GEOGRAPHIC INFORMATION ANALYSIS

second edition

DAVID O'SULLIVAN

DAVID J. UNWIN

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

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

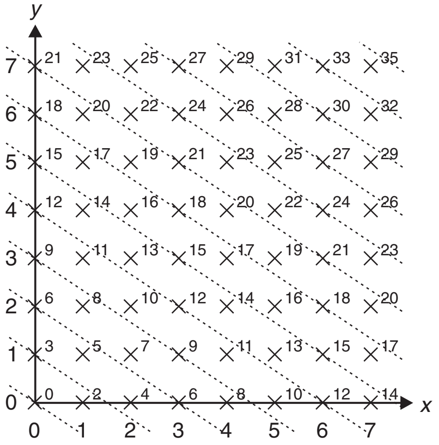

Figure 4.2 A realization of the deterministic spatial process $z = 2x + 3y$ for $0 \leq x \leq 7$, $0 \leq y \leq 7$. Contours are shown as dashed lines. This is the only possible realization because the process is deterministic.

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

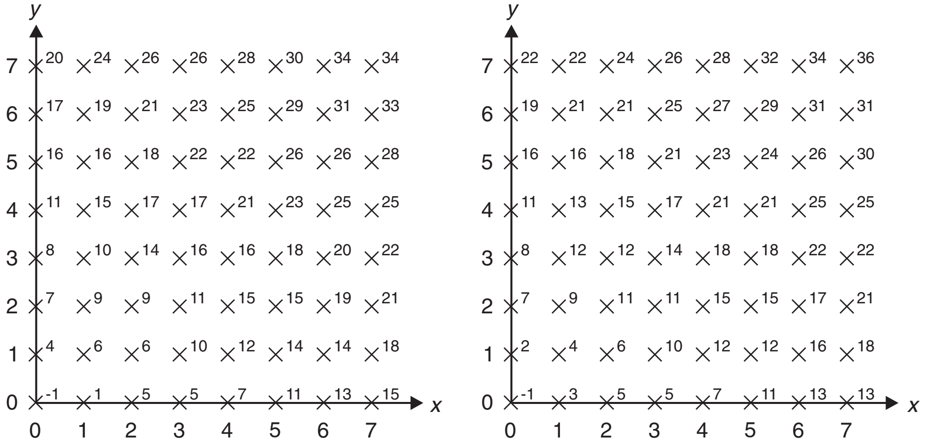

Figure 4.3 Two realizations of the stochastic spatial process $z = 2x + 3y \pm 1$ for $0 \leq x \leq 7, 0 \leq y \leq 7$.

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:

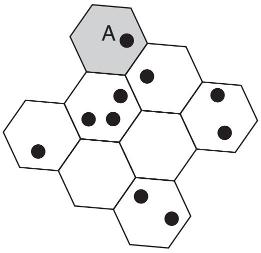

Figure 4.4 Quadrat counting for the example explained in the text.

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

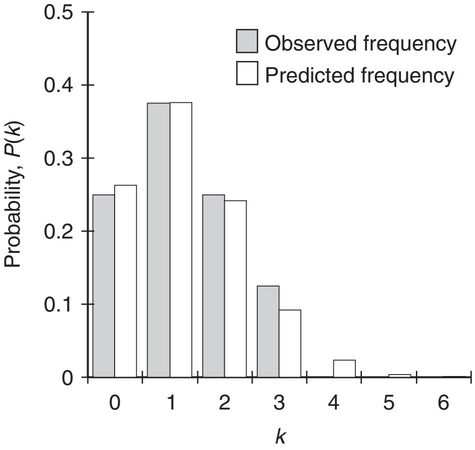

Figure 4.5 Comparison of the observed and predicted frequency distributions for the pattern in Figure 4.4.

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,

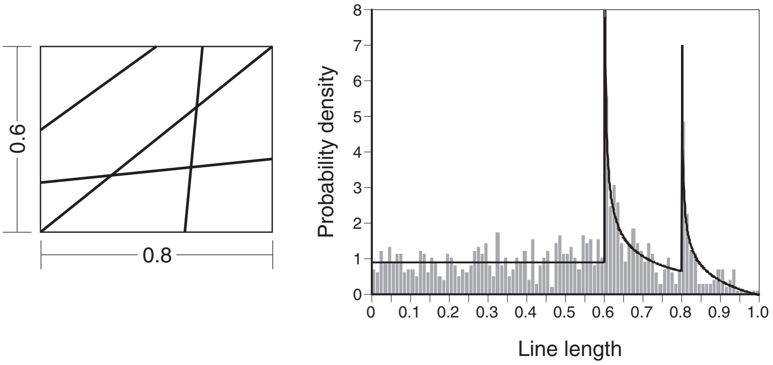

Figure 4.6 Theoretical probability density function (the line) and a single realization of the distribution of line lengths across a rectangular area (the histogram).

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.

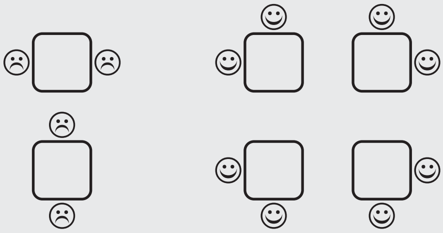

Figure 4.7 Possible ways of sitting at a coffee shop table.

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

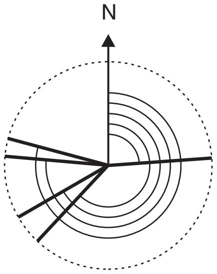

Figure 4.8 Randomly generated line segments are produced and their angle measured relative to north.

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

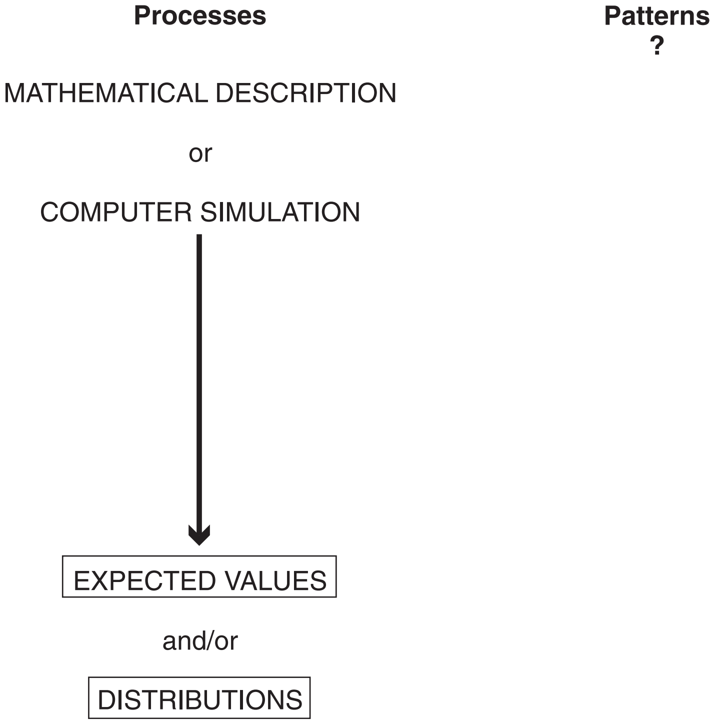

Figure 4.9 The developing framework for spatial statistical analysis. We now have a clearer picture of the meaning of a spatial process. Patterns will be tackled in the next chapter.

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.