In Section 12.3, we place GAM in its wider context of geocomputation. Many of the methods we discuss force no particular mathematical assumptions about the underlying causes of patterns, so that nonlinear phenomena may be investigated.
- *Computer modeling and simulation* are becoming increasingly important throughout geography. Such models are distinct from the statistical process models discussed in Chapter 4 in that they aim to represent the world *as it is*, in terms of the actual causal mechanisms that give rise to observable phenomena. These models usually explicitly represent the elements that constitute the complex systems being studied. We discuss the links between GIS and such models in Section 12.4.

12.3. GEOCOMPUTATION

The most direct response within the GIS and spatial analysis communities to these changes has been a set of new techniques loosely gathered under the heading of *geocomputation*. This term has given rise to a conference series, special journal issues, and two collections of articles (Longley et al., 1998; Abrahart and Openshaw, 2000). Nevertheless, it remains difficult to pin down its meaning, with a number of definitions offered in the cited collections. At its simplest, it might be defined as "the use of computers to tackle geographic problems that are too complex for manual techniques." This is a little vague, leaving open, for example, the question of whether or not day-to-day use of GIS qualifies as geocomputation. It is also unclear how this distinguishes geocomputation from earlier work in quantitative geography, since, even if they were big, sluggish, room-sized beasts programmed by cards full of punched holes, computers were almost always used.

While we use *geocomputation* in this section as a convenient catchall term for a wide variety of approaches, we suspect that any formal definition will ultimately be related to *computational complexity* (Harel, 2000). We return to this concept in more detail in Section 12.4. For now, it is sufficient to see it as making developments in programming algorithms and computer data structures central to improvement in our capacity to tackle larger and more difficult problems. Among the more ambitious of the variety of perspectives available was that initiated by Stan Openshaw and which continues to be developed by colleagues in the Centre for Computational Geography at the University of Leeds. Their perspective focuses on the question: *can we use (cheap) computer power in place of (expensive) brain power to help us discover patterns in geospatial data?* Most methods that start from this question are derived from *artificial intelligence* (AI) techniques, and this is probably what most clearly differentiates geocomputational approaches from earlier work. AI is itself a broad field, with almost as many definitions as there are researchers. For our purposes, a definition from the geography literature will serve as well as any other:

> [AI] is an attempt to endow a computer with some of the intellectual capabilities of intelligent life forms without necessarily having to imitate exactly the information processing steps that are used by human beings and other biological systems. (Openshaw and Openshaw 1997, p. 5)

Unsurprisingly, there are numerous approaches to endowing a computer with intelligence. We will not concern ourselves with the question of whether it is even possible, instead noting that in certain fields (chess, for example), computer programs have certainly been designed that can outperform any human expert. In any case, a number of AI techniques that have emerged have been applied to geographic problems, and we discuss these in the sections that follow.

First, it is instructive to consider the example of the original Geographical Analysis Machine (GAM) (see Section 6.7 and Openshaw et al., 1987) and to identify why it *is not* intelligent so that the techniques we consider in this section are a little better defined. You will recall that the GAM exhaustively searches a study area for incidences of unusually large numbers of occurrences of some phenomenon relative to an at-risk population. This not an intelligent approach, because the tool simply scans the entire study area, making no use of anything it finds to modify subsequent behavior. Equally, it does not change its definition of the problem to arrive at an answer—for example, by searching in regions other than circles. Both of these behaviors are characteristic of the way a human expert might approach the problem. For example, as an investigation proceeds, a researcher is likely to pay particular attention to areas similar to others where suspected clusters have already been identified. If a number of "linear" clusters associated with (say) overhead power transmission lines were noted early on, a human-led investigation might redirect resources to search for this phenomenon. Such *adaptability* and the ability to make effective use of previously acquired information—in other words, to *learn*—are elements of many definitions of intelligence.

Expert Systems

One of the earliest AI approaches is the *expert system* (see Naylor, 1983). The idea is to construct a formal representation of human expert knowledge in a field of interest. This *knowledge base* is stored as a set of *production rules* with the form

$$\text{IF} < \text{condition} > \text{THEN} < \text{action} > \tag{12.1}$$

A driving expert system might have a production rule

$$\text{IF} < \text{red light} > \text{THEN} < \text{stop} > \tag{12.2}$$

In practice, production rules are more complicated than this and may involve assigning weights or probabilities to intermediate actions before a final action is determined. A better example than a driving system is a medical diagnosis expert system that uses information about a patient's symptoms to arrive at a disease diagnosis. Some recommended actions may require tests for further symptoms, and a complex series of rules is followed to arrive at a final answer.

An expert system is guided through its knowledge base by an *inference engine* to determine what rules to apply and in what order. The other components of an expert system are a *knowledge acquisition system* and an *output device*. Output from an expert system can "explain" why a conclusion has been reached by storing the rules that were used to arrive at it. This is important in many applications. Expert systems have been used with some success in a number of areas, notably playing chess and medical diagnosis. The basic idea is employed in numerous embedded processor applications, where the expert system is not immediately apparent. Almost without exception, modern cars use expert systems to control functions such as fuel injection (dependent on driving conditions, temperature, engine temperature) and braking (antilock braking systems are expert systems). Some "fly-by-wire" airplanes also use expert systems to "interpret" the pilot's actions, ensuring that only changes to the control surfaces that will not cause the plane to crash are acted upon.

The major obstacle to building an expert system is construction of the knowledge base, which involves codifying complex human knowledge that may not previously have been written down. The technique has seen only limited application in geography. Applications in cartography have attracted some interest, since it seems that a cartographer's knowledge might be easily codified, but no artificial cartographer has yet been built. Instead, piecemeal contributions have been made that "solve" various aspects of the map design problem (Joao, 1993; Wadge et al., 1993). A more ambitious attempt to construct an expert GIS is discussed by Smith et al. (1987). It is not clear that an expert system could be successfully developed for the open-ended and ill-defined task of spatial analysis. A related approach is for a computer to know enough about the stages in a spatial analysis task that it can suggest candidate processing steps or *work flows* that might achieve a desired outcome (O'Brien and Gahegan, 2004).

Artificial Neural Networks (ANNs)

While expert systems are loosely based on a theory that

$$\text{knowledge} + \text{reasoning} = \text{intelligence} \tag{12.3}$$

artificial neural networks (ANNs) are based on the less immediately obvious idea that

$$\text{brain-like structure} = \text{intelligence} \tag{12.4}$$

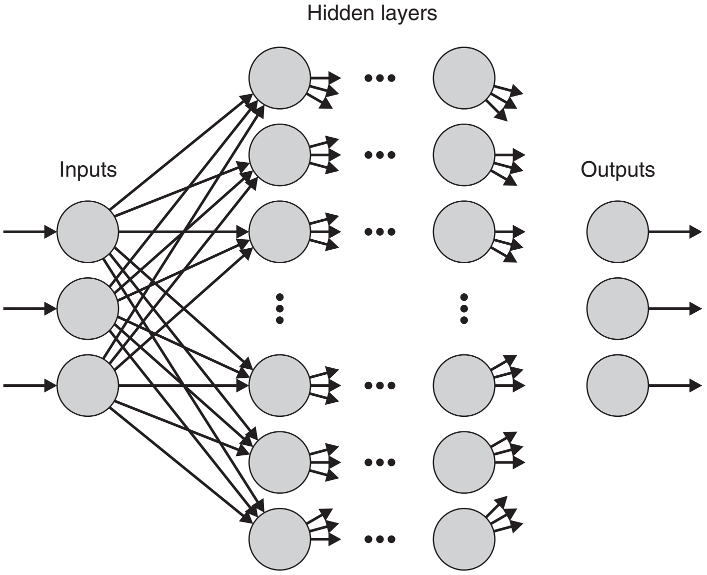


Figure 12.1 Schematic representation of an ANN.

An ANN is a very simple model of brain function. It consists of an interconnected set of artificial *neurons*. A neuron is a simple element with a number of inputs and outputs (McCulloch and Pitts, 1943). The value of the signal at each output is a function of the weighted sum of all the signals at its inputs. Usually, signal values are limited to 0 or 1 or must lie in the range 0 to 1. Various interconnection patterns are possible. A typical example is shown schematically in Figure 12.1. Note that each layer is connected to subsequent layers. For clarity, many interconnections are omitted from the diagram, and it is typical for each neuron to be connected to *all* the neurons in the next layer. One set of neurons functions as the system inputs and another as the outputs. Usually there are one or more *hidden* layers to which the inputs and outputs are connected.

Networks can operate in either *supervised* or *unsupervised* mode. A supervised network is *trained* on a set of known data. During training, the input stage is fed with data for which the desired outputs are known. The network adjusts the internal weights iteratively until a good match between its outputs and the desired outputs is obtained. This process may be thought of as *learning*. In general terms, learning proceeds by adjusting the connection weights in the network in proportion to how active they are during the training process. An unsupervised network operates more like a traditional classification procedure in that it eventually settles to a state such that different combinations of input data produce different output combinations that are similar to a cluster analysis solution.

In a typical geographic example, ANN inputs might be the signal levels on different frequency bands for a remote sensed image. The required outputs might be a code indicating what type of land cover is prevalent in each image pixel. Training data would consist of "ground truth" at known locations for part of the study area. Training stops when a sufficiently close match between the network outputs and the real data has been achieved. At this point, the network is fed new data of the same sort and will produce outputs according to the learned coding scheme. This network can now be used to classify land-cover types from the raw frequency band signal levels. Gahegan et al. (1999) provide an example of this type of application.

The final, settled state of any neural network is effectively a function that maps any combination of input data X onto some output combination of values Y, which is similar to the result of many multivariate statistical methods. Multivariate techniques that do the same thing are *discriminant analysis* and *logistic regression*. However, these are restricted to combinations of a small set of well-defined mathematical functions. The functional relation found by an ANN is not subject to this constraint and may take any form, restricted only by the complexity of the input and output coding schemes. If we imagine the variables used by the network as a multidimensional space, we can illustrate this schematically as in Figure 12.2.

Here, for simplicity, the variable space is shown as only two-dimensional. In real problems, there are many more dimensions to the dataspace, and the geometry is more complex. Cases of two different classes of observation are indicated by filled and unfilled circles. As shown in the left-hand scatterplot, the limitation of a linear classifier is that it can only "draw" straight lines through the cases as boundaries between the two classes. Except for unusually well-defined cases, numerous wrong classifications are likely. Depending on the exact structure of the ANN used, it has the potential to draw a

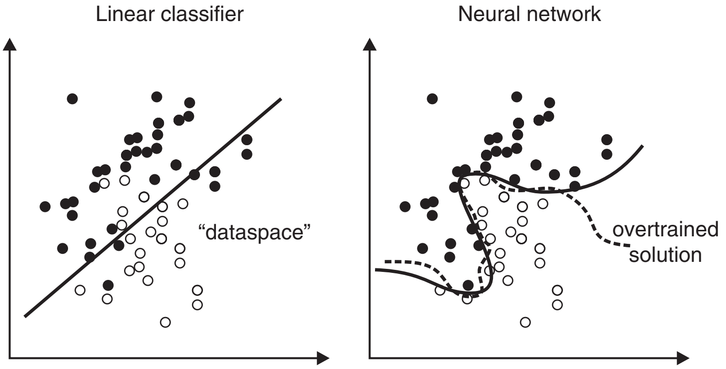


Figure 12.2 Linear classifier systems *versus* neural networks.

line of almost any shape through the cloud of observations in order to produce a much more accurate classification. ANN solutions tend to scale up better than traditional methods and are often able to handle larger, more complex problems. However, there is no way of knowing beforehand that an ANN will perform better than any other approach.

Neural networks can suffer from the problem of *overtraining* when they are matched too closely to the training data set. This means that the network has learned the particular idiosyncrasies of the training data set too well, so that when it comes to classifying other data, it performs poorly. You can think of this as analogous to the problems that may arise when a human expert becomes too familiar with a problem and tends to favor a particular diagnosis, so that it becomes hard to see other possible answers. The overtraining problem means that setting an ANN up well is a definite skill, which takes time to acquire. It also makes the selection of good training data important.

Perhaps the most troublesome thing about ANNs is just how good the answers they provide can be, even though it is hard to understand exactly how they work! In the jargon, they are *black-box* solutions, so called because we can't see what is going on "inside." Whereas in an expert system it is clear where the machine's knowledge resides and how the system arrives at its answers, with neural networks it is difficult to identify which part of the system is doing what. After applying a neural network to a problem, we may be able to solve similar problems in the future, but we may be no closer to understanding the issues involved. Whether or not this matters depends on what you are interested in. If it is your job to produce land cover maps based on several hundred 1-gigabyte satellite images, and a neural network solution works, then you may not care too much about not understanding *why* it works. If, on the other hand, you used a neural network to assess the fire risk in potential suburban development sites, you will need a better answer than "Because my neural network says so" when faced with questions from developers, landowners, and insurance companies.

A good overview of both expert systems and ANNs is found in Fischer (1994). Neural networks are just one example of a data mining technique. This broader category of approaches is well covered by Miller and Han (2008) in a geographic context.

Genetic Algorithms

Another AI technique that can generate answers without providing much information about how is *genetic algorithms* (GAs). These also adopt a simplified model of a natural process, in this case evolution (see Holland, 1975). Evolution of animal and plant life is essentially a slow process of trial and error. Over many generations, genetic adaptations and mutations that prove successful become predominant in a population. To approach a problem using GAs, we first devise a coding scheme to represent candidate solutions. At the simplest level, each solution might be represented by a string of binary digits, such as 10010000110011110101100. The GA works by assembling a large population of randomly generated strings of this type. Each potential solution is tried on the problem and scored on how successful it is using some *fitness criteria*. Many early solutions will be very poor (they are randomly generated, after all), but some will be better. In each *generation*, more successful solutions are allowed to "breed" to produce a new generation of solutions by various mechanisms. Two breeding mechanisms are:

- *Crossover* randomly exchanges partial sequences between pairs of strings to produce two new strings. The strings **10101**|001|**01** and 01011|**100**|11 might each be broken at the indicated points, and the strings crossed over to give **10101**|**100**|**01** and 01011|001|11.
- *Mutation* creates new solutions by randomly "flipping" bits in a member of the population. Thus, the string 1010100101 might mutate to 1010000101 when its fifth bit changes state.

These methods are loosely modeled on genetic mechanisms from nature, but in principle, any mechanism that "shakes things up" without completely scrambling everything can be useful. The idea is that some aspect of the relatively successful solutions must be right, but while there is room for improvement, "tinkering" is still worth while. Overly dramatic mutations are likely to lead to dysfunctional results, and many smaller mutations are likely to have no discernible effect on the quality of a solution, but a few may lead to improvements.

The new generation of solutions produced by breeding is tested and scored in the same way, and the breeding process is repeated through many generations until good solutions to the problem evolve. The net effect is an accelerated "breeding program" for a solution to the problem at hand. GA-generated problem solutions share with ANN the property that it is difficult to determine how they work. In spatial analysis, the problem is how to devise a way of applying the abstract general framework of GA to the types of problems that are of interest. A major difficulty can be devising fitness criteria for the problem at hand. After all, if we knew how to describe a good solution, we might be able to find it ourselves, without recourse to GAs. This is similar to the expert system problem of building the knowledge base. Examples of GAs are rare in the spatial analysis and

GIS literatures. They include Brookes (1997), Armstrong et al. (2003), and Conley et al. (2005).

Agent-Based Systems

One final AI approach is *agent* technology. An agent is a computer program with various properties, most importantly:

- *Autonomy*, meaning that it has the capacity for independent action
- *Reactivity,* meaning that it can react in various ways to its current environment
- *Goal direction,* meaning that it makes use of its capabilities to pursue tasks at hand

In addition, many agents are *intelligent* to the extent that this is possible given the limits of current AI. Many are also capable of *communicating* with other agents that they encounter. The best example of agent technology is the software used by Internet search engine providers to build their extensive databases of universal resource locations (URLs) and topics. These agents search for Web pages, compile details of topics and keywords as they go, and report details back to the search engine databases. Each search engine company may have many thousands of these agents, or *bots*, exploring the Internet at any given time, and this turns out to be an efficient way to index cyberspace. The application of this type of agent technology to searching large geospatial databases has been discussed by Rodrigues and Raper (1999).

The ability to communicate with other agents is a key attribute of agents employed in large numbers to solve problems in *multiagent systems*. Communication capabilities allow agents to exchange information about what they already know, so that they do not duplicate each other's activities. The space-time-attributes creature (STAC) was an innovative system using this idea coupled with GAs proposed by Openshaw (1993). The STAC would live and breed in a geospatial database and spend its time looking for repeated patterns of attributes arranged in particular configurations in space and time. Successful creatures that thrived in the database would be those that identified interesting patterns, and their breeding would enable more similar cases to be found. MacGill and Openshaw (1998) presented an implementation of this idea that was an adaptation of the basic GAM technique. Instead of systematically searching the whole study region, a flock of agents explores the space, continually communicating with one another about where interesting potential clusters are to be found. This approach is more efficient than the original GAM and has the advantage of being more easily generalized to allow searches in any space of interest (Conley et al., 2005).

12.4. SPATIAL MODELS

In this book, we have talked often about spatial process models. The models we have discussed are statistical and do not claim to represent the world as it is. The simplest spatial process model we have discussed, the independent random process, generates spatial patterns without claiming to represent any actual spatial process (IRP). Almost immediately when we start to tinker with the IRP, we describe models that are derived, at least in part, from a process that is hypothesized to be responsible for observed spatial patterns. For example, in the *Poisson cluster process*, a set of "parents" is distributed according to a standard IRP. "Offspring" for each parent are then randomly distributed around each parent, and the final distribution consists of the offspring only. It is difficult to separate this description from a relatively plausible account of the diffusion of plants by seeding (see Thomas, 1949).

This leads very naturally to the idea of developing process models that explicitly represent the real processes and mechanisms that operate to produce the observable geographic world. Such models might then be used in three different ways:

- As a basis for *pattern measurement* and hypothesis testing in the classical spatial analytic mode, as discussed in Chapter 5
- For *prediction* in an attempt to anticipate what might happen next in the real world
- To enable *exploration and understanding* of the way the process operates in the real world

Using statistical models does not raise serious questions about their nature or the way that they represent external reality, provided that we are properly cautious about our conclusions and keep in mind that statistical methods do not allow us to prove hypotheses, only to add to the evidence supporting alternative hypotheses. However, when we are serious about the process model as a representation of reality, our judgment about its plausibility becomes at least as important as the results of any statistical analysis.

Care is also required if we intend to use models for prediction or exploration. In either of these applications, it is crucial that we are confident about the model's representation of reality. This leaves us with a problem. In the real world, everything is connected to everything else: the system is *open*. Yet, if we want to model the dynamics of (say) a small stand of trees, it is impractical to include all relevant or potentially relevant factors, from global climate change to the economics of logging operations. Instead, we are forced to build a *closed* model of an open world. We can do this to an extent using, for example, probabilistic simulations of climate treated as an external factor. We might also decide to allow model users to control the climate and other parameters so as to examine the impact of different possible futures. In fact, this is often an important reason for building models—to explore different future scenarios. However, when assessing the predictive ability of models, it is important to keep the distinction between an open external world and necessarily closed models in mind. For example, 1950s models of urban housing markets in Western Europe and North America failed to anticipate later marriage, higher divorce rates and other social changes, the resulting smaller households, and the impact of these effects on the demand for apartments and smaller housing units. These issues are discussed by Peter Allen (1997) when he considers the appropriate use of models of human settlement systems.

In this section, we examine two contemporary technologies commonly applied to predictive spatial modeling. We also discuss some general issues concerned with linking such models to GIS. More traditional spatial interaction models are discussed by Wilson (2000), Fotheringham et al. (2000, Chapter 9), and Bailey and Gatrell (1995, pp. 348–366).

Cellular Automata

A simple style of spatial model well suited to raster GIS is the *cellular automaton* (CA). This consists of a regular lattice of similar cells, typically a grid. Each cell is in one of a finite number of discrete states at any particular moment, so that the cell state is a nominal variable. Cell states change simultaneously with every model time step according to a set of rules that define what cell state changes occur given the current state of a cell and its neighbors in the lattice.

To get an idea of how rich this apparently very simple framework is, examine Figure 12.3, which shows a very simple CA. Here the lattice is a one-dimensional row of 20 cells, and each row of the diagram down the page represents a single time step of the automaton's evolution. Each cell's evolution is affected by its own state and the state of its immediate neighbors on either side. Cells at the ends of a row are considered to have the cell at the opposite end as a neighbor, so that rows loop a round on themselves. This presentation of a one-dimensional automaton is convenient on the printed

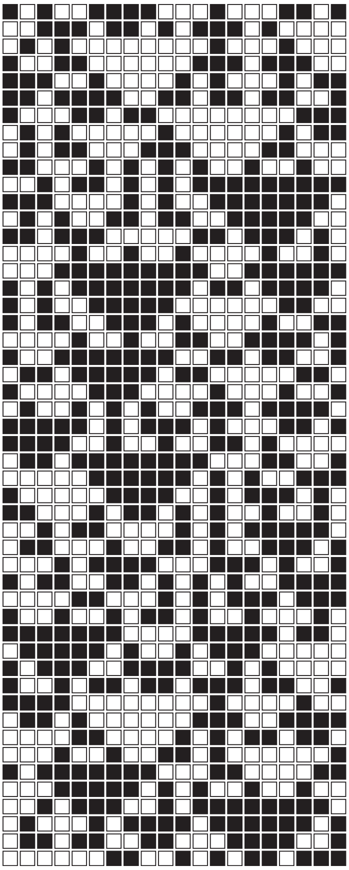


Figure 12.3 The complexity of a simple CA. The lattice state at a single moment is represented by a row of cells. Evolution of the lattice state progresses down the page.

page. The rule for this automaton is that cells with an odd number of black neighbors (counting themselves) will be black at the next time step; otherwise, they will be white. Starting from a random arrangement at the top of the diagram, the automaton rapidly develops unexpectedly rich patterns, with alternating longish sequences of exclusively black or white cells, visible as triangles in this view, as they appear and then collapse from each end over subsequent time steps.

The all-time classic CA, John Conway's Game of Life, is slightly more complex. This runs on a grid with two cell states, usually called *alive* (black) and *dead* (white). Each cell is affected by the state of its eight neighbors in the grid. The rules are simple. A dead cell comes alive if it has three live neighbors, and a live cells stays alive if it has two or three live neighbors. In print, it is hard to convey the complex behavior of this simple system, but numerous patterns of live cells have been identified that frequently occur when the CA runs. Some patterns, called *gliders* or *spaceships*, move around the lattice while preserving their shape. Others are stable configurations that spring dramatically to life when hit by a glider. Still other patterns "blink" as they cycle through a sequence of configurations before returning to their original pattern. All this rich behavior is best appreciated by watching the Life CA run on a computer. More details about the Game of Life CA can be found in Poundstone (1985).

Again, this is all very interesting but what does it have to do with geography? The point is that it is possible to build simple CA-style models in which the states and rules represent a geographic process. The important insight derived from the abstract examples above is that CA models don't have to be very complicated to do interesting things. Simple *local rules* can give rise to larger, dynamic, global structures. This suggests that, in spite of the complexity of observable geographic phenomena, it may still be possible to devise relatively simple models that replicate these phenomena and provide a better understanding of what is going on (Couclelis, 1985). In a geographic CA model, we replace the simple on/off, live/dead cell states with more meaningful states that might represent (say) different types of vegetation or land use. The rules are then based on theory about how those states change over time, depending on the context.

In practice, we must extend the CA framework considerably before we arrive at geographically plausible models. In different applications, variations on "strict" CA have been introduced. Numeric cell states, complex cell states consisting of more than one variable, rules with probabilistic effects, nonlocal neighborhoods extending to several cells in every direction, and "distance-decay" effects are among the more common adaptations. Numerous models have been developed and discussed in the research literature. Examples of urban growth and land-use models are presented by Clarke et al. (1997), Batty et al. (1999), Li and Yeh (2000), Ward et al. (2000), and White and Engelen (2000). It is also relatively easy to model phenomena such as forest fires (Takeyama, 1997) and vegetation or animal population dynamics using CA (Itami, 1994).

Agent Models

An increasingly popular alternative to CA models is *agent-based models*. These are another application of the autonomous intelligent agents already described. Instead of deploying agents in the "real world" of a spatial database, in an agent-based model the agents represent human or other actors in a simulated real-world environment. For example, the environment might be GIS data representing an urban center and agents might represent pedestrians. The model's purpose is to explore and predict likely patterns of pedestrian movement in the urban center (see Haklay et al., 2001). Other examples are provided by Westervelt and Hopkins's (1999) model of individual animal movements, models by Portugali (2000) and various coworkers examining ethnic residential segregation, models of land-use change resulting from human activity (see Evans and Kelley, 2004; Jepsen et al., 2006; Manson, 2006; and Parker et al., 2003, for an overview), and Batty's (2001) work on the evolution of settlement systems. On an altogether more ambitious scale is the *TRANSIMS* model of urban traffic developed at Los Alamos (Beckman, 1997) or models of epidemic spread (see Toroczai and Guclu, 2007, and Bian and Liebner, 2007, for a general introduction). The former model attempts to simulate urban traffic in large urban areas, such as Dallas–Fort Worth, at the scale of *individual vehicles*, using extremely detailed data sets about households and their places of work.

A good introduction to the ideas behind agent modeling is Resnick's book *Turtles, Termites and Traffic Jams* (1994). A more advanced text is Epstein and Axtell's *Growing Artificial Societies* (1996). A book that also discusses CAs and other methods we have reviewed here is Gilbert and Troitzsch's *Simulation for the Social Scientist* (2005). Agent modeling seems to have struck a chord with many researchers in various disciplines, from economics to social anthropology (see O'Sullivan, 2008, for an overview in geography), and a number of toolkits for building models are available. In general, users must write their own program code, so this is not to be undertaken lightly. Among the available systems are *StarLogo* from the MIT Media Lab; *NetLogo*, a closely related but independent project at Northwestern University; *RePast* from the University of Chicago and Argonne National Laboratories; and *Swarm* from the Santa Fe Institute. The links of each of these to GIS remain challenging to work with (although see Gimblett, 2001).

Another note of caution should be sounded. The analysis of models like these and CAs that produce detailed *and* dynamic map outputs is extremely challenging. Statistics to compare two maps (the model and the actual one) are of limited value when we don't expect the model predictions to be exact, but rather to *similar to* the way things might turn out. The reason we don't expect precise prediction is that models of these types acknowledge the complexity and inherent unpredictability of the world. It is also difficult to know how to analyze a model whose only predictions relate to the ephemeral movement of people or animals across a landscape or another environment. Cellular models are easier to handle because the predictions they make generally relate to more permanent landscape features, but they still present formidable difficulties. At present, there are few well-developed methods for addressing these problems, and this remains an area for future research (see Brown et al., 2005).

Even with better-developed map comparison techniques, the fundamental problem of any model remains: there is no way of determining statistically whether it is valid or not by examining how well it predicted past history. There are two serious problems here. First the fact that a model does well at predicting the historical record tells us nothing about how it will perform if we then run it forward into the future. Thus, if we set a model running at some known point in history (say, 1985), run it forward to a more recent known time (say, 2005), and find that the model's prediction is good, we can still have only limited confidence in what the model predicts next. This is the problem of an open world and a closed model. Second, and more fundamentally, there is no guarantee that a totally different model could not produce exactly the same result but would go on to make completely different future predictions. This is called the *equifinality* problem, and there is no escaping it except by acknowledging that, no matter what the statistics say, the theoretical plausibility of a model remains the most important criterion for judging its usefulness for forecasting.

Coupling Models and GIS

An important point to consider from a GIS perspective is how different available spatial models can be connected to available geospatial data. The issues here are similar to the general question of how GIS may be linked to spatial analytical and other statistical packages, and the fundamental problem is the same. The models used in GIS for geographic data types are different from those used in spatial modeling. Most significantly, data in GIS are generally static, whereas in spatial models they are dynamic. The raster or vector point, line, and polygon layers are not expected to change in GIS. If they do, then a whole new layer is created. If a spatial model were implemented in standard GIS, then a new layer (or layers) would be created every time anything happened in the model. For a climate model operating season by season, over a 100-year time horizon (not an unusual time span for long-range climate change studies), we end up with 400 GIS layers and all the attendant difficulties of storing, manipulating, querying, and displaying these data. Of course, we can approach the problem this way, but the real solution is to redesign the GIS data structures to accommodate the idea that objects may change over time.

As an example, consider how we might handle the changes that occur in the subdivision of land parcels over time. A plot may start out as a single unit, as shown at time 0 in Figure 12.4. It may then "grow" by acquisition at

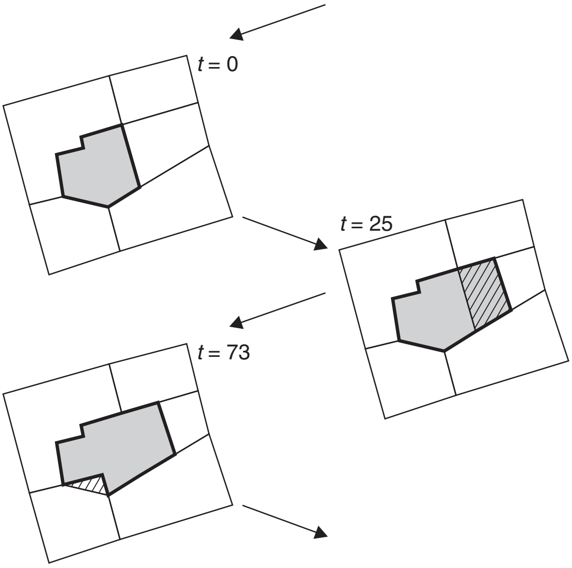


Figure 12.4 The problems of dynamic data.

time $t = 25$, only to shrink again at time $t = 73$, when a small part of the plot is sold to a neighbor. Such simple changes are only the beginning. A plot might be subdivided into several smaller parcels, some of which are retained by the same owner. Matters become more complicated when we consider the database queries that might be required in such a system. When we are only concerned with spatial relations, the fundamental relationships are "intersection," "contained within," and "within distance x of." Add to these "before," "after," "during," and, for states that persist over time, "starting before," "ending after," and so forth, and the complications for database design are obvious. This does not even consider the software design problem of how to make spatiotemporal data rapidly accessible so that animations may be easily viewed. Suffice it to say that the complexities of introducing time into GIS have yet to be widely or adequately addressed, although the issue has been on the research agenda for a long time (see Langran, 1992; O'Sullivan, 2005).

In the absence of an integrated solution to the use of models in the GIS environment, three approaches can be identified:

- *Loose coupling*, in which files are transferred between the GIS and the model and the dynamics are calculated in the model, with some display and output of results in the GIS. Usually, since the model is programmed more or less from scratch, it can be written to read and write GIS files directly. Alternatively, a good text editor, a spreadsheet, and a facility for writing script programs to do data conversions are necessary. These will remain important GIS skills for many years to come.

- *Tight coupling*, which is not very different. Data transfer is also performed by files, but each system can write files that are readable by the other. Developments in contemporary computer architectures can make this look seamless, with both programs running and continuously exchanging data. However, it is still difficult to view moving images in the GIS, and because each image requires a new file, this is still a relatively slow process.
- *Integrated model and GIS systems*, which already exist. Integration is slowly happening in three different ways: (1) by putting the required GIS functions into the model; this is easier than it sounds, because a spatial model must support spatial coordinates, measurement of distances, and so forth anyway; (2) by putting model functions into a GIS, which usually is harder, because it can be difficult to program additional functions for a GIS; and (3) by developing a generic language for building models in a GIS environment.

Distinctions between these approaches are becoming less clear-cut. As suggested above, loosely and tightly coupled solutions effectively blend into one another, depending on how and when file translation is performed, particularly given the efficiency of contemporary scripting languages such as Python, and their widespread accessibility from within GIS and other tools. In addition, the increasingly widespread availability of free open source tools, which perform many of the critical functions of a GIS, makes more or less integrated solutions that start from the modeling platform, not the GIS, the most attractive approach in many cases.

The generic modeling language approach is exemplified by *PCRaster* (Wesseling et al., 1996). Developed at the University of Utrecht, this system may be accessed online at http://pcraster.geo.uu.nl/. *PCRaster* is best understood as an extended CA-style modeling environment that also provides a GIS database, which copes with “stacks” of raster layers over time and time series. It can also produce animated maps and time series plots. The embedded dynamic modeling language (DML) makes it relatively easy to build complex geomorphologic models using built-in raster analysis functions such as aspect and slope (see Chapter 9). The system can also build surfaces from point data using kriging (see Chapter 10). Raster GIS and the CA modeling style are well suited to this integration.

More recently, there have been efforts to extend the basic CA idea of local rule-based change to irregular, non-grid-based representations of spatial data (Takeyama, 1997; O’Sullivan, 2001), and generalized spatial modeling within GIS may eventually become possible. Agent approaches seem a more likely vehicle for such integration in the long run given their greater flexibility (see Benenson and Torrens, 2004). However, it is important to realize that the problems involved extend beyond the technicalities we have been describing. In addition to the difficulties of developing appropriate spatiotemporal data structures, defining a set of dynamic spatial functions that would be required in a general system is a formidable task (see O'Sullivan, 2005).

12.5. THE GRID AND THE CLOUD: SUPERCOMPUTING FOR DUMMIES

Perceptive readers will have noticed that in this chapter on future developments, we have so far avoided much mention of perhaps the most notable technology development of the last decade or so. Even advances in desktop computing pale in comparison with the rapid growth of the Internet and the World Wide Web. The World Wide Web is itself the most visible manifestation of a move toward decentralized networked computing. These developments provide instant access to data and, increasingly, to online real-time computation. Large information technology companies routinely run tens of thousands of networked processors and can offer time on those processors as a purchasable commodity. In this setting, if you need a supercomputer, it is possible simply to rent time on someone else's network. Whoever first claimed that "the network is the computer"—it is Sun Microsystems' corporate motto, and various people associated with the company are claimed to have coined the phrase—it has proved to be a prophetic remark.

Networked computing is not a new development, but the scale of the networks now available is. While high-performance computing in specialized areas (such as computational fluid mechanics, bioinformatics, or particle physics) has historically relied first on supercomputers or, more recently, on specialized, custom-built "clusters," with specialized expertise required to make effective use of these resources, commodity processing power deployed as required relies on more diffuse architectures (the "grid" or the "cloud"). In fact, the newer architectures are highly structured behind the scenes, but the end user does not need to worry about how it all works. A layer of computer software generically termed *middleware* determines how best to partition a specific problem to run on many processors; the end users do not have to concern themselves about this question. In this environment, an application can be developed locally on a desktop computer, and then, once satisfied that the analysis is conceptually sound, the user can scale up by simply requesting that the processing be performed by whatever grid or cloud computing resources are available.

Many spatial analysis tasks are well suited to this environment. The key to effective use of multiprocessor architectures is to be able to partition the problem into smaller problems that can be processed separately and then recombined to provide a final result. Many spatial analysis methods, such as interpolation, local indicators of spatial autocorrelation, geographically weighted regression, and kernel smoothing, are readily partitioned simply by sending all the data to every processor and asking each processor to perform local operations on one part of the data set. Thus, for example, cloud or grid computing architectures are valuable for generating digital elevation models from LIDAR data sets, which would present significant challenges to even the most powerful desktop machines. The common problem of generating many synthetic data sets by permutation or randomization as part of a Monte Carlo procedure can also be run conveniently on multiple processors, without any complex analysis of the problem structure.

It is important to realize, however, that there are some problems that will not become more tractable using this technology. If you have ever performed a GIS analysis that runs on a test data set of (say) 100 data points, taking (say) 5 seconds to complete, but that seems to take forever when applied to the real data set of 10,000 data points, then you have encountered the issues central to the study of *computational complexity*. Computational complexity, according to Fortnow and Homer (2003), was first formulated by Hartmanis and Stearns (1965) and is concerned with the analysis of how computer solutions or *algorithms* scale with the size of the problem. Problems are defined by their size in terms of the data, usually denoted n, although several symbols may be required for different aspects of the problem, and algorithms are characterized in terms of how their time and space (i.e., computer memory) requirements scale with respect to the problem size. Analysis of an algorithm estimates how much time or memory a particular programmatic solution to a problem will require, depending on the size of the data set.

"Big O" notation is used to summarize the computational complexity of an algorithm. An $\mathrm{O}(n)$ algorithm is one whose run time increases linearly with the problem size, so that if the problem size doubles, the run time doubles. An $\mathrm{O}(\log n)$ problem takes time, which increases only with the log of the problem size, a slower than linear rate of increase. $\mathrm{O}(n^2)$ and $\mathrm{O}(n^3)$ algorithms are termed *polynomial*. Many spatial analysis problems lie in this area. For example, calculating all the interevent distances as part of an analysis using Ripley's K is an $\mathrm{O}(n^2)$ process, where n is the number of events. If we double the number of events, then the time required for the analysis increases four-fold. Denoting the required number of Monte Carlo simulations for some required significance level by k, then, the full analysis will be roughly $O(kn^2)$.

A more challenging class of problems is those for which the time or space requirements increase exponentially with the size of the problem—that is, $O(c^n)$ problems, where c is some constant greater than 1. Here, doubling the size of the problem, depending on the value of c, can lead to an explosion in the time or space requirements. If (say) $c = 1.5$, then for $n = 100$ and increasing to 200, the time requirements for a solution increase by a factor of 1.5^{100} or around 4×10^{17}, a very large number indeed. To put this in perspective, even if your test problem ($n = 100$) runs in 1 microsecond, a twofold increase in the problem size results in a runtime of about 12,700 years! Regardless of the value of c, there will always be some point at which relatively small increases in the size of the problem lead to dramatic increases in the time or space requirements (or both) for a solution. In this example, even if the $n = 10{,}000$ problem size runs in 1 microsecond, increasing n to just 10,100 (i.e., by 1%) will produce the same computational explosion.

Roughly speaking, polynomial problems are easy and problems with worse than polynomial characteristics are hard (which really means "impossible except for small data sets"). In fact, even polynomial problems can be problematic. Suppose your computer can produce a solution to a $n = 100$ test case in 1 minute, and the algorithm in use is $O(n^2)$; then the real data set, with $n = 10{,}000$, will take $100^2 = 10{,}000$ minutes to complete. Ten thousand minutes is almost a full seven-day week, and an $O(n^2 \log n)$ or $O(n^3)$ problem will be even worse. The reason such problems are regarded as easy is that, in terms of computational complexity, polynomial requirements can usually be met. Faster processors, more memory, and access to many processors bring the computational demands of such problems within reach. Exponential problems are hard because they do not scale so nicely, and even with improving technology, they will remain a challenge. It is these problems that demand the innovative methods discussed elsewhere in this chapter.

12.6. CONCLUSIONS: NEOGEOGRAPHIC INFORMATION ANALYSIS?

Finally, so far, we have ignored yet another development of the last few years. In the realm of complexity studies, it is common to argue that systems are more than the sum of their parts, that at some point, "more" is not merely "more" but that it fundamentally alters the nature of the system under discussion, so that "more is different" (Anderson, 1972). The enormous, rapid, and highly visible popularity of GoogleEarth and other "virtual earth" products is surely an example of such a moment in the realm of geographic information. As is often the case with commercially led developments, it can be hard to separate the hype from the reality and the significant implications of such developments from the superficial. Perhaps the most important consequence is that there is now widespread awareness of the importance of the spatial or geographic aspect of all (or almost all) data. Almost everybody now realizes that "where" is as important an attribute of our data as "what."

Placement of the spatial aspect of data firmly in the domain of the World Wide Web has opened up the field of geographic information analysis to a much wider and less specialist audience than ever before. The sheer number of "mashup" Web sites that combine spatial data from one source with map backgrounds from another, and overlay data from several sources to produce unique, new, and dynamic map products, is extraordinary. It is difficult to keep up with these developments. One Web site that tries can be found at http://googlemapsmania.blogspot.com/ and gives some idea of the diversity of end user–generated mapping now being produced. Such maps are examples of so-called *neogeography*, which is dominated by user-generated mapping and often includes spatial data generated by geolocated devices such as mobile phones or digital cameras. Geotagged photographs are a good example, as are geographic diaries, where a phone or another device has been used to track a person's position on Earth over time. Many of the data generated by such processes are difficult to accommodate in the traditional framework of geographic information analysis: can point pattern analysis be conducted on a collection of photographs, with appropriate attention paid to the content of the photographs?

Of course, not all such data call for spatial analysis at all, but the rapid proliferation of ways in which data can be generated, geolocated, and subsequently mapped brings us back to a remark made in Chapter 1, where we commented that "we often need geographic information analysis to answer questions about the significance or importance of the apparently obvious." This comment was originally made about the many user-generated maps made in GISs, but it must apply with even more force to the less formal products of neogeography, particularly where such products are used to argue a point or make a case.

Equally, the explosion of mapping-related content online has increased the rate at which free and open source software tools are being developed for the manipulation and analysis of the associated data. The resulting potential for innovative and interesting analysis to be carried out without the need for complex, expensive GIS infrastructure is an exciting development that we warmly welcome. Needless to say, we think that such developments only reinforce the need for greater awareness of many of the topics and themes in this book.

In that context, the many new techniques and approaches discussed in this chapter, which aim in one way or another to enhance our ability to carry out spatial analysis on ever-larger and more complex datasets, are important. We are certain that at least some of the techniques we have mentioned will make it into the everyday toolbox of the geographic information analyst, whether as additions to conventional GIS or, as seems more likely, as standalone tools in the loosely coupled modern computing environment. In any case, whatever the fate of artificially "intelligent" analysis techniques and their kindred simulations of the real world, we firmly believe that there will always be space for the sensitive application of human intelligence to spatial analysis within the GIS and wider communities. And that seems a good note on which to close this book.

CHAPTER REVIEW

- There have been profound changes in both the computational environment and the scientific environment in which we analyze geographic information. These changes have led to the development of methods of analysis and modeling that rely on what has been called *geocomputation*.
- Computers are now far more powerful than they were when most of the techniques discussed in this book were developed.
- *Complexity theory*, and a recognition of the need to model nonlinear effects, mean that explicit spatial prediction is rarely possible.
- In developing models, biological analogies have often been used, mimicking how humans reason in so-called *expert systems*, how brains work in *artificial neural networks* (ANNs), how species evolve by trial and error in *genetic algorithms*, (GAs), and how individuals respond to their environment and communicate with each other in *agent-based systems*. All of these have been experimented with in recent geographic research.
- True spatial models are dynamic—for example, in *cellular automata* (CA) and *agent models*.
- Coupling these types of models to existing GIS is not easy. It has been done loosely by file transfer, tightly by "wrapping" model software and GIS together, and only rarely in fully integrated systems.
- *Cloud and grid computing* bring the resources of supercomputing within reach of many more users than ever before, and many spatial analysis methods lend themselves to these approaches.

- *Computational complexity* demonstrates that many problems remain intractable, even with virtually limitless computational resources, and reinforces the need for continued innovation in spatial analysis methods.

REFERENCES

Abrahart, R. J. and Openshaw, S., eds. (2000) *GeoComputation* (London: Taylor & Francis).

Allen, P. M. (1997) *Cities and Regions as Self-Organizing Systems: Models of Complexity* (Amsterdam: Gordon Breach).

Anderson, P. W. (1972) More is different: broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047): 393–396.

Armstrong, M. P., Xiao, N., and Bennett, D. A. (2003) Using genetic algorithms to create multicriteria class intervals for choropleth maps. *Annals of the Association of American Geographers*, 93(3): 595–623.

Bailey, T. C. and Gatrell, A. C. (1995) *Interactive Spatial Data Analysis* (Harlow, England: Longman).

Batty, M. (2001) Polynucleated urban landscapes. *Urban Studies*, 38(4): 635–655.

Batty, M., Xie, Y., and Sun, Z. (1999) Modelling urban dynamics through GIS-based cellular automata. *Computers, Environment and Urban Systems*, 23: 205–233.

Beckman, R. J., ed. (1997) *The TRansportation ANalysis SIMulation System (TRANSIMS). The Dallas-Ft. Worth Case Study* (Los Alamos National Laboratory Unclassified Report LAUR-97–4502LANL).

Benenson, I. and Torrens, P. M. (2004) *Geosimulation: Automata-based Modeling of Urban Phenomena* (Chichester, England: Wiley).

Bian, L. and Liebner, D. (2007) A network model for dispersion of communicable diseases. *Transactions in GIS*, 11: 155–173.

Brookes, C. (1997) A genetic algorithm for locating optimal sites on raster suitability maps. *Transactions in GIS*, 2: 201–212.

Brown, D. G., Page, S., Riolo, R., Zellner, M., and Rand, W. (2005) Path dependence and the validation of agent-based spatial models of land use. *International Journal of Geographical Information Science*, 19(2): 153–174.

Byrne, D. (1998) *Complexity Theory and the Social Sciences: An Introduction* (London: Routledge).

Clarke, K. C., Hoppen, S., and Gaydos, L. (1997) A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environment and Planning B: Planning and Design*, 24(2): 247–262.

Conley, J. F., Gahegan, M. N., and Macgill, J. (2005) A genetic approach to detecting clusters in point data sets. *Geographical Analysis*, 37(3): 286–314.

Couclelis, H. (1985) Cellular worlds: a framework for modelling micro-macro dynamics. *Environment and Planning A*, 17: 585–596.

Epstein, J. M. and Axtell, R. (1996) *Growing Artificial Societies: Social Science from the Bottom Up* (Cambridge, MA: MIT Press).

Evans, T. P. and Kelley, H. (2004) Multi-scale analysis of a household level agent-based model of landcover change. *Journal of Environmental Management*, 72: 57–72.

Fischer, M. M. (1994) Expert systems and artificial neural networks for spatial analysis and modelling: essential components for knowledge-based geographical information systems. *Geographical Systems*, 1: 221–235.

Fortnow, L. and Homer, S. (2003) A short history of computational complexity. *Bulletin of the European Association for Theoretical Computer Science*, 80: 95–133.

Fotheringham, S., Brunsdon, C., and Charlton, M. (2000) *Quantitative Geography: Perspectives on Spatial Data Analysis* (London: Sage).

Gahegan, M., German, G., and West, G. (1999) Improving neural network performance on the classification of complex geographic data sets. *Journal of Geographical Systems*, 1: 3–22.

Gilbert, N. and Troitzsch, K. G. (2005) *Simulation for the Social Scientist*, 2nd ed. (Buckingham, England: Open University Press).

Gimblett, R., ed. (2001) *Integrating Geographic Information Systems and Agent-Based Modeling: Techniques for Understanding Social and Ecological Processes* (New York: Oxford University Press).

Haklay, M., O'Sullivan, D., Thurstain-Goodwin, M., and Schelhorn, T. (2001) "So go down town": simulating pedestrian movement in town centres. *Environment and Planning B: Planning and Design*, 28(3): 343–359.

Harel, D. (2000) *Computers Ltd: What They Really Can't Do* (Oxford: Oxford University Press).

Harrison, S. (1999) The problem with landscape: some philosophical and practical questions. *Geography*, 84(4): 355–363.

Hartmanis, J. and Stearns, R. (1965) On the computational complexity of algorithms. *Transactions of the American Mathematical Society*, 117: 285–306.

Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems* (Ann Arbor: University of Michigan Press).

Itami, R. M. (1994) Simulating spatial dynamics: cellular automata theory. *Landscape and Urban Planning*, 30(1–2): 27–47.

Jepsen, M. R., Leisz, S., Rasmussen, K., Jakobsen, J., Moller-Jensen, L., and Christiansen, L. (2006) Agent-based modelling of shifting cultivation field patterns, Vietnam. *International Journal of Geographical Information Science*, 20: 1067–1085.

Joao, E. (1993) Towards a generalisation machine to minimise generalisation effects within a GIS. In: P. M. Mather, ed., *Geographical Information Handling—Research and Applications* (Chichester, England: Wiley), pp. 63–78.

Kauffman, S. A. (1993) *The Origins of Order* (Oxford, England: Oxford University Press).

Langran, G. (1992) *Time in Geographic Information Systems* (London: Taylor & Francis).

Li, X. and Yeh, A. G.-O. (2000) Modelling sustainable urban development by the integration of constrained cellular automata and GIS. *International Journal of Geographical Information Science*, 14(2): 131–152.

Longley, P. A., Brooks, S. M., McDonnell, R., and Macmillan, B., eds. (1998) *Geocomputation: A Primer* (Chichester, England: Wiley).

MacGill, J. and Openshaw, S. (1998) The use of flocks to drive a geographic analysis machine. Presented at the *Third International Conference on GeoComputation,* School of Geographical Science, University of Bristol, England, 17–19 September.

Malanson, G. P. (1999) Considering complexity. *Annals of the Association of American Geographers*, 89(4): 746–753.

Manson, S. M. (2001) Simplifying complexity: a review of complexity theory. *Geoforum*, 32(3): 405–414.

Manson, S. M. (2006) Land use in the southern Yucatan Peninsular region of Mexico: scenarios of population and institutional change. *Computers, Environment and Urban Systems*, 30: 230–253.

Matheron, G. (1963) Principles of geostatistics. *Economic Geology*, 58: 1246–1266.

McCulloch, W. S. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Journal of Mathematical Biophysics*, 5: 115–133.

Miller, H. J. and Han, J., eds. (2008) *Geographic Data Mining and Knowledge Discovery*, 2nd ed. (London: Taylor & Francis).

Naylor, C. (1983) *Build Your Own Expert System* (Bristol, England: Sigma).

O'Brien, J., and Gahegan, M. N. (2004) Knowledge framework for representing, manipulating, and reasoning with geographic semantics. In: Z. Li, Q. Zhou, and W. Kainz, eds., *Advances in Spatial Analysis and Decision Making* (Lisse, The Netherlands: Swetz & Zeitlinger), pp. 31–44.

Openshaw, S. (1993), Exploratory space-time-attribute pattern analysers. In: A. S. Fotheringham and P. Rogerson, eds., *Spatial Analysis and GIS* (London: Taylor & Francis), pp. 147–163.

Openshaw, S., Charlton, M., Wymer, C., and Craft, A. (1987) Developing a mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, 1: 335–358.

OpenshawS. and Openshaw, C. (1997) *Artificial Intelligence in Geography* (Chichester, England: Wiley).

O'Sullivan, D. (2001) Graph-cellular automata: a generalised discrete urban and regional model. *Environment and Planning B: Planning and Design*, 28(5): 687–705.

O'Sullivan, D. (2004) Complexity science and human geography. *Transactions of the Institute of British Geographers*, 29(3): 282–295.

O'Sullivan, D. (2005) Geographical information science: time changes everything. *Progress in Human Geography*, 29(6): 749–756.

O'Sullivan, D. (2008) Geographical information science: agent-based models. *Progress in Human Geography*, 32(2): 541–550.

Parker, D. C., Manson, S. M., Janssen, M. A., Hoffmann, M. J., and Deadman, P. (2003) Multiagent systems for the simulation of land-use and land-cover change: a review. *Annals of the Association of American Geographers*, 93: 316–340.

Phillips, J. D. (1999) Spatial analysis in physical geography and the challenge; of deterministic uncertainty. *Geographical Analysis*, 31(4): 359–372.

Portugali, J. (2000) *Self-Organisation and the City* (Berlin: Springer-Verlag).

Poundstone, W. (1985) *The Recursive Universe* (New York: Morrow).

Prigogine, I., and Stengers, I. (1985) *Order out of Chaos: Man's New Dialogue with Nature* (London, England: Fontana Press).

Resnick, M. (1994) *Turtles, Termites, and Traffic Jams* (Cambridge, MA: MIT Press).

Ripley, B. D. (1976) The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13: 255–266.

Rodrigues, A. and Raper, J. (1999) Defining spatial agents. In A. S. Camara and J. Raper, eds., *Spatial Multimedia and Virtual Reality (*London: Taylor & Francis), pp. 111–129.

Sayer, A. (1992) *Method in Social Science: A Realist Approach* (London: Routledge).

Smith, T., Peuquet, D., Menon, S., and Agarwal, P. (1987) KBGIS-II: a knowledge based geographical information system. *International Journal of Geographical Information Systems*, 1: 149–172.

Takeyama, M. (1997) Building spatial models within GIS through Geo-Algebra. *Transactions in GIS*, 2: 245–256.

Thomas, M. (1949) A generalisation of Poisson's binomial limit for use in ecology. *Biometrika*, 36: 18–25.

Toroczkai, Z., and Guclu, H. (2007) Proximity networks and epidemics, *Physica A: Statistical Mechanics and Its Applications*, 378: 68–75.

Wadge, G., Wislocki, A., and Pearson, E. J. (1993) Mapping natural hazards with spatial modelling systems. In P. Mather, ed., *Geographic Information Handling—Research and Applications* (Chichester, England: Wiley), pp. 239–250.

Waldrop, M. (1992) *Complexity: The Emerging Science at the Edge of Chaos* (New York: Simon and Schuster).

Ward, D. P., Murray, A. T., and Phinn, S. R. (2000) A stochastically constrained cellular model of urban growth. *Computers, Environment and Urban Systems*, 24: 539–558.

Weaver, W. (1948) Science and complexity. *American Scientist*, 36: 536–544.

Webster, R. and Oliver, M. A. (2007) *Geostatistics for Environmental Scientists*, 2nd ed. (Chichester, England: Wiley).

Wesseling, C. G., Karssenberg, D., Van Deursen, W., and Burrough, P.A. (1996) Integrating dynamic environmental models in GIS: the development of a Dynamic Modelling language. *Transactions in GIS*, 1: 40–48.

Westervelt, J. O. and Hopkins, L. D. (1999) Modeling mobile individuals in dynamic landscapes. *International Journal of Geographical Information Science*, 13(3): 191–208.

White, R. and Engelen, G. (2000) High-resolution integrated modelling of the spatial dynamics of urban and regional systems. *Computers, Environment and Urban Systems*, 24: 383–400.

Wilson, A. G. (2000) *Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis* (Harlow, England: Prentice Hall/Pearson Education).

Youden, W. J. and Mehlich, A. (1937) Selection of efficient methods for soil sampling, *Contributions of the Boyce Thompson Institute for Plant Research*, 9: 59–70.

GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Appendix A

Notation, Matrices, and Matrix Mathematics

A.1. INTRODUCTION

In this appendix, we outline the notation that we use in this book and then some of the mathematics of *matrices* and closely related *vectors*. This material is worth mastering, because notation is important in ensuring consistency in many of the materials we present and, as will be discovered, matrices are vital to pursuing many topics in spatial analysis (and many other disciplines). In some cases, they provide a compact way of expressing questions and problems, but they also provide a useful generic way of representing the extremely important concept of adjacency in spatial systems.

We have two aims: (1) that you acquire familiarity with the notation and terminology of matrices and (2) that you become used to the way simple arithmetic operations are performed with them.

Before starting, we must introduce the basics of mathematical notation.

A.2. SOME PRELIMINARY NOTES ON NOTATION

In using mathematical notation in an introductory book, such as this, one has to steer a course between two extremes. Too rigorous adherence to a particular notation scheme can mystify the reader just as easily as a too casual approach can confuse. A further complication is that there are standard uses in the literature that need to be followed if possible. In developing this book, we have tried to be as consistent as possible and to follow some relatively straightforward basics. We hope that readers unfamiliar with the field will find this description of these basics useful.

A single instance of some variable or quantity is usually denoted by a *lowercase italicized* letter. Sometimes this is the initial letter of the quantity we're talking about—say, h for height or d for distance. More often, in

Table A.1 Commonly Used Symbols and Their Meaning in This Book

Symbol	*Meaning*
x	The Easting geographic coordinate or a general data value
y	The Northing geographic coordinate or a general data value
z, a, b	The numerical value of some measurement recorded at the geographic coordinates (x, y)
n, m	The number of observations in a data set
k	Either an arbitrary constant or the number of entities in a spatial neighborhood
d	Distance
w	The strength or "weight" of interaction between locations
$\mathbf{s}$	An arbitrary (x, y) location

introducing a statistical measure, we don't really care what the numbers represent (because they could be *anything*), so we employ one of the commonly used mathematical letters, say, x or y. Commonly used letters are x, y, z, n, m, and k. In the main text, these occur frequently, and generally have the meanings described in Table A.1. In addition to these six, you will note that d, w, and $\mathbf{s}$ also occur frequently in spatial analysis. The reason for use of an **upright bold** symbol for $\mathbf{s}$ is made clear later, where vectors and matrices are discussed.

A familiar aspect of mathematical notation is that letters from the Greek alphabet are used alongside the Roman alphabet letters that you are used to. You may already be familiar with mu (μ) for a population mean, sigma (σ) for population standard deviation, chi (χ) for a particular statistical distribution, and pi (π) for . . . well, just for "pi." In general, we try to avoid using any Greek symbols other than these, although lambda (λ) is commonly used for the intensity of a spatial process. In statistical logic, it is important to keep in mind the distinction between some parameter of an entire defined population and any estimate of that same parameter arrived at by analysis of a sample from that population. Usually, which is which will be evident from the context, but we also use Greek letters (as above) to indicate population parameters. Estimates of parameters are indicated by a "hat" symbol above the letter used for the parameter. Thus, the unknown intensity of a spatial process is indicated as λ and an estimate of it as $\hat{\lambda}$.

Symbols are introduced so that we can use mathematical notation to talk about related values or to indicate mathematical operations that we want to perform on sets of values. So, if h (or z) represents our height value, then h^2 (or z^2) indicates "height value squared." The symbols are a concise way of saying the same thing, and that's very important when we describe more complex operations on data sets.

Two symbols that you will see often are i and j. However, i and j normally appear in a particular way. To describe complex operations on sets of values, we need another notational device: the *subscript*. Subscripts are small italic letters or numbers below and to the right of normal mathematical symbols: the i in z_i is a subscript. A subscript is used to signify that there may be more than one item of the type denoted by the symbol, so z_i stands for a series or set of z values: z_1, z_2, z_3, and so on. This has various uses:

- A set of values is written between *braces*, so that $\{z_1, z_2, \ldots, z_{n-1}, z_n\}$ tells us that there are n elements in this set of z values. If required, the set as a whole may be denoted by a capital letter: Z. A typical value from the set Z is denoted z_i, and we can abbreviate the previous partial listing to simply $Z = \{z_i\}$, where it is understood that the set has n elements.
- In spatial analysis, it is common for the subscripts to refer to locations at which observations have been made and for the same subscripts to be used across a number of different data sets. Thus, h_7 and t_7 refer to the values of two different observations—say, height and temperature—at the same location ("location 7").
- Subscripts may also be used to distinguish different calculations of (say) the same statistic on different populations or samples. Thus, μ_A and μ_B denote the means of two different data sets, A and B.

The symbols i and j usually appear as subscripts in one of these ways. A particularly common usage is to denote *summation* operations, indicated by the Σ symbol (another Greek letter, this time capital sigma). This is where subscripts come into their own, because we can specify a range of values that are summed to produce a result. Thus, the sum

$$a_1 + a_2 + a_3 + a_4 + a_5 + a_6 \tag{A.1}$$

is denoted

$$\sum_{i=1}^{i=6} a_i \tag{A.2}$$

indicating that summation of a set of a values should be carried out on all the elements from a_1 to a_6. For a set of n "a" values, this becomes

$$\sum_{i=1}^{i=n} a_i \tag{A.3}$$

which is usually abbreviated to either

$$\sum_{i=1}^{n} a_i \tag{A.4}$$

or

$$\sum_i a_i \tag{A.5}$$

where the number of values in the set of a's is understood to be n. If, instead of the simple sum, we wanted the sum of the squares of the a values, then we would have

$$\sum_{i=1}^{n} a_i^2 \tag{A.6}$$

instead. Or perhaps we have two data sets, A and B, and we want the sum of the products of the a and b values at each location. This would be denoted

$$\sum_{i=1}^{n} a_i b_i \tag{A.7}$$

In spatial analysis, more complex operations might be carried out *between* two sets of values, and then we may need two summation operators. For example,

$$c = k \sum_{i=1}^{n} \sum_{j=1}^{n} \left(z_i - z_j\right)^2 \tag{A.8}$$

indicates that c is to be calculated in two stages. First, we take each z value in turn (the outer i subscript) and sum the square of its value minus every z value in turn (the j subscript). You can figure this out by imagining first setting i to 1 and calculating the inner sum, which would be $\sum_j \left(z_1 - z_j\right)^2$. We then set i to 2, and do the summation $\sum_j \left(z_2 - z_j\right)^2$, and so on all the way to $\sum_j \left(z_n - z_j\right)^2$. The final "double summation" is the sum of all of these individual sums, and c is equal to this sum multiplied by k. This will seem complex at first, but you will get used to it.

A.3. MATRIX BASICS AND NOTATION

A *matrix* is a rectangular array of numbers arranged in rows and columns; for example,

$$\begin{bmatrix} 2 & 4 & 7 & -2 \\ 0 & 1 & -3 & 3 \\ 5 & -1 & 7 & 1 \end{bmatrix} \tag{A.9}$$

As shown above, a matrix is usually written enclosed in square brackets. This matrix has three *rows* and four *columns*. The size of a matrix is described in terms of the number of rows by the number of columns, so the example above is a "3 by 4" matrix. A *square matrix* has equal numbers of rows and columns. For example,

$$\begin{bmatrix} 3 & 1 & 2 \\ 1 & -3 & 4 \\ 6 & -1 & 0 \end{bmatrix} \tag{A.10}$$

is a 3 by 3 square matrix. When we wish to talk about matrices in general terms, it is usual to represent them using uppercase **ROMAN BOLD** characters:

$$\mathbf{A} = \begin{bmatrix} 2 & 4 & 7 & -2 \\ 0 & 1 & -3 & 3 \\ 5 & -1 & 7 & 1 \end{bmatrix} \tag{A.11}$$

Individual elements in a matrix are generally referred to using *lowercase italic* characters, with their row and column numbers written as subscripts. The element in the top left corner of the above matrix is $a_{11} = 2$, and element a_{24} is the entry in row 2, column 4, and is equal to 3. In general, the subscripts i and j are used to represent rows and columns, and a general matrix has n rows and p columns, so we have

$$\mathbf{B} = \begin{bmatrix} b_{11} & \cdots & b_{1j} & \cdots & b_{1p} \\ \vdots & \ddots & & & \vdots \\ b_{i1} & & b_{ij} & & b_{ip} \\ \vdots & & & \ddots & \vdots \\ b_{n1} & \cdots & b_{nj} & \cdots & b_{np} \end{bmatrix} \tag{A.12}$$

Vectors and Matrices

A *vector* is a quantity that has size and direction. It is convenient to represent a vector graphically by an arrow of length equal to its size, pointing in the vector's direction. Typical vectors are shown in Figure A.1. In geography, vectors might be used to represent winds or current flows. In a more abstract application, they might represent migration flows. In terms of a typology of spatial data (see Chapter 1), we can add vectors to our list of types of quantity so that we have nominal, ordinal, interval, ratio, and vector types. In

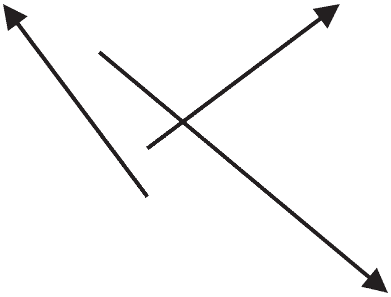


Figure A.1 Typical vectors.

particular, we can imagine a *vector field* representing, for example, the wind patterns across a region, as shown in Figure A.2.

How do we represent a vector mathematically, and what do vectors have to do with matrices? In two-dimensional space (as in the diagrams), we can use two numbers, representing the vector *components* in two perpendicular directions. This should be familiar from geographic grid coordinate systems

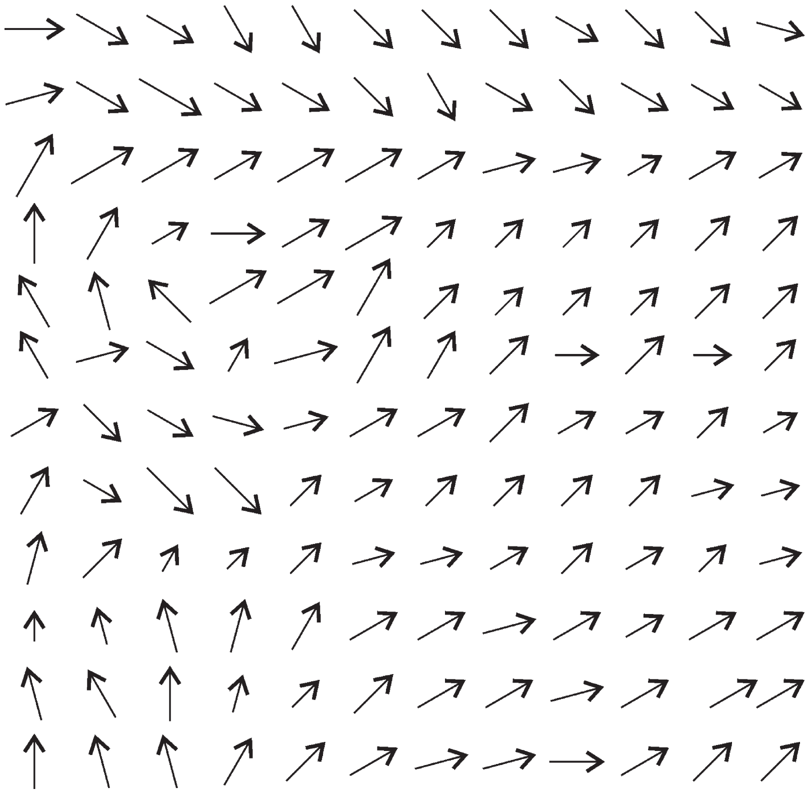


Figure A.2 A vector field.

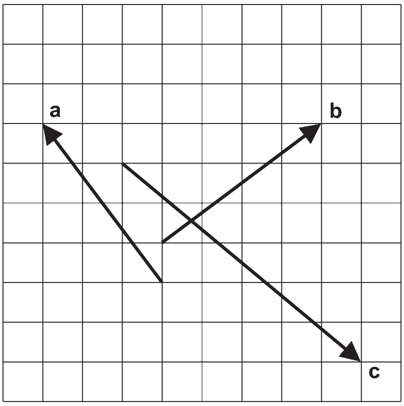


Figure A.3 Vectors in a coordinate space.

and is shown in Figure A.3. The three vectors shown have components **a** = (−3, 4), **b** = (4, 3), and **c** = (6, −5) in the east–west and north–south directions, respectively, relative to the coordinate system shown on the grid.

An alternative way to represent vectors is as *column matrices*, that is, as 2 by 1 matrices:

$$\mathbf{a} = \begin{bmatrix} -3 \\ 4 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \text{ and } \mathbf{c} = \begin{bmatrix} 6 \\ -5 \end{bmatrix} \tag{A.13}$$

Thus, a vector is a particular type of matrix with only one column. As here, vectors are usually denoted by a lowercase **roman bold** symbol. In the same way, point locations relative to an origin can be represented as vectors. This is why we sometimes use the notation in the main text where a point is represented as

$$\mathbf{s} = \begin{bmatrix} x \\ y \end{bmatrix} \tag{A.14}$$

Note also that we can represent a location in three dimensions in exactly the same way. Instead of a 2 by 1 column matrix, we use a 3 by 1 column matrix. More abstractly, in n-dimensional space, a vector will have n rows, so that it is an n by 1 matrix.

A.4. SIMPLE MATRIX MATHEMATICS

Now let us review the mathematical rules by which matrices are manipulated.

Addition and Subtraction

Matrix addition and subtraction are straightforward. Corresponding elements in the matrices in the operation are simply added (or subtracted) to produce the result. Thus, if

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \tag{A.15}$$

and

$$\mathbf{B} = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} \tag{A.16}$$

then

$$\begin{aligned} \mathbf{A} + \mathbf{B} &= \begin{bmatrix} 1+5 & 2+6 \\ 3+7 & 4+8 \end{bmatrix} \\ &= \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix} \end{aligned} \tag{A.17}$$

Subtraction is defined similarly. It follows from this that $\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$. It also follows that $\mathbf{A}$ and $\mathbf{B}$ must each have the same number of rows and columns for addition (or subtraction) to be possible.

For vectors, subtraction has a specific useful interpretation. If $\mathbf{s}_1$ and $\mathbf{s}_2$ are two locations, then the vector from $\mathbf{s}_1$ to $\mathbf{s}_2$ is given by $\mathbf{s}_2 - \mathbf{s}_1$. This is illustrated in Figure A.4, where the vector $\mathbf{x}$ from $\mathbf{s}_1$ to $\mathbf{s}_2$ is given by

$$\begin{aligned} \mathbf{x} &= \mathbf{s}_2 - \mathbf{s}_1 \\ &= \begin{bmatrix} 5 \\ 7 \end{bmatrix} - \begin{bmatrix} 8 \\ 3 \end{bmatrix} \\ &= \begin{bmatrix} -3 \\ 4 \end{bmatrix} \end{aligned} \tag{A.18}$$

Multiplication

Multiplication of matrices and vectors is more involved. The easiest way to think of the multiplication operation is that we "multiply rows into columns." Mathematically, we can define multiplication as follows: If

$$\mathbf{C} = \mathbf{AB} \tag{A.19}$$

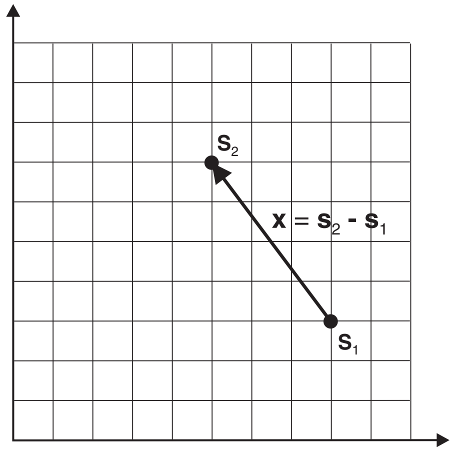


Figure A.4 Vector subtraction gives the vector between two point locations.

then the element in row i, column j of $\mathbf{C}$ is given by

$$c_{ij} = \sum_k a_{ik} b_{kj} \tag{A.20}$$

Thus, element in the ith row and jth column of the product of $\mathbf{A}$ and $\mathbf{B}$ is the sum of the products of corresponding elements from the ith row of $\mathbf{A}$ and the jth column of $\mathbf{B}$. Working through an example will make this clearer. If

$$\mathbf{A} = \begin{bmatrix} 1 & -2 & 3 \\ -4 & 5 & -6 \end{bmatrix} \tag{A.21}$$

and

$$\mathbf{B} = \begin{bmatrix} 6 & -5 \\ 4 & -3 \\ 2 & -1 \end{bmatrix} \tag{A.22}$$

then, for the element in *row 1, column 1* of the product $\mathbf{C}$, we have the sum of products of corresponding elements in *row 1* of $\mathbf{A}$ and *column 1* of $\mathbf{B}$, that is,

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ &= (1 \times 6) + (-2 \times 4) + (3 \times 2) \\ &= 6 - 8 + 6 \\ &= 4 \end{aligned} \tag{A.23}$$

Similarly, we have

$$\begin{aligned} c_{12} &= (1 \times -5) + (-2 \times -3) + (3 \times -1) \\ &= -5 + 6 + (-3) \\ &= -2 \\ c_{21} &= (-4 \times 6) + (5 \times 4) + (-6 \times 2) \\ &= -24 + 20 + (-12) \\ &= -16 \\ c_{22} &= (-4 \times -5) + (5 \times -3) + (-6 \times -1) \\ &= 20 + (-15) + 6 \\ &= 11 \end{aligned} \tag{A.24}$$

This gives us the final product matrix

$$\mathbf{C} = \begin{bmatrix} 4 & -2 \\ -16 & 11 \end{bmatrix} \tag{A.25}$$

Figure A.5 shows how multiplication works schematically. Corresponding elements from a row of the first matrix and a column of the second are multiplied together and summed to produce a single element of the product

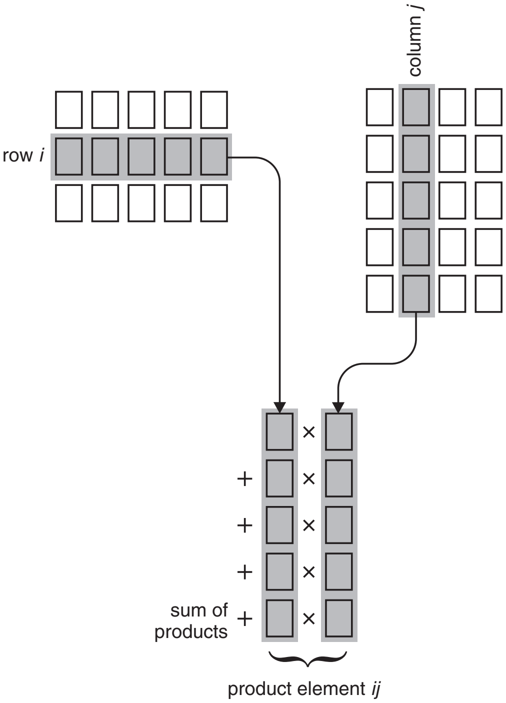


Figure A.5 Matrix multiplication.

matrix. This element's position in the product matrix corresponds to the row number from the first matrix and the column number from the second. Because of the way matrix multiplication works, it is necessary that the first matrix has the same number of columns as the second has rows. If this is not the case, then the matrices cannot be multiplied. If you write the matrices you want to multiply as ${}_n\mathbf{A}_p$ (n rows, p columns) and ${}_x\mathbf{B}_y$ (x rows, y columns), then you can determine whether they multiply by checking that the subscripts between the two matrices are equal:

$$ {}_n\mathbf{A}_p \ {}_x\mathbf{B}_y \tag{A.26}$$

If $p = x$, then this multiplication is possible and the product **AB** exists. Furthermore, the product matrix has dimensions given by the "outer" subscripts, n and y, so that the product will be an n by y matrix. On the other hand, for

$$ {}_x\mathbf{B}_y \ {}_n\mathbf{A}_p \tag{A.27}$$

if $y \neq n$, then **BA** *does not exist* and multiplication is not possible. Note that this means that, in general, for matrices

$$\mathbf{AB} \neq \mathbf{BA} \tag{A.28}$$

and multiplication is not *commutative*: it is *order dependent*. This is important when matrices are used to transform between coordinate spaces (see Section A.6).

In the example above,

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 4 & -2 \\ -16 & 11 \end{bmatrix} \tag{A.29}$$

but

$$\mathbf{D} = \mathbf{BA} = \begin{bmatrix} 26 & -37 & 48 \\ 16 & -23 & 30 \\ 6 & -9 & 12 \end{bmatrix} \tag{A.30}$$

Here the product **D** is not even the same size as **C**, and this is not unusual. However, it is useful to know that (**AB**)**C** = **A**(**BC**). The rule is that, provided the written order of multiplications is preserved, multiplications may be carried out in any sequence.

Matrix Transposition

The *transpose* of a matrix is obtained by swapping rows for columns. This operation is indicated by a superscript $^{\mathrm{T}}$, so that the transpose of $\mathbf{A}$ is written $\mathbf{A}^{\mathrm{T}}$. Hence,

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \tag{A.31}$$

Note that this definition, combined with the row-column requirement for multiplication, means that $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ always exist. The product $\mathbf{a}^{\mathrm{T}}\mathbf{a}$ is of particular interest when $\mathbf{a}$ is a vector, because it is equal to the sum of the squares of the components of the matrix. This means that the *length* of a vector $\mathbf{a}$ is given by $\sqrt{(\mathbf{a}^{\mathrm{T}}\mathbf{a})}$, from Pythagoras's Theorem. See Section A.6 for more on this topic.

A.5. SOLVING SIMULTANEOUS EQUATIONS USING MATRICES

We now come to one of the major applications of matrices. Suppose we have a pair of equations in two unknowns, x and y, for example:

$$\begin{aligned} 3x + 4y &= 11 \\ 2x - 4y &= -6 \end{aligned} \tag{A.32}$$

The usual way to solve this is to add a multiple of one of the equations to the other, so that one of the unknown variables is eliminated, leaving an equation in one unknown, which we can solve. The second unknown is then found by substituting the first known value back into one of the original equations. In this example, if we add the second equation to the first, we get

$$(3 + 2)x + (4 - 4)y = 11 + (-6) \tag{A.33}$$

which gives us

$$5x = 5 \tag{A.34}$$

so that $x = 1$. Substituting this into (say) the first equation, we get

$$3(1) + 4y = 11 \tag{A.35}$$

so that

$$4y = 11 - 3 \tag{A.36}$$

which we easily solve to get $y = 2$. This is simple enough. But what if we have 3 unknowns, or 4, or 100, or 10,000? This is where matrix algebra comes into its own. To understand how, we must introduce two more matrix concepts: the identity matrix and the inverse matrix.

The Identity Matrix and the Inverse Matrix

The identity matrix, written **I**, is defined such that

$$\mathbf{IA} = \mathbf{AI} = \mathbf{A} \tag{A.37}$$

Think of the identity matrix as the matrix equivalent of the number 1, since $1 \times z = z \times 1 = z$, where z is any number. It turns out that the identity matrix is always a square matrix with the required number of rows and columns for the multiplication to go through. Elements in **I** are all equal to 1 on the *main diagonal* from top left to bottom right. All other elements are equal to 0. The 2 by 2 identity matrix is

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{A.38}$$

The 5 by 5 identity matrix is

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{A.39}$$

and so on.

We now define the *inverse* $\mathbf{A}^{-1}$ of matrix **A**, such that

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \tag{A.40}$$

Finding the inverse of a matrix is tricky and is not always possible. For 2 by 2 matrices it is simple:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{A.41}$$

For example if

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \tag{A.42}$$

then we have

$$\begin{aligned} \mathbf{A}^{-1} &= \frac{1}{(1 \times 4) - (2 \times 3)} \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix} \\ &= -\frac{1}{2} \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix} \\ &= \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} \end{aligned} \tag{A.43}$$

We can check that this really is the inverse of **A** by calculating $\mathbf{AA}^{-1}$:

$$\begin{aligned} \mathbf{AA}^{-1} &= \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} \\ &= \begin{bmatrix} (1 \times -2) + \left(2 \times \frac{3}{2}\right) & (1 \times 1) + \left(2 \times -\frac{1}{2}\right) \\ (3 \times -2) + \left(4 \times \frac{3}{2}\right) & (3 \times 1) + \left(4 \times -\frac{1}{2}\right) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{A.44}$$

We leave it to you to check that the product $\mathbf{A}^{-1}\mathbf{A}$ also equates to **I**.

Unfortunately, finding the inverse for bigger matrices rapidly becomes much more difficult as the matrix gets bigger. Fortunately, it isn't necessary for you to know how to perform matrix inversion. The important things to remember are its definition and its relation to the identity matrix. Almost invariably, computer routines using well-known and reliable algorithms will be employed to invert any large matrices you come across.

Some other points are also worth noting:

- The quantity $ad - bc$ is known as the matrix *determinant* and is usually denoted $|\mathbf{A}|$. If $|\mathbf{A}| = 0$, then the matrix **A** has no inverse. The determinant of a larger square matrix can be found recursively from the determinants of smaller matrices known as the *cofactors* of the matrix. You will find details in books on linear algebra (Strang, 1988, is recommended).

- It is also useful to know that

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{A.45}$$

You can verify this from

$$\begin{aligned}\mathbf{B}^{-1}\mathbf{A}^{-1}\mathbf{AB} &= \mathbf{B}^{-1}\left(\mathbf{A}^{-1}\mathbf{A}\right)\mathbf{B} \\ &= \mathbf{B}^{-1}(\mathbf{I})\mathbf{B} \\ &= \mathbf{B}^{-1}\mathbf{B} \\ &= \mathbf{I}\end{aligned} \tag{A.46}$$

- Also useful is

$$\left(\mathbf{A}^{\mathrm{T}}\right)^{-1} = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}} \tag{A.47}$$

Now, Back to the Simultaneous Equations

Now we know about inverting matrices, we can get back to the simultaneous equations:

$$\begin{aligned}3x + 4y &= 11 \\ 2x - 4y &= -6\end{aligned} \tag{A.48}$$

The key is to realize that these can be rewritten as the matrix equation:

$$\begin{bmatrix} 3 & 4 \\ 2 & -4 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 11 \\ -6 \end{bmatrix} \tag{A.49}$$

Now, to solve the original equations, if we can find the inverse of the first matrix on the left-hand side, we can premultiply both sides of the matrix equation by the inverse matrix to obtain a solution for x and y directly. The inverse of

$$\begin{bmatrix} 3 & 4 \\ 2 & -4 \end{bmatrix} \tag{A.50}$$

is

$$-\frac{1}{20}\begin{bmatrix} -4 & -4 \\ -2 & 3 \end{bmatrix} \tag{A.51}$$

Doing the premultiplication on both sides, we get

$$-\frac{1}{20}\begin{bmatrix} -4 & -4 \\ -2 & 3 \end{bmatrix}\begin{bmatrix} 3 & 4 \\ 2 & -4 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = -\frac{1}{20}\begin{bmatrix} -4 & -4 \\ -2 & 3 \end{bmatrix}\begin{bmatrix} 11 \\ -6 \end{bmatrix} \tag{A.52}$$

which gives us

$$\begin{aligned}
\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} &= -\frac{1}{20}\begin{bmatrix} (-4 \times 11) + (-4 \times -6) \\ (-2 \times 11) + (3 \times -6) \end{bmatrix} \\
\begin{bmatrix} x \\ y \end{bmatrix} &= -\frac{1}{20}\begin{bmatrix} -44 + 24 \\ -22 - 18 \end{bmatrix} \\
&= -\frac{1}{20}\begin{bmatrix} -20 \\ -40 \end{bmatrix} \\
\begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} 1 \\ 2 \end{bmatrix}
\end{aligned} \tag{A.53}$$

which is the same solution for x and y that we obtained before. This all probably seems a bit laborious for just two equations! The point is that this approach can be scaled up very easily to much larger sets of equations, and provided we can find the inverse of the matrix on the left-hand side, the equations can be solved. We can generalize this result. Any system of equations can be written

$$\mathbf{Ax} = \mathbf{b} \tag{A.54}$$

and the solution is given by premultiplying both sides by $\mathbf{A}^{-1}$ to get

$$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\mathbf{b} \tag{A.55}$$

Since $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, we then have

$$\mathbf{Ix} = \mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{A.56}$$

This is an amazingly compressed statement of the problem of solving any number of equations. Remember that the matrix equation $\mathbf{Ax} = \mathbf{b}$ can represent a system of hundreds or even thousands of equations, not just two or three. Note also that if we calculate the determinant of $\mathbf{A}$ and find that it is zero, then we know that the equations cannot be solved, since $\mathbf{A}$ has no inverse. Furthermore, having solved this system once by finding $\mathbf{A}^{-1}$, we can quickly solve it for any values on the right-hand side of the equation.

Because of this general result, matrices have become central to modern mathematics, statistics, computer science, and engineering. In a smaller way, they are important in spatial analysis, as will become clear in the main text.

A.6. MATRICES, VECTORS, AND GEOMETRY

Another reason for the importance of matrices is their usefulness in representing coordinate geometry. We have already seen that a vector (in two or more dimensions) may be considered a *column vector* where each element represents the vector's length parallel to each of the axes of the coordinate space. We expand here on a point that we have already touched on relating to the calculation of the quantity $\mathbf{a}^{\mathrm{T}}\mathbf{a}$ for a vector. As we have already mentioned, this quantity is equal to the sum of the squares of the components of $\mathbf{a}$, so that the length of $\mathbf{a}$ is given by

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{a}} \tag{A.57}$$

This result applies regardless of the number of dimensions of $\mathbf{a}$.

We can use this result to determine the angle between any two vectors $\mathbf{a}$ and $\mathbf{b}$. In Figure A.6, the vector $\mathbf{a}$ forms an angle A with the positive x axis, and $\mathbf{b}$ forms angle B. The angle between the two vectors $(B - A)$ we label θ.

Using the well-known trigonometric equality

$$\cos(B - A) = \cos A\,\cos B + \sin A\,\sin B \tag{A.58}$$

we have

$$\begin{aligned}\cos\theta &= \cos A\,\cos B + \sin A\,\sin B \\ &= \left(\frac{x_a}{\|\mathbf{a}\|} \times \frac{x_b}{\|\mathbf{b}\|}\right) + \left(\frac{y_a}{\|\mathbf{a}\|} \times \frac{y_b}{\|\mathbf{b}\|}\right) \\ &= \frac{x_a x_b + y_a y_b}{\|\mathbf{a}\|\|\mathbf{b}\|} \\ &= \frac{\mathbf{a}^{\mathrm{T}}\mathbf{b}}{\sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{a}}\sqrt{\mathbf{b}^{\mathrm{T}}\mathbf{b}}}\end{aligned} \tag{A.59}$$

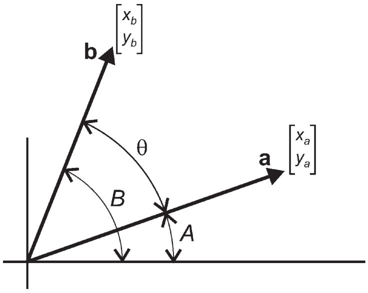

$\begin{bmatrix} x_b \\ y_b \end{bmatrix}$
b
θ
a
$\begin{bmatrix} x_a \\ y_a \end{bmatrix}$
B
A

Figure A.6 Derivation of the expression for the angle between two vectors (see text).

The quantity $\mathbf{a}^{\mathrm{T}}\mathbf{b}$ is known as the *dot product* or *scalar product* of the two vectors and is simply the sum of products of corresponding vector components. One of the most important corollaries of this result is that two vectors whose dot product is equal to zero are perpendicular or *orthogonal*. This follows directly from the fact that cos 90° is equal to zero. Although we have derived this result in two dimensions, it again scales to any number of dimensions, even if we have trouble understanding what "perpendicular" means in nine dimensions! The result is also considered to apply to matrices, so that if $\mathbf{A}^{\mathrm{T}}\mathbf{B} = 0$, then we say that matrices $\mathbf{A}$ and $\mathbf{B}$ are orthogonal.

The Geometric Perspective on Matrix Multiplication

In this context, it is useful to introduce an alternative way of understanding the matrix multiplication operation. Consider the the 2 × 2 matrix, **A**, and the spatial location vector, **s**

$$\mathbf{A} = \begin{bmatrix} 0.6 & 0.8 \\ -0.8 & 0.6 \end{bmatrix}, \mathbf{s} = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \tag{A.60}$$

The product, **As**, of these matrices is

$$\mathbf{As} = \begin{bmatrix} 5 \\ 0 \end{bmatrix} \tag{A.61}$$

We can look at a diagram of this operation in two-dimensional coordinate space, as shown on the left-hand side of Figure A.7. The vector **As** is a rotated version of the original vector **s**. If we perform the same multiplication on a series of vectors, collected together in the two-row matrix **S** so that each column of **S** is a vector,

$$\begin{aligned}\mathbf{AS} &= \begin{bmatrix} 0.6 & 0.8 \\ -0.8 & 0.6 \end{bmatrix}\begin{bmatrix} 1 & 3 & 0 & -1 & -2.5 \\ 1 & -2 & 5 & 4 & -4 \end{bmatrix} \\ &= \begin{bmatrix} 1.4 & 0.2 & 4 & 2.6 & -4.7 \\ -0.2 & -3.6 & 3 & 3.2 & -0.4 \end{bmatrix}\end{aligned} \tag{A.62}$$

then we can see that multiplication by the matrix **A** may be considered equivalent to a clockwise *rotation* of the vectors (through 53.13° for the record). These operations are shown on the right-hand side of Figure A.7 for confirmation.

In fact, *any* matrix multiplication may be thought of as a transformation of some coordinate space. This property of matrices has ensured their

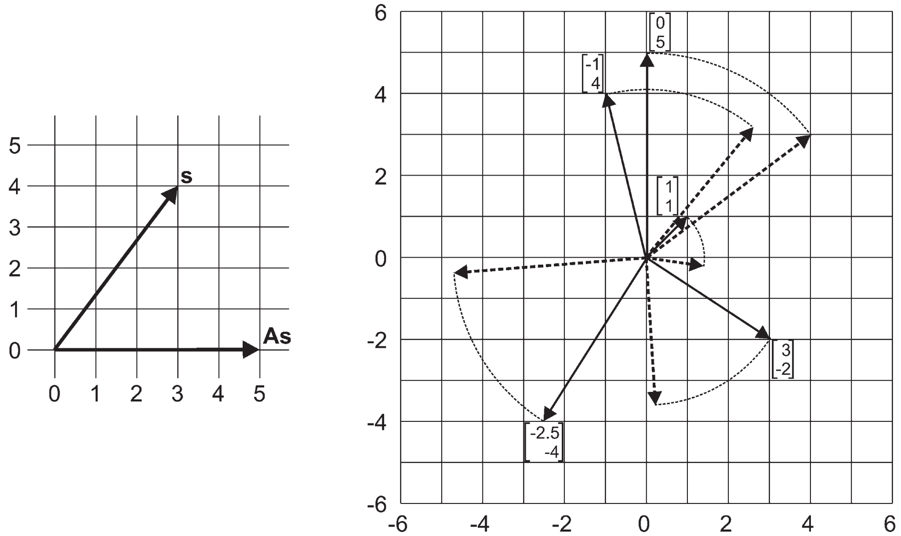


Figure A.7 Matrix multiplication as a transformation of coordinate space. In the left-hand grid, the multiplication **As** is shown. In the right-hand grid, each column of **S** is shown as a vector that is rotated after multiplication by **A** (see text).

widespread use in computer graphics, where they are an efficient way of doing the calculations required for drawing perspective views. Transformation matrices have the special property that they *project* the three dimensions of the objects displayed into the two dimensions of the screen. By changing the projection matrices used, we change the viewer's position relative to the displayed objects. This perspective on matrices is also important for transforming between geographic projections (see Chapter 11).

This perspective also provides an interpretation of the inverse of a matrix. Since multiplication of a vector **s** by a matrix, followed by multiplication by its inverse, returns **s** to its original value, the inverse of a matrix performs the opposite coordinate transformation to that of the original matrix. The inverse of the matrix above therefore performs a 53.13° *counterclockwise* rotation. You may care to try this on some examples.

A.7. EIGENVECTORS AND EIGENVALUES

Two properties important in statistical analysis are the *eigenvectors* and *eigenvalues* of a matrix. These only make intuitive sense in light of the geometric interpretation of matrices we have just introduced—although you will probably still find it a stretch. The eigenvectors $\{\mathbf{e}_1 \ldots \mathbf{e}_n\}$ and eigenvalues $\{\lambda_1 \ldots \lambda_n\}$ of an $n \times n$ matrix $\mathbf{A}$ each satisfy the following equation:

$$\mathbf{A}\mathbf{e}_i = \lambda \mathbf{e}_i \tag{A.63}$$

Seen in terms of the multiplication-as-transformation view, this means that the eigenvectors of a matrix are directions in coordinate space that are unchanged under transformation by that matrix. Note that the equation means that the eigenvalues and eigenvectors are associated with one another in pairs $\{(\lambda_1, \mathbf{e}_1), (\lambda_1, \mathbf{e}_1), \ldots (\lambda_n, \mathbf{e}_n)\}$. The scale of the eigenvectors is arbitrary, since they appear on both sides of the above equation, but normally they are scaled so that they have unit length. We won't worry too much about how the eigenvectors and eigenvalues of a matrix are determined (see Strang, 1988, for details). As an example, the eigenvalues and eigenvectors of the matrix in our simultaneous equations

$$\begin{bmatrix} 3 & 4 \\ 2 & -4 \end{bmatrix} \tag{A.64}$$

are

$$\left(\lambda_1 = 4,\ \mathbf{e}_1 = \begin{bmatrix} 0.9701 \\ 0.2425 \end{bmatrix}\right) \text{ and } \left(\lambda_2 = -5,\ \mathbf{e}_2 = \begin{bmatrix} -0.4472 \\ 0.8944 \end{bmatrix}\right) \tag{A.65}$$

It is straightforward to check this result by substitution into the defining equation above.

Figure A.8 may help to explain the meaning of the eigenvectors and eigenvalues. The unit circle shown is transformed to the ellipse shown under multiplication by the matrix we have been discussing. However, the eigenvectors have their direction unchanged by this transformation. Instead, they are each scaled by a factor equal to the corresponding eigenvalue.

An important result (again, see Strang, 1988) is that *the eigenvectors of a symmetric matrix are mutually orthogonal*. That is, if $\mathbf{A}$ is symmetric about its main diagonal, then any pair of its eigenvectors $\mathbf{e}_i$ and $\mathbf{e}_j$ have a dot product $\mathbf{e}_i^T \mathbf{e}_j = 0$. For example, the symmetric matrix

$$\begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \tag{A.66}$$

has eigenvalues and eigenvectors

$$\left(4.541,\ \begin{bmatrix} 0.6464 \\ 0.7630 \end{bmatrix}\right) \text{ and } \left(-1.541,\ \begin{bmatrix} -0.7630 \\ 0.6464 \end{bmatrix}\right) \tag{A.67}$$

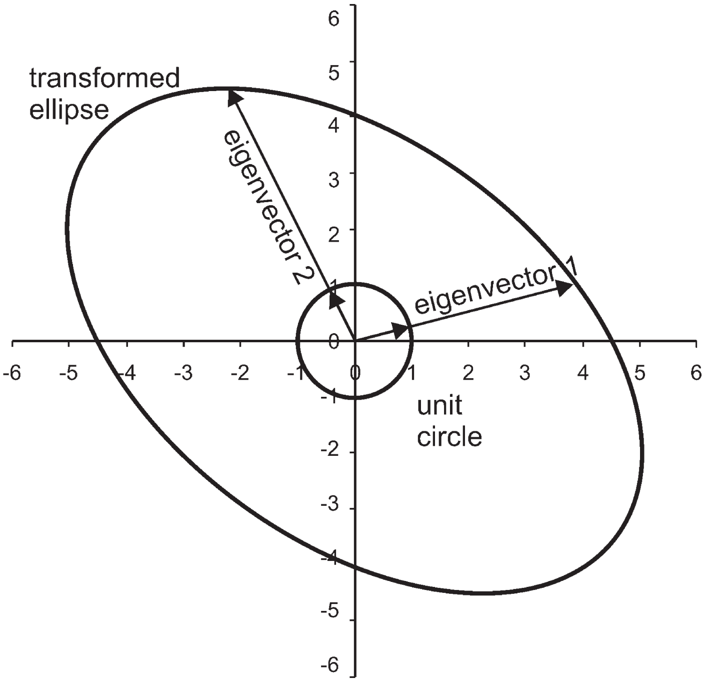


Figure A.8 The geometric interpretation of eigenvectors and eigenvalues (see text).

and it is easy to confirm that these vectors are orthogonal. The widely used method, *principal components analysis*, makes use of this result.

REFERENCES

Strang, G. (1988) *Linear Algebra and Its Applications*, 3rd ed., (Fort Worth, TX: Harcourt Brace Jovanovich).

Index