13 Spatial Data Analysis

This chapter is the first in a set of three dealing with geographic analysis and modeling methods. The chapter begins with a review of the relevant terms and presents an outline of the major topics covered in the three chapters. Analysis and modeling are grounded in spatial concepts, allowing the investigator to examine the role those concepts play in understanding the world around us. This chapter also examines methods constructed around the concepts of location, distance, and area. Location provides a common key between datasets, allowing the investigator to discover relationships and correlations between properties of places. Distance defines the separation of places in geographic space and acts as an important variable in many of the processes that impact the geographic landscape. Areas define the neighborhood context of processes and events and also capture the important concept of scale.

LEARNING OBJECTIVES

After studying this chapter you will understand:

- Definitions of spatial data analysis and related terms and tests to determine whether a method is spatial.
- Techniques for detecting relationships between the various properties of places and for preparing data for such tests.
- Methods to examine distance effects, in the creation of clusters, hotspots on *heat maps*, and anomalies.
- The applications of convolution in GI systems, including density estimation and the characterization of neighborhoods.

13.1 Introduction: What Is Spatial Analysis?

Many terms are used to describe the techniques that are the focus of these chapters. Although spatial analysis and spatial data analysis are the ones preferred here for the techniques discussed in Chapters 13 and 14, it is also common to find them being described as data *mining* and data *analytics*. Data mining often implies large volumes of data, so is often encountered in discussions of Big Data and the general search for patterns without particular reference to hypotheses about what those patterns should look like. The term data *analytics* is increasingly used in business applications and again is popular in the context of Big Data. To all intents and purposes, however, the terms are equivalent. The definition of spatial modeling is left to Chapter 15.

The techniques covered in these three chapters are generally termed *spatial* rather than *geographic* because they can be applied to data arrayed in any space, not only geographic space (see Section 1.1.2). Many of the methods might potentially be used in analysis of outer space by astronomers, or in analysis of brain scans by neuroscientists. So the term *spatial* is used consistently throughout these chapters. It is important to recognize at the outset that spatial analysis is *different* because the standard assumptions made in many statistical tests are not valid when dealing with spatial data, a topic addressed at length in Chapter 14. It is important, therefore, to keep a clear distinction between statistical analysis and spatial analysis.

Spatial analysis is in many ways the crux of geographic information science and systems (GISS) because it includes all the transformations, manipulations, and methods that can be applied to geographic data to add value to them, to support decisions, and to reveal patterns and anomalies that are not immediately obvious. In other words, spatial analysis is the process by which we turn raw data into useful information, in pursuit of scientific discovery, or more effective decision making. If a geographic information (GI) system is a method of communicating information about the Earth's surface from one person to another, then the transformations of spatial analysis are ways in which the sender tries to inform the receiver, by adding greater informative content and value, and by revealing things that the receiver might not otherwise see.

Some methods of spatial analysis were developed long before the advent of GI systems, when they were carried out by hand or by the use of measuring devices like the ruler. The term *analytical cartography* is sometimes used to refer to methods of analysis that can be applied to maps to make them more useful and informative, and spatial analysis using GI systems is in many ways its logical successor. But it is much more powerful because it covers not only the contents of maps but also any type of geographic data.

Spatial analysis can reveal things that might otherwise be invisible—it can make what is implicit explicit.

In this chapter we will look first at some definitions and basic concepts of spatial analysis. Concepts such as location, distance, and area—the topics discussed in this chapter—have already been encountered at various points in this book, but here they serve to provide an organizing framework for the vast array of methods that fall under the heading of spatial analysis. More advanced concepts, as well as the methods used to elucidate them, are discussed in Chapter 15, along with the use of GI systems in design, when their power is directed not to understanding the world but to improving it according to specific goals and objectives. Chapter 16 addresses the use of GI systems to examine dynamic processes, primarily by simulation.

Spatial analysis is the crux of GISS, the means of adding value to geographic data and of turning data into useful information.

Methods of spatial analysis can be very sophisticated, but they can also be very simple. A large body of methods of spatial analysis has been developed over the past century or so, and some methods are highly mathematical—so much so that it might sometimes seem that mathematical complexity is an indicator of the importance of a technique. But the human eye and brain are also very sophisticated processors of geographic data and excellent detectors of patterns and anomalies in maps and images. So the approach taken here is to regard spatial analysis as spread out along a continuum of sophistication, ranging from the simplest types that occur very quickly and intuitively when the eye and brain look at a map to the types that require complex software and sophisticated mathematical understanding. Spatial analysis is best seen as a *collaboration* between the computer and the human, in which both play vital roles.

Effective spatial analysis requires an intelligent user, not just a powerful computer.

There is an unfortunate tendency in the GI systems community to regard the making of a map using a GI system as somehow less important than the performance of a mathematically sophisticated form of spatial analysis. According to this line of thought, *real* use of GI systems involves number crunching, and users who *just* use GI systems to make maps are not serious users. But every cartographer knows that the design of a map can be very sophisticated and that maps are excellent ways of conveying geographic information and knowledge by revealing patterns and processes to us (see Chapter 11). We agree; moreover, we believe that mapmaking is potentially just as important as any other application of GI systems.

Spatial analysis may be defined in many possible ways, but all definitions in one way or another express the basic idea that information on locations is essential—that analysis carried out without knowledge of locations is not spatial analysis. One fairly formal statement of this idea is the following:

- Spatial analysis is a set of methods whose results are not invariant under changes in the locations of the objects being analyzed.

The double negative in this statement follows convention in mathematics, but for our purposes we can remove it:

- Spatial analysis is a set of methods whose results change when the locations of the objects being analyzed change.

On this test the calculation of an average income for a group of people is not spatial analysis because it in no way depends on the locations of the people. But the calculation of the center of the U.S. population is spatial analysis because the results depend on knowing where all U.S. residents are located. A GI system is an ideal platform for spatial analysis because its data structures accommodate the storage of object locations, as will have become clear from Chapters 6 and 7.

The techniques discussed in these chapters are no more than the tip of the spatial analysis iceberg. Some GI systems are more sophisticated than others in the range of techniques they support, and others are strongly oriented toward certain domains of application. For a more advanced review including many techniques not mentioned here, and for a detailed analysis of the techniques supported by each package, the reader is encouraged to examine the book *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools* by De Smith, Goodchild, and Longley, which is available online at www.spatialanalysisonline.com.

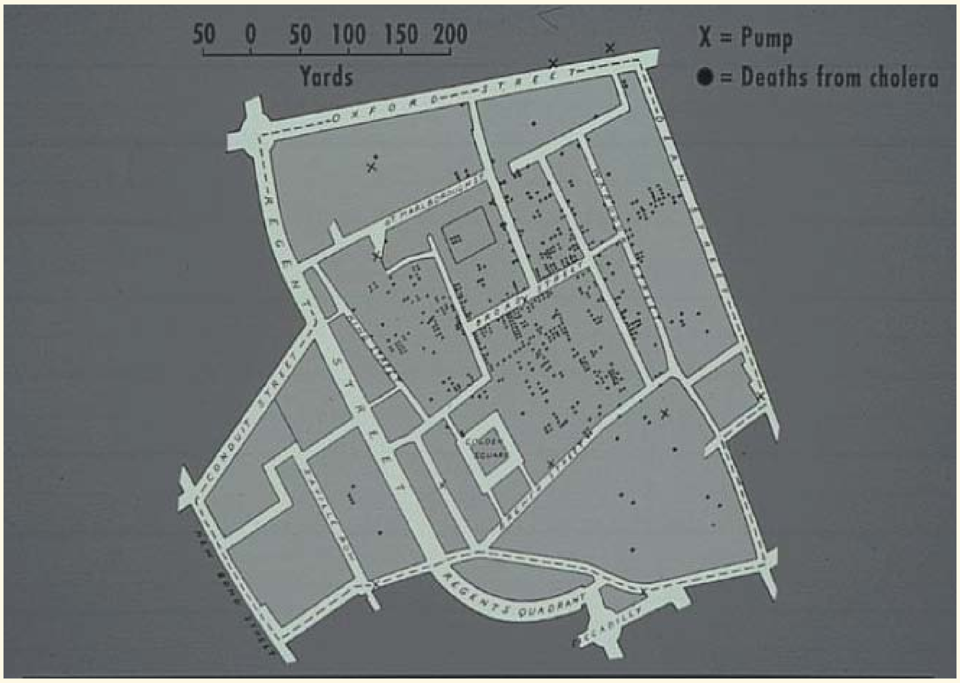
13.1.1 Examples

Spatial analysis can be used to further the aims of science, by revealing patterns that were not previously recognized and that hint at undiscovered generalities and laws. Patterns in the occurrence of a disease may hint at the mechanisms that cause the disease, and some of the most famous examples of spatial analysis are of this nature, including the work of Dr. John Snow in unraveling the causes of cholera (see Box 13.1).

Biographical Box 13.1

Dr. John Snow and the Causes of Cholera

In the 1850s cholera was poorly understood, and massive outbreaks were a common occurrence in major industrial cities. (Even today cholera remains a significant health hazard in many parts of the world, despite progress in understanding its causes and advances in treatment.) An outbreak in London in 1854 in the Soho district was typical of the time, and the deaths it caused are mapped in Figure 13.1. The map was made by Dr. John Snow (Figure 13.2), who had conceived the hypothesis that cholera was transmitted through the drinking of polluted water rather than through the air, as was commonly believed. He noticed that the outbreak appeared to be centered on a public drinking water pump in Broad Street (Figure 13.3)—and if his hypothesis was correct, the pattern shown on the map would reflect the locations

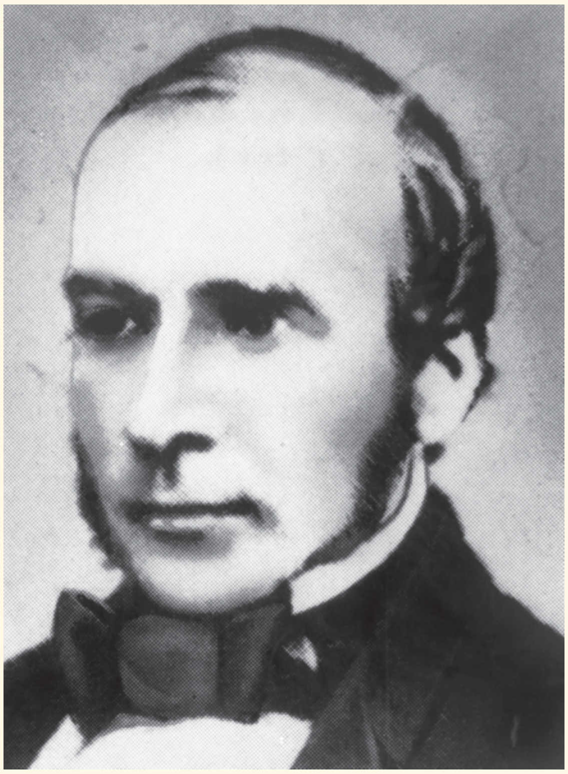


Courtesy: Gilbert E. W. 1958. Pioneer maps of health and disease in England. *Geographical Journal* 124: 172–183.

Figure 13.1 A redrafting of the map made by Dr. John Snow in 1854, showing the deaths that occurred in an outbreak of cholera in the Soho district of London. The existence of a public water pump in the center of the outbreak (the cross in Broad Street) convinced Snow that drinking water was the probable cause of the outbreak. Stronger evidence in support of this hypothesis was obtained when the water supply was cut off, and the outbreak subsided.

of people who drank the pump's water. There appeared to be anomalies, in the sense that deaths had occurred in households that were located closer to other sources of water, but he was able to confirm that these households also drew their water from the Broad Street pump. Snow had the handle of the pump removed, and the outbreak subsided, providing direct causal evidence in favor of his hypothesis. The full story is much more complicated than this largely apocryphal version, of course; much more information is available at www.jsi.com.

Today, Snow is widely regarded as the father of modern epidemiology.



Courtesy: Wellcome Library, London

Figure 13.2 Dr. John Snow.

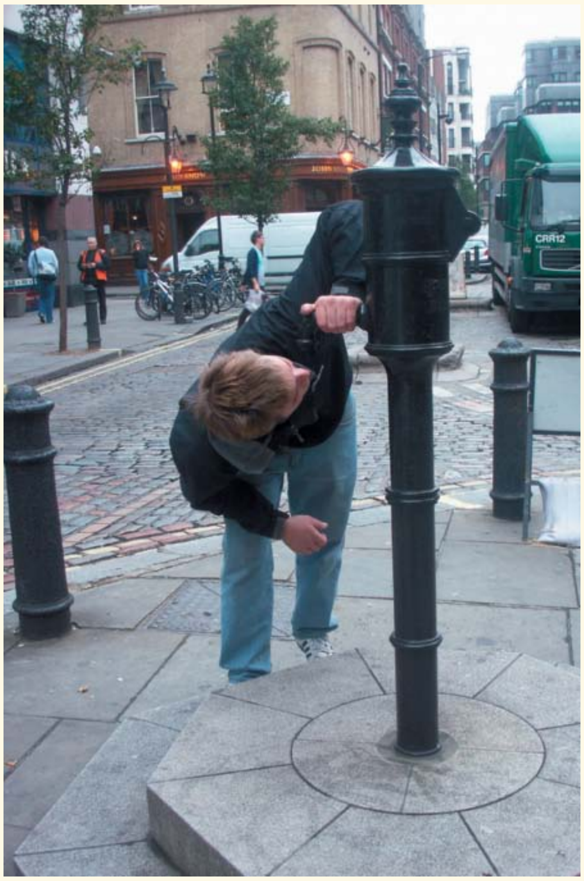


Figure 13.3 A modern replica of the pump that led Snow to the inference that drinking water transmitted cholera, located in what is now Broadwick Street in Soho, London.

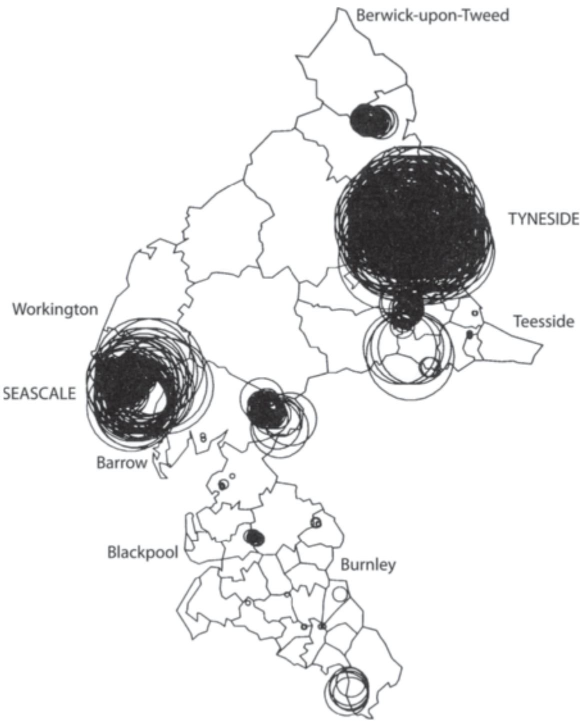
It is interesting to speculate on what would have happened if early epidemiologists like Snow had had access to a modern GI system. The rules governing research today would not have allowed Snow to remove the pump handle, except after lengthy review, because the removal constituted an experiment on human subjects. To get approval, he would have had to have shown persuasive evidence in favor of his hypothesis, and it is doubtful that the map would have been sufficient because several other hypotheses might have explained the pattern equally well. First, it is conceivable that the population of Soho was inherently at risk of cholera, perhaps by being comparatively elderly or because of poor housing conditions. The map would have been more convincing if it had *normalized* the data by showing the *rate* of incidence relative to the population at risk. For example, if cholera was highest among the elderly, the map could have shown the number of cases in each small area of Soho as a proportion of the population over 50 in each area. Second, it is still conceivable that the hypothesis of transmission through the air between carriers could have produced the same observed pattern, if the first carrier happened to live in the center of the outbreak near the pump, though requiring a coincidence makes this hypothesis less attractive. Snow could have eliminated this alternative if he had been able to produce a sequence of maps, showing the locations of cases as the outbreak developed. Both of these options involve simple spatial analysis of the kind that is readily available today in GI systems. He might also have cited the scientific principle known as *Occam's Razor*, which favors accepting the simplest hypothesis when more than one is available.

GI systems provide tools that are far more powerful than the map in suggesting causes of disease.

Today the causal mechanisms of diseases like cholera, which results in short, concentrated outbreaks, have long since been worked out. Much more problematic are the causal mechanisms of diseases that are rare and not so sharply concentrated in space and time. The now-classic work of Stan Openshaw at the University of Leeds, using one of his Geographical Analysis Machines, illustrates the kinds of applications that make good use of the power of GI systems in a more contemporary context.

Figure 13.4 shows an application of one of Openshaw's techniques to a comparatively rare but devastating disease whose causal mechanisms remain largely a mystery—childhood leukemia. The study area is northern England, the region from the Mersey to the Tyne Rivers. The analysis begins with two datasets: one of the locations of cases of the disease and the other of the numbers of people at risk in standard census reporting zones. Openshaw's technique then generates a large number of circles, of random sizes, and places them randomly over the map. The computer generates and places the circles and then analyzes their contents by dividing the number of cases found in the circle by the size of the population at risk within the circle. If the ratio is anomalously high, the circle is drawn. After a large number of circles have been generated, and a small proportion have been drawn, a pattern emerges. Two large concentrations, or clusters of cases, are evident in the figure. The one on the left is located around Sellafield, the location of the British Nuclear Fuels processing plant and a site of various kinds of past leaks of radioactive material. The other, in the upper right, is in the Tyneside region, and Openshaw and his colleagues discuss possible local causes.

Both of these examples are instances of the use of spatial analysis for scientific discovery and decision making, specifically for public health.



Source: Openshaw S., Charlton M., Wymer C., and Craft A. 1987. A Mark I geographical analysis machine for the automated analysis of point datasets. *International Journal of Geographical Information Systems* 1: 335–358.

Figure 13.4 The map made by Openshaw and colleagues by applying their Geographical Analysis Machine to the incidence of childhood leukemia in northern England. A very large number of circles of random sizes is randomly placed on the map, and a circle is drawn if the number of cases it encloses substantially exceeds the number expected in that area given the size of its population at risk.

Box 13.2 is devoted to Sara McLafferty, who teaches and researches the use of GI systems in public health.

Sometimes spatial analysis is used *inductively*, to examine empirical evidence in the search for patterns that might support new theories or general principles, in this case with regard to disease causation (we suggested earlier that the term *data mining* is often used in this context). Inductive reasoning is often identified as characteristic of Big Data, as we noted in Section 1.4. Other uses of spatial analysis are *deductive*, focusing on the testing of known theories or principles against data. (Snow already had a theory of how cholera was transmitted and used the map as powerful confirmation to convince others.) Induction and deduction were also discussed in Section 2.7 in the context of the nature of spatial data. A third type of application is *normative*, using spatial analysis to develop or prescribe new or better designs, for the locations of new retail stores, or new roads, or a new manufacturing plant. Examples of this type appear in Section 14.4.

Biographical Box 13.2

Sara McLafferty

Sara McLafferty (Figure 13.5) is Professor of Geography and Geographic Information Science at the University of Illinois at Urbana-Champaign. Throughout her career she has been fascinated by the question: How do the places where people work and interact affect their health and well-being? Where we live influences our exposure to environmental pollution; our access to health services, well-paying jobs, parks, and recreational spaces; our exposure to crime and violence; and our ability to have supportive social interactions. Two decades ago, Sara and her students began using GI systems and spatial analysis methods to better understand the connections between place environments and health. An early study looked at the spatial clustering of breast-cancer cases in the town of West Islip in Long Island, NY, and analyzed whether spatial clusters were in close proximity to certain environmental hazards. Community members had an important role in the project, raising concerns about environmental factors underlying the high incidence of breast cancer that Sara's team investigated using GI systems. As the field of GI systems and health began to take off during the 1990s, Sara and Ellen Cromley decided to write a book describing the dynamic new field of GI systems and health. Published in 2002, with a second edition in 2011, their book, *GIS and Public Health*, presents a foundation for the field and discusses applications of GI systems in exploring the determinants of health and planning health interventions in fields like infectious disease control, environmental health, and access to health services. Public-health researchers and planners from across the globe have adopted the book.

Courtesy: Sara McLafferty

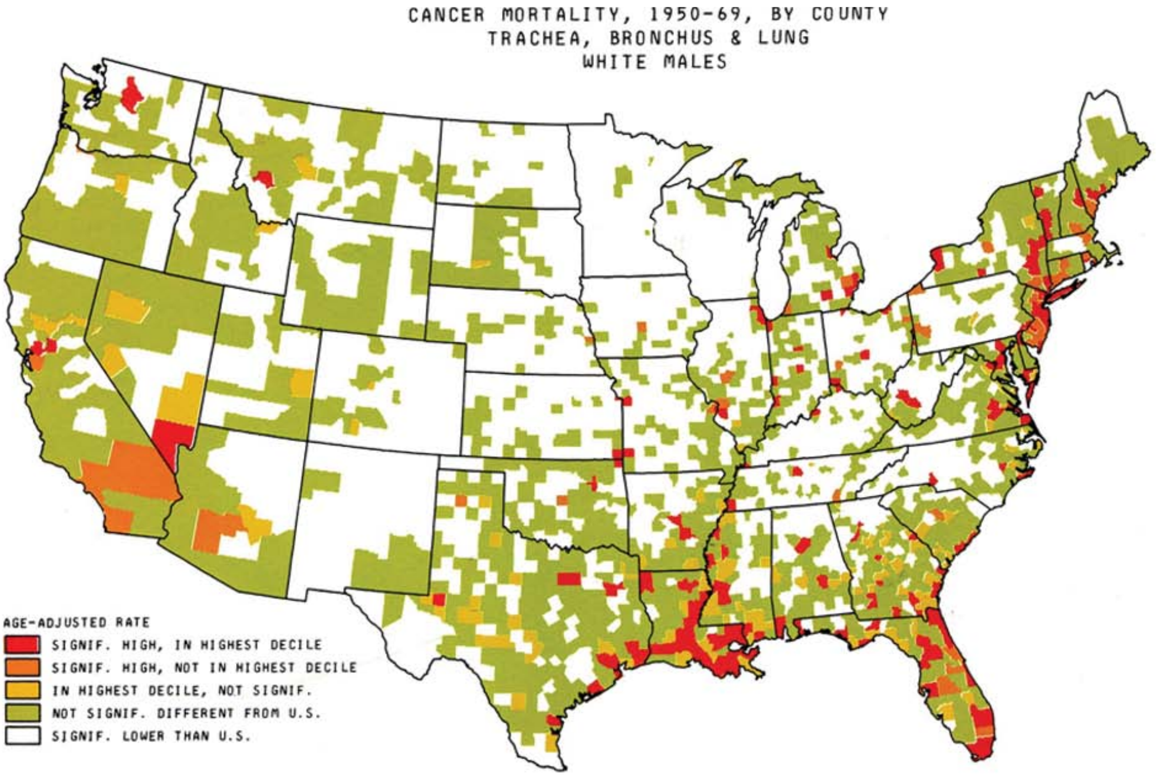
Figure 13.5 Sara McLafferty.

13.2 Analysis Based on Location

The concept of location—identifying *where* something exists or happens—is central to GISS, and the ability to compare different properties of the same place, and as a result to discover relationships and correlations and perhaps even explanations, is often presented as the field's greatest advantage. Take Figure 13.6 as an example. It shows the age-adjusted rate of death in the United States due to cancers of the throat and lung among adult males in a 20-year period 1950–1969, compiled by county. Maps such as this immediately prompt us to ask: Do I know of other properties of these counties that might explain their rates? Within a fraction of a second the mind is busy identifying counties, recalling general knowledge that might suggest cause, and checking other counties to see if they confirm suspicions. All of this depends, of course, on having a basis of knowledge (Section 1.2) in one's mind that is sufficient to identify particular counties and their general characteristics. One might note, for example, that Chicago, Detroit, Cleveland, and several other major cities are *hotspots* where cancer rates are significantly high, but that other major cities such as Minneapolis are not—and ask why? It is important to note that rather than the *rate* of cancer as a proportion of some estimate of the total population at risk, the map highlights counties where the rate is *statistically significant* according to standard statistical tests. To reach this threshold, a county needs both a high rate and a large population because rates in counties with small populations can be extreme just by chance. But in some counties, notably Silver Bow County in Montana, the home of Butte and a massive copper-mining operation, the rate in that period was so high as to be significant even within a comparatively small population.

GI systems allow us to look for explanations among the different properties of a location.

The map shows many intriguing patterns, but the pattern that is of greatest significance to the history of cancer research concerns counties distributed around the Pacific, Gulf, and Atlantic coasts. These were counties involved in ship construction during World War II, when large amounts of asbestos were used for insulation. In the period 5 to 25 years later, the consequences of


CANCER MORTALITY, 1950-69, BY COUNTY
TRACHEA, BRONCHUS & LUNG
WHITE MALES
AGE-ADJUSTED RATE
🟥 SIGNIF. HIGH, IN HIGHEST DECILE
🟧 SIGNIF. HIGH, NOT IN HIGHEST DECILE
🟨 IN HIGHEST DECILE, NOT SIGNIF.
🟩 NOT SIGNIF. DIFFERENT FROM U.S.
□ SIGNIF. LOWER THAN U.S.

Source: Mason et al., *Atlas of Cancer Mortality for U.S. Counties.* Bethesda, MD: National Cancer Institute, 1975

Figure 13.6 Age-adjusted rates of mortality due to cancers of the trachea, bronchus, and lung, among white males between 1950 and 1969, by county.

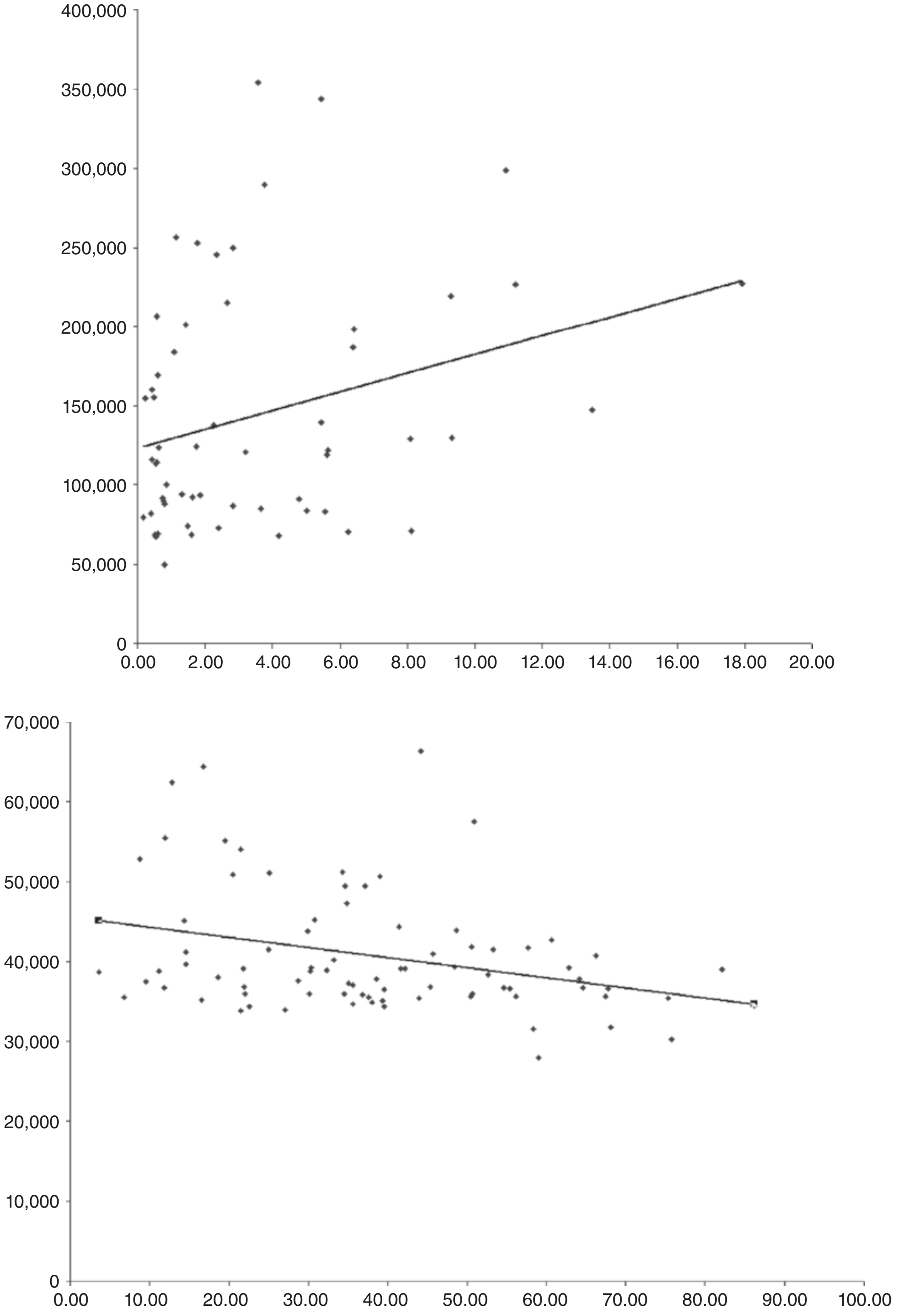
breathing asbestos fibers into the respiratory system are plain to see in the high rates of cancers in the counties containing Mobile, Alabama; Norfolk, Virginia; Jacksonville, Florida; and many other port cities.

It may seem that comparing the properties of places should be straightforward in a technology that insists on giving locations to all the information it stores. The next subsection discusses examples of such cases. In other cases, however, comparison can be quite difficult and complex. The data models adopted in GI systems and described in Chapter 8 are designed primarily to achieve efficiency in representation and to emphasize storing *information about one property for all places* over *information about all properties for one place*. We sometimes term this *horizontal* rather than *vertical* integration of data. For example, it is traditional to store all the elevation data for a given county together, perhaps as a digital elevation model, and all the soil data for the same county together, perhaps as a set of topologically related soil polygons. These are very efficient approaches, but they are not designed for a point-by-point comparison of elevation and soil type, or for answering questions such as "Are certain soil types more likely to be found at high elevations in this county?" Subsequent subsections discuss some of the GI system techniques designed specifically for situations such as this.

13.2.1 Analysis of Attribute Tables

In the example shown in Figure 13.6, it is quite likely that the kinds of factors responsible for high rates of cancer are already available in the attribute table of the counties, along with the cancer rates. In such cases our interest is in comparing the contents of two columns of the table, looking for possible relationships or correlations—are there counties for which cancer rates and the values of potentially causative variables are both high, or both low? Figure 13.7 shows a suitable example, where the interest lies in a possible relationship between two columns of a county attribute table. In this case the investigator suspects a pattern in the relationship between average value of house and percent black, variables that are collected and disseminated by the U.S. Bureau of the Census as county attributes. One way to examine this suspicion is to plot one variable against the other as a *scatterplot*. In Figure 13.7 median house value is shown on the vertical or *y*-axis, and percent black on the horizontal or *x*-axis.

In a formal statistical sense, these scatterplots allow us to examine in detail the *dependence* of one variable on one or more *independent* variables. For example, we might hypothesize that the median value of houses in a county is correlated with a number of variables such



Source: U.S. Bureau of the Census

Figure 13.7 Scatterplots of median house value (*y*-axis) versus percent black (*x*-axis) for U.S. counties in 1990, with linear regressions: (A) California; and (B) Mississippi.

as percent black, average county income, percent 65 and over, average county air pollution levels, and so forth, all stored in the attribute table associated with the properties. Formally this may be written as:

$$Y = f(X_1, X_2, X_3, \ldots, X_k)$$

where Y is the dependent variable and X_1 through X_k are all the possible independent variables that might affect housing value or be correlated with it. It is important to note that it is the independent variables that together predict the dependent variable, and that any hypothesized causal relationship is one way—that is, that median house value is *responsive* to average income, percent 65 and over, and so forth, and not vice versa. For this reason the dependent variable is termed the *response* variable, and the independent variables are termed *predictor* variables in some statistics textbooks.

In our case there is only a single predictor variable, and it is clear from the scatterplots that the relationship is far from perfect—that many other unidentified factors contribute to the determination of median housing value. In general this will always be true in the social and environmental sciences because it will never be possible to capture all the factors responsible for a given outcome. So we modify the model by adding an error term ε to represent that unknown variation.

Regression analysis is the term commonly used to describe this kind of investigation, and it focuses on finding the simplest relationship indicated by the data. That simplest relationship is linear, implying that a unit increase in percent black always corresponds to a constant corresponding increase or decrease in the dependent variable. Linear relationships plot as straight lines on a scatterplot and have the equation:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k + \varepsilon$$

though in the case of Figure 13.7 there is only one independent variable X_1. b_1 through b_k are termed regression *parameters* and measure the direction and strength of the influence of the independent variables X_1 through X_k on Y. b_0 is termed the *constant* or *intercept* term.

Figure 13.7A shows the data for the counties of California, and Figure 13.7B for the counties of Mississippi. Notice the difference—counties with a high percentage of blacks have *lower* housing values in Mississippi but *higher* housing values in California. In Figure 13.7 *best fit* lines have been drawn through the scatters of points. The gradient of this line is calculated as the b_1 parameter of the regression; it is positive in Figure 13.7A and negative in Figure 13.7B, indicating positive and negative trends, respectively. The value where the regression line intersects the y-axis identifies the (hypothetical) median housing value when percent black is zero and gives us the intercept value b_0. The more general multiple regression case works by extension of this principle, and each of the b parameters gauges the marginal effects of its respective X variable.

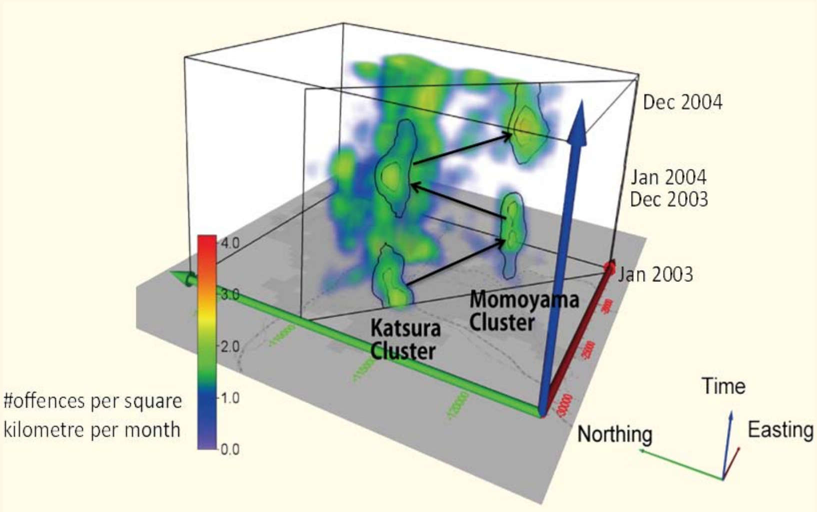
There are two lessons to be learned here. First, the relationship we have uncovered varies across space, from state to state, being an example of the general issue termed *spatial heterogeneity* that we discussed briefly in Section 2.2. Traditionally, science has concerned itself with finding patterns and relationships that exist everywhere—with *general* principles and laws that we described in Section 1.3 as *nomothetic*. In that sense what we have uncovered here is something different, a general principle (housing values are correlated with percent black) that varies in its specifics from state to state (a positive relationship in California, a negative one in Mississippi). Recently, geographers have developed a set of techniques that recognize such heterogeneity explicitly and focus not so much on what is true everywhere, but on how things vary across the geographic world. A prominent example is *geographically weighted regression* (GWR), a technique originally developed by Stewart Fotheringham, Martin Charlton, and Chris Brunsdon at the University of Newcastle-upon-Tyne. Rather than look for a single regression line, it examines how the slope and intercept vary across space in ways that may be related to other factors. Box 13.3 is devoted to Tomoki Nakaya, who among other contributions to spatial analysis has developed some of the widely used software for GWR.

The second lesson to be learned concerns the use of counties to unveil this relationship. Counties in the United States are particularly awkward units of observation because they vary enormously in size and population and mask variations in space that are often dramatic. The list of 58 counties of California, for example, includes one (Alpine) with a population of about 1,000 and another (Los Angeles) with a population of about 10 million. In Virginia individual cities are often their own counties and may be tiny in comparison with the size of San Bernardino County, California, the largest county in the continental United States. The lesson to be learned here is similar to that for English counties in Section 5.4.3: The results of this analysis depend intimately on the units of analysis chosen. Thus if we had repeated the analysis with other units, such as watersheds, our results might have been entirely different. California reverses the Mississippi trend in this case because the wealthy urban counties of California (San Francisco, Los Angeles, Alameda, etc.) are also the counties where most blacks live, whereas in Mississippi it is the rural counties with low housing values that house most blacks. But this pattern

Biographical Box 13.3

Tomoki Nakaya

Tomoki Nakaya is Professor of Geography at Ritsumeikan University in Japan and codirector of the Institute of Disaster Mitigation for Urban Cultural Heritage. He obtained his PhD from Tokyo Metropolitan University in 1997. His research specializes in spatial statistics and mathematical modeling in human geography, and he played a principal role in developing the software for fitting geographically weighted regression (GWR4 is available from the National Centre for Geocomputation in Ireland).



Source: Tomoki Nakaya

Figure 13.8 Alternating occurrence of bag-snatching clusters in a pair of areas in Kyoto.

One of his current research themes is the integration of spatial or space–time statistics and geovisualization, for better understanding of such phenomena as crime, health, and disaster hazards. A good example of such visualization based on space–time statistical analysis is found in the study with the local police department for epidemiological analysis on snatching crime in the city of Kyoto. Figure 13.8 shows the results of one of his analyses, in this case the space–time density of bag-snatching crimes in Kyoto. An outbreak of crimes in one area is followed by increased policing and citizen vigilance, displacing the outbreak to a new area. As policing and vigilance increases in the second area and relaxes in the first, crime returns to the first area—and the pattern repeats itself. Visualization of data in space and time can thus lead to new insights into criminal behavior, as well as providing a good storytelling tool.

Nakaya published the first book on applications of GI systems in spatial epidemiology in Japan in 2004, and since then has been involved in numerous related projects, such as GI systems-based research into cancer-registry data, focusing on social inequalities in the cancer burden at the local level. With his colleague Keiji Yano he also has led the Virtual Kyoto project, which collects and stores detailed 2-D and 3-D geographical data about the city of Kyoto in different historical periods.

might not hold at all if we were to use watersheds rather than counties as the units of analysis. This issue is known in general as the *modifiable areal unit problem* and is also discussed in Section 5.4.3.

Earlier we described spatial analysis as a set of techniques whose results depend on the locations of the objects of study. But note that in this case the locations of the counties do not affect the results, and latitude and longitude nowhere play a role. Locations are useful for making maps of the inputs, but this analysis hardly rates the term *spatial*. In fact, many types of statistical analysis can be applied to the contents of attribute tables, without ever needing to know the locations or geometric shapes of features. On the other hand space can enter into models such as the ones discussed in this section if the researcher feels that some of the influences on *Y* come from the values of variables in nearby areas rather than from variables measured in the same area. For example, we might believe that housing values in county *i* are influenced by conditions both in county *i* and also in neighboring counties. This kind of model is explicitly spatial, embodying the concept of influence *at a distance*, and calls for a special kind of regression termed *spatial regression* (a technique strongly related to GWR) in which values of independent variables in nearby areas are added to the model. A full discussion of these methods is outside the scope of this chapter, but can be found in several of the books listed in the Further Reading section.

13.2.2 Spatial Joins

In the previous section the variables needed for the analysis were all present in the same attribute table. Often, however, it is necessary to perform some basic operations first to bring the relevant variables together; recall the earlier discussion of how GI systems favor horizontal over vertical integration. Suppose, for example, that we wish to conduct a nationwide analysis of the average income of major cities and have a GI database containing cities as points, together with some relevant attributes of those cities: the dependent variable average income, plus some independent variables such as percent with college degrees and percent retired people. But other variables that are potentially relevant to the analysis are available only at the state level, including the state's percent unemployed. In Section 9.3 we described a *relational join* or simply a *join*—a fundamental operation in databases that is used to combine the contents of two tables using a common key. In this case the key is "state," a variable that exists in each city record to indicate the state containing the city, and in each state record as an identifier. The result of the join will be that each city record now includes the attributes of the city's containing state, allowing us to add the state-level variables to the analysis. Figure 9.2 shows a simple example of a join. One of the most powerful features of a GI system is the ability to join tables in this way based on common geographic location. To return to the point made at the outset of this discussion of location, we have now achieved what GI systems have always promised—vertical integration, or the ability to link disparate information based on location.

But this example is simple because common location is represented here by a key indicating the state containing the city. This spatial relationship was explicitly coded in the city attribute table in this case, but in other cases it will need to be computed using the functions of a GI system. The next two subsections describe circumstances in which the performance of a spatial join is a much more complex operation.

13.2.3 The Point-in-Polygon Operation

Comparing the properties of points to those of containing areas is a common operation in a GI system. It occurs in many areas of social science, when an investigator is interested in the extent to which the behavior of an individual is determined by properties of the individual's neighborhood. It occurs when point-like events, such as instances of a disease, must be compared to properties of the surrounding environment. In other applications of GI systems it occurs when a company needs to determine the property on which an asset such as an oil well or a transformer lies. In its simplest form, the point-in-polygon operation determines whether a given point lies inside or outside a given polygon. In more elaborate forms there may be many polygons, and many points, and the task is to assign points to polygons. If the polygons overlap, it is possible that a given point lies in one, many, or no polygons, depending on its location. Figure 13.9 illustrates the task.

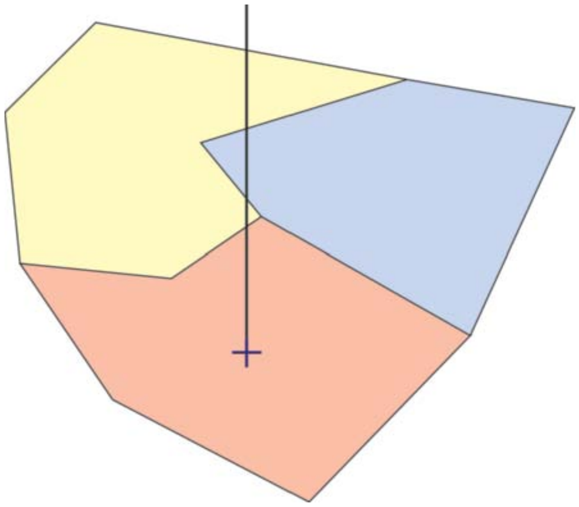


Figure 13.9 The point-in-polygon problem, shown in the continuous-field case (the point must by definition lie in exactly one polygon, or outside the project area). In only one instance (the orange polygon) is there an odd number of intersections between the polygon boundary and a line drawn vertically upward from the point.

The point-in-polygon operation makes sense from both the discrete-object and the continuous-field perspectives (see Section 3.5 for a discussion of these two perspectives). From a discrete-object perspective both points and polygons are objects, and the task is simply to determine enclosure. From a continuous-field perspective, polygons representing a variable such as land ownership cannot overlap because each polygon represents the land owned by one owner, and overlap would imply that a point is owned simultaneously by two owners. Similarly from a continuous-field perspective there can be no gaps between polygons. Consequently, the result of a point-in-polygon operation from a continuous-field perspective must assign each point to exactly one polygon.

The point-in-polygon operation is used to determine whether a point lies inside or outside a polygon.

Although the actual methods used by programmers to perform standard GI system operations are not normally addressed in this book, the standard approach or *algorithm* for the point-in-polygon operation is sufficiently simple and interesting to merit a short description. In essence, it consists of drawing a line from the point to infinity (see Figure 13.9), in this case parallel to the *y*-axis, and determining the number of intersections between the line and the polygon's boundary. If the number is odd, the point is inside the polygon, and if it is even, the point is outside. The algorithm must deal successfully with special cases—for example, if the point lies directly below a vertex (corner point) of the polygon. Some algorithms extend the task to include a third option, when the point lies exactly on the boundary. But others ignore this, on the grounds that it is never possible in practice to determine location with perfect accuracy, and so it is never possible to determine whether an infinitely small point lies on or off an infinitely thin boundary line.

13.2.4 Polygon Overlay

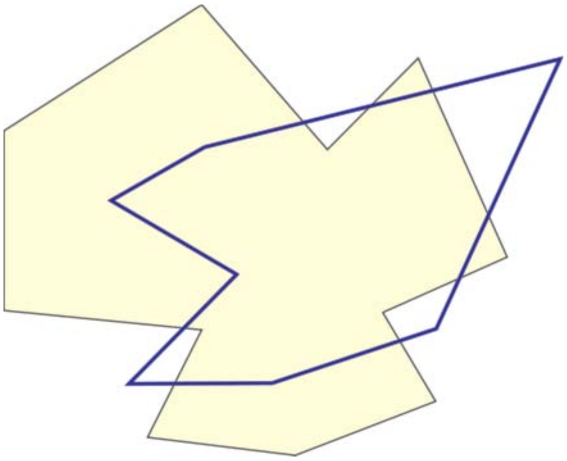
Polygon overlay is similar to the point-in-polygon operation in the sense that two sets of objects are involved, but in this case both are polygons. It again exists in two forms, depending on whether a continuous-field or discrete-object perspective is taken. The development of effective algorithms for polygon overlay was one of the most significant challenges of early GI systems, and the task remains one of the most complex and difficult to program.

The complexity of computing a polygon overlay was one of the greatest barriers to the development of vector GI systems.

From the discrete-object perspective, the task is to determine whether two area objects overlap, to determine the area of overlap, and to define the area formed by the overlap as one or more new area objects (the overlay of two polygons can produce a large number of distinct area objects; see Figure 13.10). This operation is useful for determining answers to such queries as:

- How much of this proposed clear-cut lies in this riparian zone?
- How much of the projected catchment area of this proposed retail store lies in the catchment of this other existing store in the same chain?
- How much of this land parcel is affected by this easement?

Figure 13.10 Polygon overlay, in the discrete-object case. Here the overlay of two polygons produces nine distinct polygons. One has the properties of both polygons, four have the properties of the yellow shaded polygon but not the blue (bounded) polygon, and four are outside the yellow polygon but inside the blue polygon.



- What proportion of the land area of the United States lies in areas managed by the Bureau of Land Management?

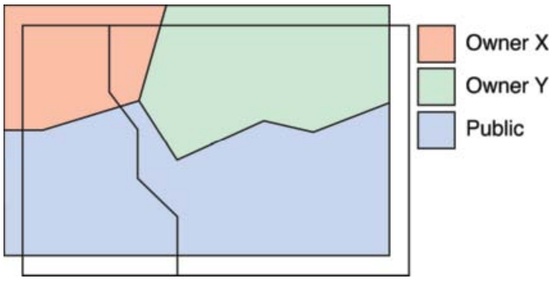
From the continuous-field perspective the task is somewhat different. Figure 13.11 shows two datasets, both of which are representations of fields: one differentiates areas according to land ownership, and the other differentiates the same region according to land-cover class. In the terminology of Esri's ArcGIS, both datasets are instances of *area coverages*, or fields of nominal variables represented by nonoverlapping polygons. The methods discussed earlier in this chapter could be used to interrogate either dataset separately, but there are numerous queries that require simultaneous access to both datasets. For example:

- What is the land-cover class, and who is the owner of the point indicated by the user?
- What is the total area of land owned by X and with land cover class A?
- Where are the areas that lie on publicly owned land and have land-cover class B?

None of these queries can be answered by interrogating one of the datasets alone—the datasets must somehow be combined (vertically integrated) so that interrogation can be directed simultaneously at both of them.

The continuous-field version of polygon overlay does this by first computing a new dataset in which the region is partitioned into smaller areas that have uniform characteristics on both variables. Each area in the new dataset will have two sets of attributes—those obtained from one of the input datasets and those obtained from the other. In effect, then, we will have performed a spatial join by creating a new table that combines both sets of attributes, though in this case

Figure 13.11 Polygon overlay in the continuous-field case. Here a dataset representing two types of land cover (A on the left, B on the right) is overlaid on a dataset representing three types of ownership (the two datasets have been offset for visual clarity). The result will be a single dataset in which every point is identified with one land cover type and one ownership type. It will have five polygons because land cover A intersects with two ownership types, and land cover B intersects with three.



we will also have created a new set of features. All the boundaries will be retained, but they will be broken into shorter fragments by the intersections that occur between boundaries in one input dataset and boundaries in the other. Note the unusual characteristics of the new dataset shown in Figure 13.11. Unlike the two input datasets, where boundaries meet in junctions of three lines, the new map contains a new junction of four lines, formed by the new intersection discovered during the overlay process. Because the results of overlay are distinct in this way, it is almost always possible to discover whether a geographically referenced dataset was formed by overlaying two earlier datasets.

Polygon overlay has different meanings from the continuous-field and discrete-object perspectives.

With a single dataset that combines both inputs, it is an easy matter to answer all the queries listed earlier through simple interrogation, or to look for relationships between the attributes. It is also easy to reverse the overlay process. If neighboring areas that share the same land-cover class are merged, for example, the result is the land-ownership map, and vice versa.

Polygon overlay is a computationally complex operation, and as noted earlier, much work has gone into developing algorithms that function efficiently for large datasets. One of the issues that must be tackled by a practically useful algorithm is known as the *spurious polygon* or *coastline weave* problem. It is almost inevitable that there will be instances in any practical application where the same line on the ground occurs in both datasets. This happens, for example, when a coastal region is being analyzed because the coastline is almost certain to appear in every dataset of the region. Rivers and roads often form boundaries in many different datasets—a river may function both as a land-cover-class boundary and as a land-ownership boundary, for example. But although the same line is represented in both datasets, its representations will almost certainly not be the same. They may have been digitized from different maps, digitized using different numbers of points, subjected to different manipulations, obtained from entirely different sources (an air photograph and a topographic map, for example), and subjected to different measurement errors. When overlaid, the result is a series of small slivers. Paradoxically, the more care one takes in digitizing or processing, the worse the problem appears, as the result is simply more slivers, albeit smaller in size.

In two vector datasets of the same area, there will almost certainly be instances where lines in each dataset represent the same feature on the ground.

Table 13.1 shows an example of the consequences of slivers and how a GI system can be rapidly overwhelmed if it fails to anticipate and deal with them adequately. Today, a GI system will offer various methods for dealing with the problem, the most common of which is the specification of a *tolerance*. If two

Table 13.1 Numbers of polygons resulting from an overlay of five datasets, illustrating the spurious polygon problem. The datasets come from the Canada Geographic Information System discussed in Section 1.5.1, and all are representations of continuous fields. Dataset 1 is a representation of a map of soil capability for agriculture, Datasets 2 through 4 are land-use maps of the same area at different times (the probability of finding the same real boundary in more than one such map is very high), and Dataset 5 is a map of land capability for recreation. The final three columns show the numbers of polygons in overlays of three, four, and five of the input datasets.

Acres	1	2	3	4	5	1+2+5	1+2+3+5	1+2+3+4+5
0–1	0	0	0	1	2	2,640	27,566	77,346
1–5	0	165	182	131	31	2,195	7,521	7,330
5–10	5	498	515	408	10	1,421	2,108	2,201
10–25	1	784	775	688	38	1,590	2,106	2,129
25–50	4	353	373	382	61	801	853	827
50–100	9	238	249	232	64	462	462	413
100–200	12	155	152	158	72	248	208	197
200–500	21	71	83	89	92	133	105	99
500–1,000	9	32	31	33	56	39	34	34
1,000–5,000	19	25	27	21	50	27	24	22
>5,000	8	6	7	6	11	2	1	1
Totals	88	2,327	2,394	2,149	487	9,558	39,188	90,599

lines fall within this distance of each other, the GI system will treat them as a single line, and not create slivers (see also Section 8.3.2.2). The resulting overlay contains just one version of the line, not two. But at least one of the input lines has been moved, and if the tolerance is set too high, the movement can be substantial and can lead to problems later.

13.2.5 Raster Analysis

Many of the complications addressed in the previous subsections disappear if the data are structured in raster form and if the cells in each layer of data are geometrically identical. For example, suppose we are interested in the agricultural productivity of land and have data in the form of a raster, each 10 m by 10 m cell giving the average annual yield of corn in the cell. We might investigate the degree to which these yield values are predictable from other properties of each cell, using rasters of fertilizer quantity applied, depth to water table, percent organic matter, and so forth, each provided for the same set of 10 m by 10 m cells.

In such cases there is no need for complicated overlay operations because all the attributes are already available for the same set of spatial features, the cells of the raster. Overlay in raster is thus an altogether simpler operation, and this has often been cited as a good reason to adopt raster rather than vector structures. Attributes from different rasters can be readily combined for a variety of purposes, as long as the rasters consist of identically defined arrays of cells. The power of this raster-based approach is such that it deserves its own section; it is discussed in Section 15.2.4 under the heading "Cartographic Modeling and Map Algebra."

In other cases, however, the different variables needed for an analysis may not use identical rasters. For example, it may be necessary to compare data derived from the AVHRR (Advanced Very High Resolution Radiometer) sensor with a cell size of 1 km by 1 km with other data derived from the MODIS (Moderate Resolution Imaging Spectroradiometer) sensor with a cell size of 250 m by 250 m—and the two rasters will also likely be at different orientations and may even use different map projections to flatten the Earth (see Section 8.2.1 for a discussion of satellite remote sensing and Section 4.8 for a discussion of map projections). In such cases it is necessary to employ some form of *resampling* to transform each dataset to a common raster. Box 13.4 shows a simple example of the use of resampling to join point data. Resampling is a form of spatial interpolation, a technique discussed at length in Section 13.3.6.

Application Box 13.4

Comparing Attributes When Points Do Not Coincide Spatially

GI system users often encounter situations in which attributes must be compared, but for different sets of features. Figure 13.12 shows a study area with two sets of points, one being the locations where levels of ambient sound were measured using recorders mounted on telephone poles, the other the locations of interviews conducted with local residents to determine attitudes to noise. We would like to know about the relationship between sound and attitudes, but the locations and numbers of cases (the *spatial supports*) are different. A simple expedient is to use spatial interpolation (Section 13.3.6), a form of intelligent guesswork that provides estimates of the values of continuous fields at locations where no measurements have been taken. In other words, it *resamples* each field at different points. Given such a method it would be possible to conduct the analysis in any of three ways:

- By interpolating the second dataset to the locations at which the first attribute was measured
- By the reverse—interpolating the first dataset to the locations at which the second attribute was measured
- By interpolating both datasets to a common geometric base, such as a raster

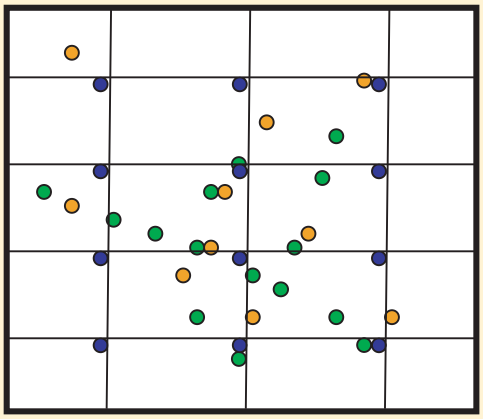


Figure 13.12 Coping with the comparison of two sets of attributes when the respective objects do not coincide. In this instance, attitudes regarding ambient noise have been obtained through a household survey of 15 residents (green dots) and are to be compared to ambient noise levels measured at 10 observation points (brown dots). The solution is to interpolate both sets of data to 12 comparison points (blue dots), using the methods discussed in Section 13.3.6.

▶

In the third case, note that it is possible to create a vast amount of data by using a sufficiently detailed raster. In essence, all these options involve manufacturing information, and the results will depend to some extent on the nature and suitability of the method used to do the spatial interpolation. We normally think of a relationship that is demonstrated with a large number of observations as stronger and more convincing than one based on a small number of observations. But the ability to manufacture data upsets this standard view. The implications of this power of GI systems to manufacture data are addressed in Section 14.5.

Figure 13.12 shows a possible solution using this third option. The number of grid points has been determined by the smaller number of cases—in this case, approximately 10 (12 are shown)—to minimize concerns about manufacturing information.

13.3 Analysis Based on Distance

The second fundamental spatial concept considered in this chapter is *distance*, the separation of places on the Earth's surface. The ability to calculate and manipulate distances underlies many forms of spatial analysis, some of the most important of which are reviewed in this section. All are based on the concept that the separation of features or events on the Earth's surface can tell us something useful, either about the mechanisms responsible for their presence or properties—in other words, to *explain* their patterns—or as input to decision-making processes. The first subsection looks at the measurement of distance and length in a GI system, as well as some of the issues involved. The second discusses the construction of buffers, together with their use in a wide range of applications. The identification of an anomalous concentration of points in space was the trigger that led to Dr. Snow's work on the causes of cholera transmission, and it is discussed in its general case as *cluster detection* in the third subsection. The concept of *spatial dependence* or *dependence at a distance*, first introduced in Chapter 2 as a fundamental property of geographic data, is given operational meaning in the fourth subsection. The fifth subsection addresses *density estimation*, based on the concept of averaging over defined distances, and the final subsection discusses the related distance-based operation of *spatial interpolation*.

13.3.1 Measuring Distance and Length

A *metric* is a rule for determining distance between points in a space. Several kinds of metrics are used in GI systems, depending on the application. The simplest is the rule for determining the shortest distance between two points in a flat plane, called the Pythagorean or straight-line metric. If the two points are defined by the coordinates (x_1,y_1) and (x_2,y_2), then the distance D between them is the length of the hypotenuse of a right-angled triangle (Figure 13.13). Pythagoras's theorem tells us that the square of this length is equal to the sum of the squares of the lengths of the other two sides. So a simple formula results:

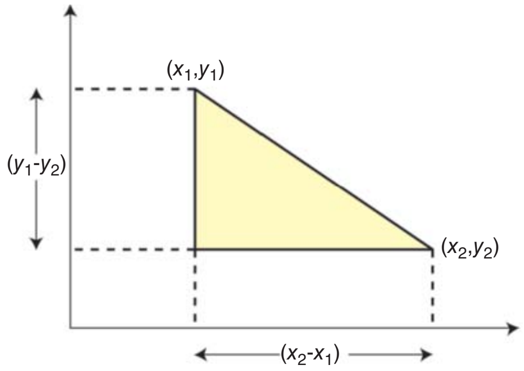
$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

A metric is a rule for determining the distance between points in space.

The Pythagorean metric gives a simple and straightforward solution for a plane, if the coordinates x and y are comparable, as they are in any coordinate system based on a conformal projection, such as the Universal Transverse Mercator (UTM) or Web Mercator (see Chapter 4). But the metric will not work for latitude and longitude, reflecting a common source of problems in GI systems—the temptation to treat latitude and longitude as if they were equivalent to plane coordinates. This issue is discussed in detail in Section 4.8.1.

For points widely separated on the curved surface of the Earth the assumption of a flat plane leads to

Figure 13.13 Pythagoras's Theorem and the straight-line distance between two points on a plane. The square of the length of the hypotenuse is equal to the sum of the squares of the lengths of the other two sides of the right-angled triangle.



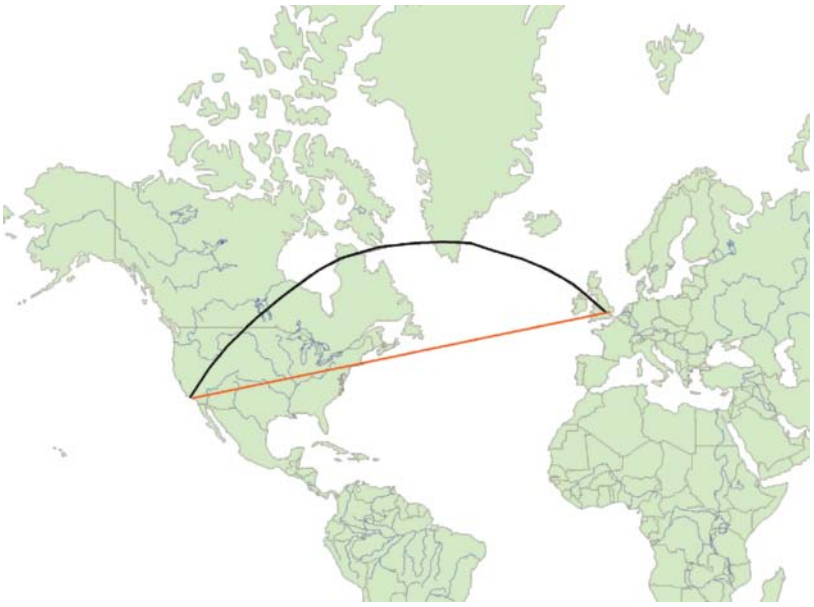


Figure 13.14 The effects of the Earth's curvature on the measurement of distance, and the choice of shortest paths. The map shows the North Atlantic on the Mercator projection. The red line shows the track made by steering a constant course of 79 degrees from Los Angeles and is 9807 km long. The shortest path from Los Angeles to London is actually the black line, the trace of the great circle connecting them, with a length of roughly 8800 km. This is typically the route followed by aircraft flying from London to Los Angeles. Flying in the other direction, aircraft may sometimes follow other, longer tracks, such as the red line, if by doing so they can take advantage of jet stream winds.

significant distortion, and distance must be measured using the metric for a spherical Earth given in Section 4.6 and based on a great circle. For some purposes even this is not sufficiently accurate because of the nonspherical nature of the Earth, and even more complex procedures must be used to estimate distance that take nonsphericity into account. Figure 13.14 shows an example of the differences that the curved surface of the Earth makes when flying long distances.

In many applications the simple rules—the Pythagorean and great-circle equations—are not sufficiently accurate estimates of actual travel distance, and we are forced to resort to summing the actual lengths of travel routes. In GI systems this normally means summing the lengths of links in a network representation, and many forms of spatial analysis use this approach, along with Web services for obtaining driving directions. If a line is represented as a polyline, or a series of straight segments, then its length is simply the sum of the lengths of each segment, and each segment length can be calculated using the Pythagorean formula and the coordinates of its endpoints. But it is worth being aware of two problems with this simple approach.

First, a polyline is often only a rough version of the true object's geometry. A river, for example, never makes the sudden changes of direction of a polyline, and Figure 13.15 shows how smoothly curving streets

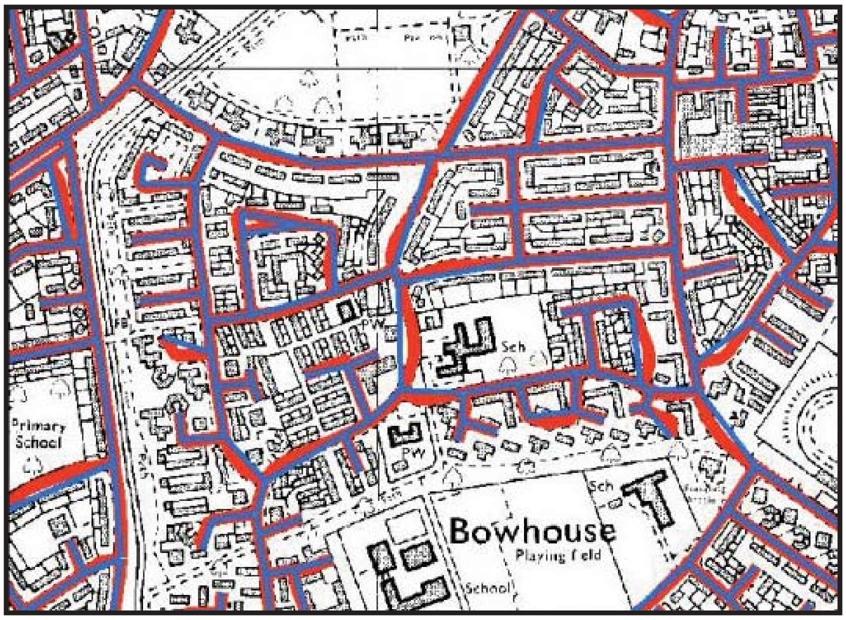


Figure 13.15 The polyline representations of smooth curves tend to be shorter in length, as illustrated by this street map (note how curves are replaced by straight-line segments). But estimates of area tend not to show systematic bias because the effects of overshoots and undershoots tend to cancel out to some extent.

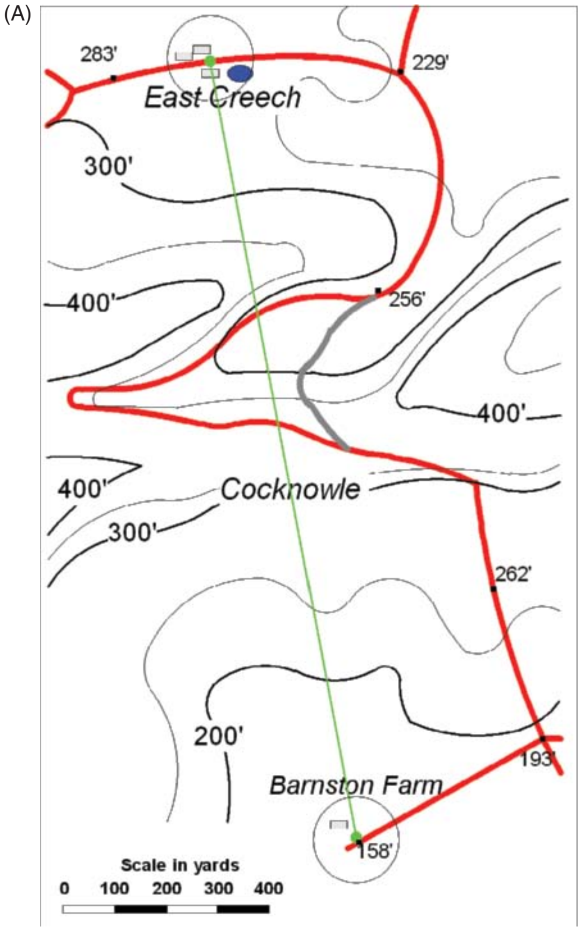
have to be approximated by the sharp corners of a polyline. Because there is a general tendency for polylines to shortcut corners, *the length of a polyline tends to be shorter than the length of the object it represents*. There are some exceptions, of course—surveyed boundaries are often truly straight between corner points, and streets are often truly straight between intersections. But in general the lengths of linear objects estimated in a GI system—and this includes the lengths of the perimeters of areas represented as polygons—are often substantially shorter than their counterparts on the ground (and note the discussion of fractals in Box 2.7). Note that this is not similarly true of area estimates because shortcutting corners tends to produce both underestimates and overestimates of area, and these tend to cancel out. So although estimates of line length tend to be systematically biased, estimates of area are not.

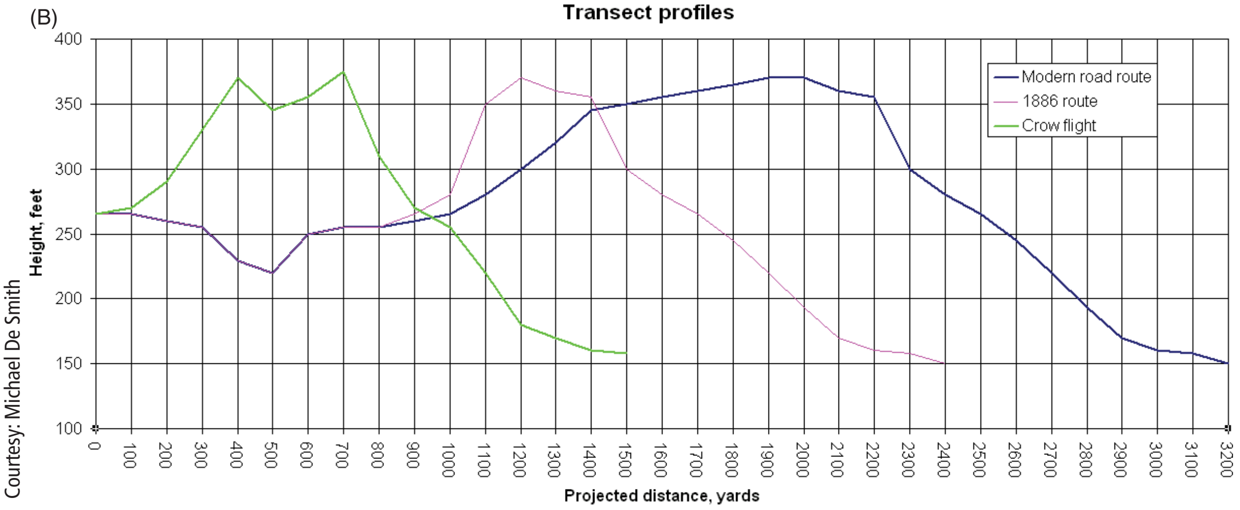
> **A GI system will almost always underestimate the true length of a geographic line.**

Second, the length of a line in a two-dimensional representation will always be the length of the line's planar projection, not its true length in three dimensions, and the difference can be substantial if the line is steep (Figure 13.16). In most jurisdictions the area of a parcel of land is the area of its horizontal projection, not its true surface area. A system that stores the third dimension for every point is able to calculate both versions of length and area, but not a system that stores only the two horizontal dimensions.

13.3.2 Buffering

One of the most important GI system operations available to the user is the calculation of a *buffer* (see also Section 9.5). Given any set of objects, which may include points, lines, or areas, a buffer operation builds a new object or objects by identifying all areas that are within a certain specified distance of the original objects. Figure 13.17 shows instances of a point,





Courtesy: Michael De Smith

Figure 13.16 The length of a path as traveled on the Earth's surface (red line) may be substantially longer than the length of its horizontal projection as evaluated in a two-dimensional GIS. (A) shows three paths across part of Dorset in the UK. The green path is the straight route, the red path is the modern road system, and the gray path represents the route followed by the road in 1886. (B) Shows the vertical profiles of all three routes, with elevation plotted against the distance traveled horizontally in each case. 1 ft = 0.3048 m, 1 yd = 0.9144 m.

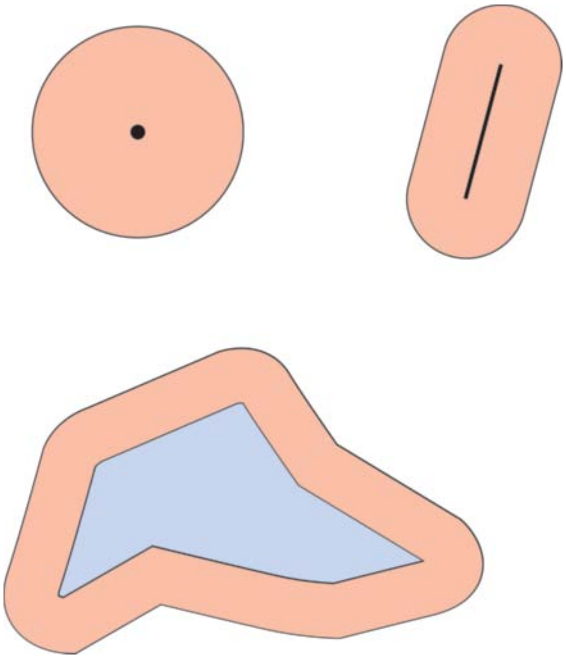


Figure 13.17 Buffers (dilations) of constant width drawn around a point, a polyline, and a polygon.

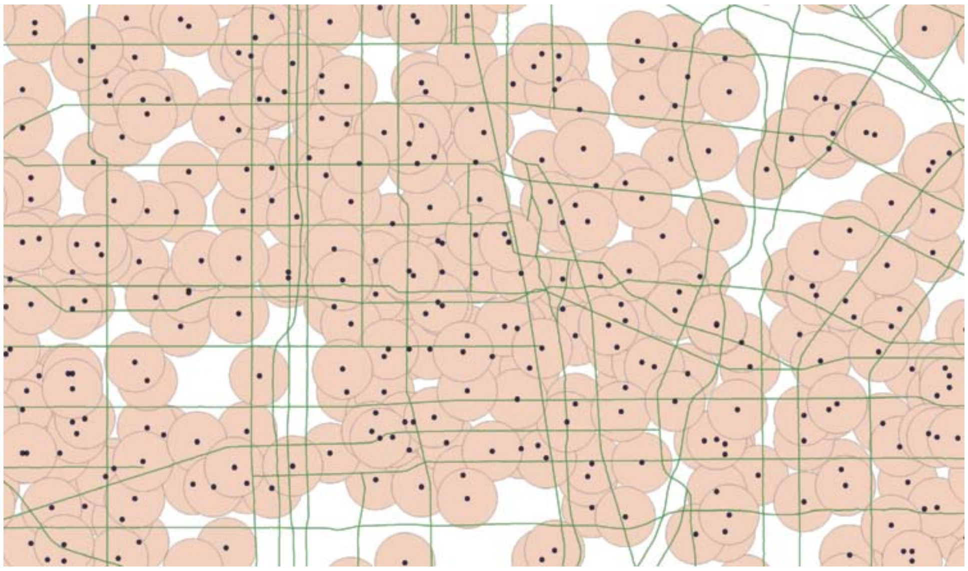
a line, and an area, as well as the results of buffering. Buffers have many uses, and they are among the most popular of GI system functions:

- The owner of a land parcel has applied for planning permission to rebuild—the local planning authority could build a buffer around the parcel in order to identify all homeowners who live within the legally mandated distance for notification of proposed redevelopments.
- A logging company wishes to clear-cut an area but is required to avoid cutting in areas within 100 m of streams; the company could build buffers 100 m wide around all streams to identify these protected riparian areas.
- A retailer is considering developing a new store on a site, of a type that is able to draw consumers from up to 4 km away from its stores. The retailer could build a buffer around the site to identify the number of consumers living within 4 km of the site, in order to estimate the new store's potential sales.
- A new law is passed requiring people once convicted of sexual offenses involving young children to live more than ½ mile (0.8 km) from any school. Figure 13.18 shows the implications of such a law for an area of Los Angeles. The buffers are based on point locations; if the school grounds had been represented as polygons the buffered areas would be larger.

Buffering is possible in both raster and vector GI systems. In the raster case, the result is the classification of cells according to whether they lie inside or outside the buffer, whereas the result in the vector case is a new set of objects (Figure 13.17). But there is an additional possibility in the raster case that makes buffering more useful in some situations. Rather than buffer according to distance, we can ask a raster GI system to *spread* outward from one or more features at rates determined by *friction*, *travel speed*, or *cost* values stored in each cell. This form of analysis is discussed in Section 14.3.2.

Buffering is one of the most useful transformations in a GI system and is possible in both raster and vector formats.

Figure 13.18 Buffers representing 1/2-mile exclusion zones around all schools in part of Los Angeles.



13.3.3 Cluster Detection

One of the questions most commonly asked about distributions of features, particularly point-like features, is whether they display a random pattern (see Figure 2.1), in the sense that all locations are equally likely to contain a point, or whether some locations are more likely than others—and particularly, whether the presence of one point makes other points either more or less likely in its immediate neighborhood. This leads to three possibilities:

- The pattern is *random* (points are located independently, and all locations are equally likely).
- The pattern is *clustered* (some locations are more likely than others, and the presence of one point may attract others to its vicinity).
- The pattern is *dispersed* (the presence of one point may make others less likely in its vicinity).

Establishing the existence of clusters is often of great interest because it may point to possible causal factors, as, for example, with the case of childhood leukemia studied by Stan Openshaw (Figure 13.4). Dispersed patterns are the typical result of competition for space, as each point establishes its own territory and excludes others. Thus such patterns are commonly found among organisms that exhibit territorial behavior, as well as among market towns in rural areas and among retail outlets.

Point patterns can be identified as clustered, dispersed, or random.

It is helpful to distinguish two kinds of processes responsible for point patterns. *First-order* processes involve points being located independently, but may still result in clusters because of varying point density. For example, the drinking-water hypothesis investigated by Dr. John Snow and described in Box 13.1 led to a higher density of points around the pump because of greater access. Similarly, the density of organisms of a particular species may vary over an area because of varying suitability of habitat. *Second-order* processes involve interaction between points, leading to clusters when the interactions are attractive in nature and to dispersion when they are competitive or repulsive. In the cholera case, the contagion hypothesis rejected by Snow is a second-order process and results in clustering even in situations when all other density-controlling factors are perfectly uniform. Unfortunately, as argued in Section 13.1, Snow's evidence did not allow him to resolve with complete confidence between first-order and second-order processes, and in general it is not possible to determine whether a given clustered point pattern was created by varying density factors or by interactions. On the other hand, dispersed patterns can only be created by second-order processes.

Clustering can be produced by two distinct mechanisms, identified as first-order and second-order.

Many tests are available for clusters, and some excellent books have been written on the subject. Only one test will be discussed in this section, to illustrate the method. This is the *K* function, and unlike many such statistics it provides an analysis of clustering and dispersion over a range of scales. In the interests of brevity the technical details will be omitted, but they can be found in the texts listed at the end of this chapter. They include procedures for dealing with the effects of the study area boundary, which is an important distorting factor for many pattern statistics.

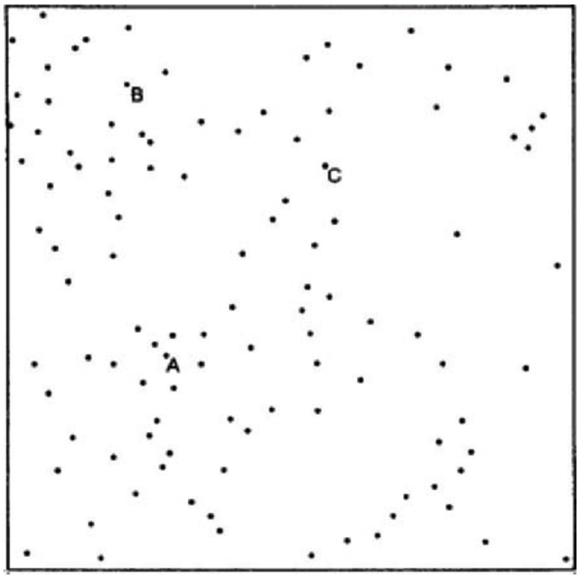
K(*d*) is defined as the expected number of points within a distance *d* of an arbitrarily chosen point, divided by the density of points per unit area. When the pattern is random, this number is πd^2, so the normal practice is to plot the function:

$$\hat{L}(d) = \sqrt{K(d)/\pi}$$

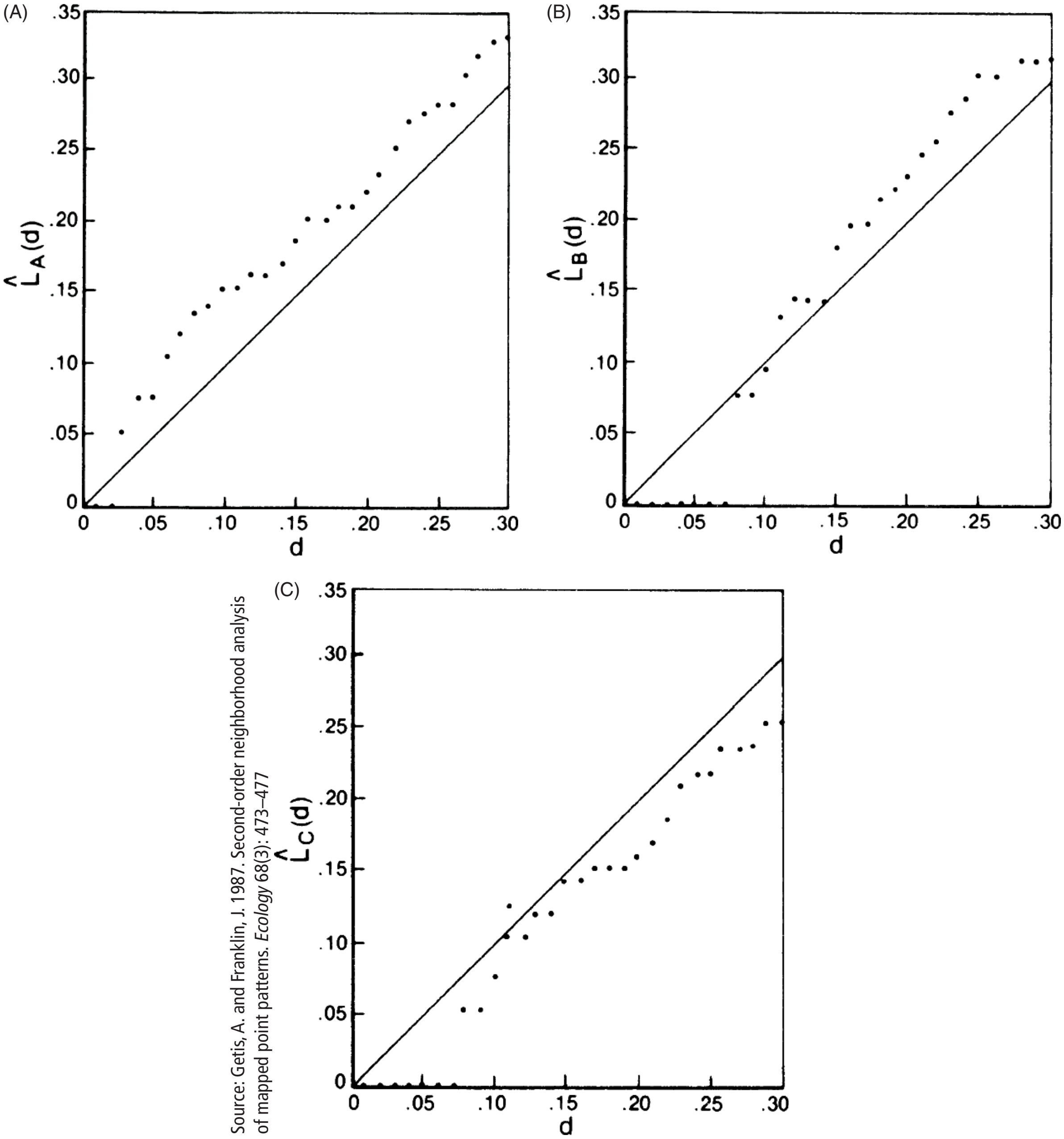
because $\hat{L}(d)$ will equal *d* for all *d* in a random pattern, and a plot of $\hat{L}(d)$ against *d* will be a straight line with a slope of 1. Clustering at certain distances is indicated by departures of $\hat{L}(d)$ above the line and dispersion by departures below the line.

Figures 13.19 and 13.20 show a simple example, used to discover how trees are spaced relative to

Figure 13.19 Point pattern of individual tree locations. A, B, and C identify the individual trees analyzed in Figure 13.20.



Source: Getis, A. and Franklin, J. 1987. Second-order neighborhood analysis of mapped point patterns. *Ecology* 68(3): 473–477.



Source: Getis, A. and Franklin, J. 1987. Second-order neighborhood analysis of mapped point patterns. *Ecology* 68(3): 473–477

Figure 13.20 Analysis of the local distribution of trees around three reference trees in Figure 13.18 (see text for discussion).

each other in a forest. The locations of trees are shown in Figure 13.19. In Figure 13.20A locations are analyzed in relation to Tree A. At very short distances there are fewer trees than would occur in a random pattern, but for most of the range up to a distance of 30% of the width (and height) of the study area there are *more* trees than would be expected, showing a degree of clustering. Tree B (Figure 13.20B) has no nearby neighbors, but shows a degree of clustering at longer distances. Tree C (Figure 13.20C) shows fewer trees than expected in a random pattern over most distances, and it is evident in Figure 13.19 that C is in a comparatively widely spaced area of forest.

13.3.4 Dependence at a Distance

The concept of distance decay—that interactions and similarities decline over space in ways that are often systematic—was introduced in Section 2.6 as a fundamental property of geographic data. It is so fundamental in fact that it is often identified somewhat informally as Tobler's First Law of Geography, first discussed in Chapter 2. This subsection examines techniques for measuring spatial dependence effects—in other words, the ways in which the characteristics of one location are correlated with characteristics of other nearby locations. Underlying this are concepts similar to those examined in the previous subsection on cluster detection—the potential for *hotspots* of anomalously high (or low) values of some property.

One way to look for such effects is to think of the features as fixed and their attributes as displaying interesting or anomalous patterns. Are attribute values randomly distributed over the features, or do extreme values tend to cluster: high values surrounded by high values, and low values surrounded by low values? In such investigations, the processes that determined the locations of features, and were the major concern in the previous section, tend to be ignored and may have nothing to do with the processes that created the pattern of their attributes. For example, the concern might be with some attribute of counties—their average house value, or percent married—and hypotheses about the processes that led to counties having different values on these indicators. The processes that led to the locations and shapes of counties, on the other hand, which were political processes operating perhaps 100 years ago, would be of no interest. The usefulness, or otherwise, of different areal units was discussed in Chapter 5.

The Moran statistic described in Chapter 2 is designed precisely for this purpose, to indicate the general properties of the pattern of attributes. It distinguishes between positively autocorrelated patterns, in which high values tend to be surrounded by high values and low values by low values; random patterns, in which neighboring values are independent of each other; and dispersed patterns, in which high values tend to be surrounded by low and vice versa. Section 2.7 described various ways of defining the weights needed to calculate the Moran statistic and also showed how it is possible to use various measures of separation or distance as a basis for the weights. A common expedient, described in Box 2.6, is to use a simple binary indicator of whether or not two areas are adjacent as a surrogate for the distance between them.

The Moran statistic looks for patterns among the attributes assigned to features.

In recent years there has been much interest in going beyond these global measures of spatial dependence to identify dependences locally. Is it possible, for example, to identify individual hotspots, areas where high values are surrounded by high values, and coldspots where low values are surrounded by low? Is it possible to identify anomalies, where high values are surrounded by low or vice versa? Local versions of the Moran statistic are among this group, and along with several others now form a useful resource that is easily implemented in GI systems.

Figure 13.21 shows an example, using the Local Moran statistic to differentiate states according to their roles in the pattern of one variable, the median value of housing. The weights have been defined by adjacency, such that pairs of states that share a common boundary are given a weight of 1 and all other pairs a weight of zero.

13.3.5 Density Estimation

One of the strongest arguments for spatial analysis is that it is capable of addressing *context*, of asking to what extent events or properties at some location are related to or determined by the location's surroundings. Does living in a polluted neighborhood make a person more likely to suffer from diseases such as asthma; does living near a liquor store make binge drinking more likely among young people; can obesity be blamed in part on a lack of nearby parks or exercise facilities? Buffering is one approach, defining a neighborhood as contained within a circle of appropriate radius around the person's location. The liquor store question, for example, could be addressed by building buffers around each individual and through a point-in-polygon operation determining the number of liquor stores within a defined distance. But this would imply that any liquor store within the buffer was equally important regardless of its distance, and every store outside the buffer was irrelevant. Instead some kind of attenuating effect of distance, such as the options shown in Figure 2.7, would seem to be more appropriate and realistic.

The general term for this kind of technique is *convolution*. A weighting function is chosen, based on a suitable distance decay function, and applied to the nearby features. So the net impact *P* of a set of liquor stores on a person at a given location **x** might be represented as the sum:

$$P(\mathbf{x}) = \sum_i w_i z_i$$

where i denotes a liquor store, z_i is a measure of the liquor store's relative size or importance, and w_i is a weight that declines with distance according to a distance decay function. Functions such as *P* defined in this way are often termed *potential* functions and have many uses in spatial analysis. They can be used to measure the potential impact of different population centers on a recreational facility, or the potential expenditures of the surrounding populations at a proposed retail facility. In all such cases the intent is to capture influence *at a distance*.

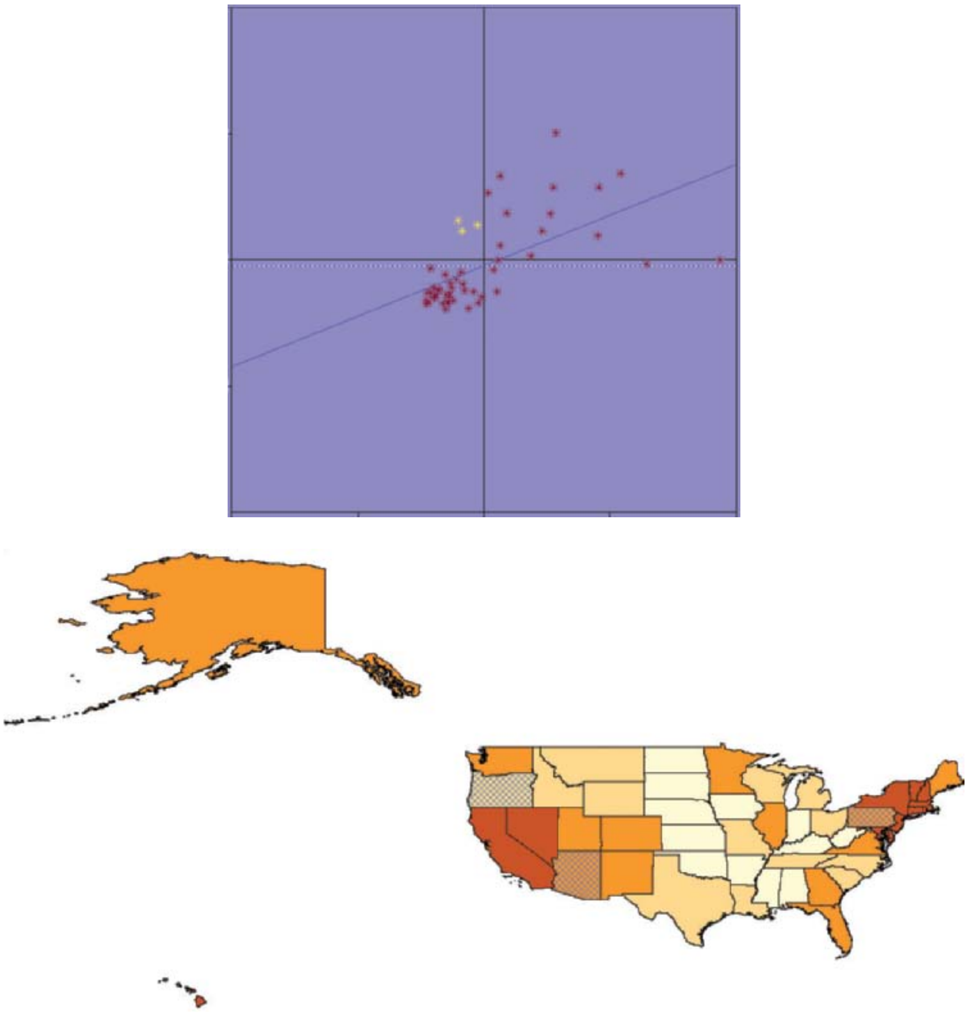


Figure 13.21 The Local Moran statistic, applied using the GeoDa software (available via geodacenter.asu.edu) to describe local aspects of the pattern of housing value among U.S. states. In the map window, the states are colored according to median value, with the darker shades corresponding to more expensive housing. In the scatterplot window, the median value appears on the *x*-axis, whereas the *y*-axis is the weighted average of neighboring states. The three points colored yellow are instances where a state of below-average housing value is surrounded by states of above-average value. The windows are linked, and the three points are identified as Oregon, Arizona, and Pennsylvania. The global Moran statistic is +0.4011, indicating a general tendency for clustering of similar values.

In effect, *P* measures the density of features in the neighborhood of **x**. The most obvious example is the estimation of population density, and that example is used in this discussion. But it could be equally well applied to the density of different kinds of diseases, or animals, or any other set of well-defined points. Consider a collection of point objects, such as those shown in Figure 13.22. The surface shown in the figure is an example of a *kernel function*, the central idea in density estimation. Any kernel function has an associated length measure, and in the case of the function shown, which is a Gaussian distribution, the length measure is a parameter of the distribution. We can generate Gaussian distributions with any value of this parameter, and they become flatter and wider as the value increases.

In density estimation, each point is replaced by its kernel function, and the various kernel functions are added to obtain an aggregate surface, a continuous field of density. If one thinks of each kernel as a pile of sand, then each pile has the same total weight of one unit. The total weight of all piles of sand is equal to the number of points, and the total weight of sand within a given area, such as the area shown in the figure, is an estimate of the total population in that area. Mathematically, if the population density is represented by a field $\rho(x,y)$, then the total population within area *A* is the integral of the field function over that area, that is:

$$P = \int_A \rho dA$$

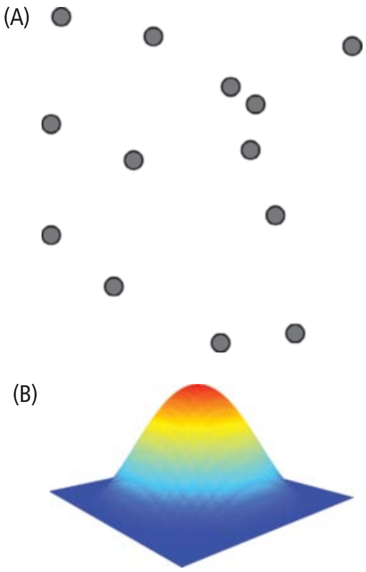


Figure 13.22 (A) A collection of point objects, and (B) a kernel function. The kernel's shape depends on a distance parameter—increasing the value of the parameter results in a broader and lower kernel, and reducing it results in a narrower and sharper kernel. When each point is replaced by a kernel and the kernels are added, the result is a density surface whose smoothness depends on the value of the distance parameter.

A variety of kernel functions are used in density estimation, but the form shown in Figure 13.22B is perhaps the most common. This is the traditional bell curve or Gaussian distribution of statistics and is encountered elsewhere in this book in connection with errors in the measurement of position in two dimensions (Section 5.3.2.2). By adjusting the width of the bell, it is possible to produce a range of density surfaces of different amounts of smoothness. Figure 13.23 contrasts two density estimations from the same data, one using a comparatively narrow bell to produce a complex surface and the other using a broader bell to produce a smoother surface.

Both parts of Figure 13.23 are illustrations of the concept of a *heat map*. Such mappings are commonly employed in studies of public health and in crime mapping to identify areas where the density of cases is especially high. Note, however, that the patterns shown in such maps may be confounded by other factors. A hot spot on a heat map of disease, for example, indicates the presence of a high density of the disease. But a hot spot or local high on a map of density of crime may not be of much interest if it simply mirrors a high density of population—we expect more crime where there are more people and do not expect much crime in rural areas where there are few people, other things being equal, so crime density that is simply proportional to local population density will be nothing special. To echo the discussion of Snow's map earlier, we cannot tell from the map whether the hot spot is due to a higher local rate of crime, or a constant rate superimposed on a higher density of people. When looking for hot spots, therefore, it is important to carefully select an appropriate *denominator* or *normalizing* variable and to show a

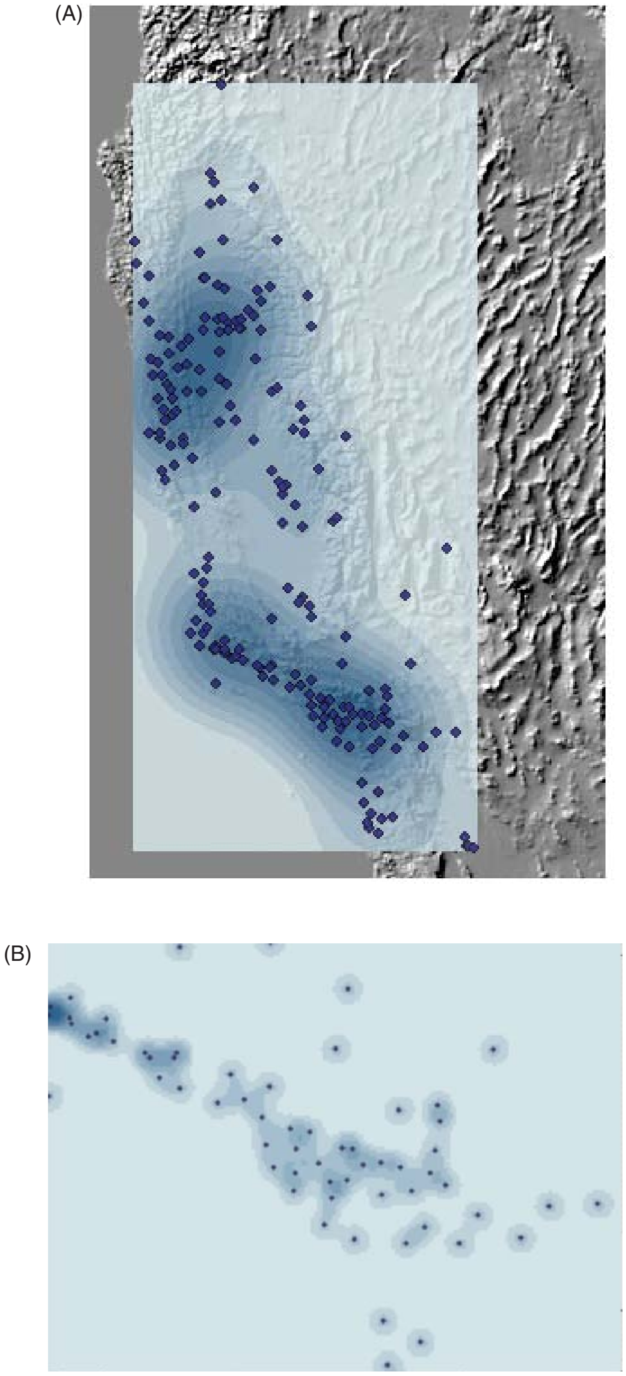


Figure 13.23 Density estimation using two different distance parameters in the respective kernel functions. In (A), the surface shows the density of ozone-monitoring stations in California, using a kernel radius of 150 km. In (B), zoomed to an area of Southern California, a kernel radius of 16 km is too small for this dataset, as it leaves each kernel isolated from its neighbors.

rate rather than a simple density. A hot spot on a map of crimes *per capita* will be much more insightful than one on a map of crime density alone. Figure 13.6 resolved this issue by mapping the rate (actually the statistical significance of the rate) rather than the density of disease.

An important lesson to extract from this discussion concerns the importance of scale. Density is an abstraction, created by taking discrete objects and convolving their distribution with a kernel function. The result depends explicitly on the width of the kernel, which we recognize as a length measure. Change the measure and the resulting density surface changes, as Figure 13.23 shows. Density estimation is just one example of a common phenomenon in geographic data—the importance of scale in the definition of many data types and hence the importance of knowing what that scale is explicitly when dealing with geographic data.

Many geographic data types have a scale built into their definition; it is important therefore to know what that scale is as precisely as possible.

13.3.6 Spatial Interpolation

Spatial interpolation is a pervasive operation in GI systems. Although it is often used explicitly in analysis, it is also used implicitly in various operations such as the preparation of a contour map display, where spatial interpolation is invoked without the user's direct involvement. Spatial interpolation is a process of intelligent guesswork in which the investigator (and the GI system) attempt to make a reasonable estimate of the value of a continuous field at places where the field has not actually been measured (spatial interpolation deals only with measured variables, that is, interval or ratio, not nominal or ordinal variables). All the methods use distance, based on the belief that the value at a location is more similar to the values measured at nearby sample points than to the values at distant sample points, a direct use of Tobler's First Law (discussed throughout Chapter 2). Spatial interpolation is an operation that makes sense only from the continuous-field perspective. The principles of spatial interpolation are discussed in Section 2.6; here the emphasis is on practical applications of the technique and on commonly used implementations of the principles.

Spatial interpolation finds applications in many areas:

- In estimating rainfall, temperature, and other attributes at places that are not weather stations, and where no direct measurements of these variables are available.
- In estimating the elevation of the surface between the measured locations of a digital elevation model (DEM).
- In *resampling* rasters (Section 13.2.5), the operation that must take place whenever raster data must be transformed to another grid.
- In contouring, when it is necessary to guess where to place contours between measured locations.

In all these instances spatial interpolation calls for intelligent guesswork, and the one principle that underlies all spatial interpolation is the Tobler Law (Section 2.6): "nearby things are more related than distant things." In other words, the best guess as to the value of a field at some point is the value measured at the closest observation points—the rainfall *here* is likely to be more similar to the rainfall recorded at the nearest weather stations than to the rainfall recorded at more distant weather stations. A corollary of this same principle is that in the absence of better information, it is reasonable to assume that any continuous field exhibits relatively smooth variation; fields tend to vary slowly and to exhibit strong positive spatial autocorrelation, a property of geographic data discussed in Section 2.6.

Spatial interpolation is the GISS version of intelligent guesswork.

In this section three methods of spatial interpolation are discussed, all distance-based: Thiessen polygons; inverse-distance weighting (IDW), which is the simplest commonly used method; and Kriging, a popular statistical method that is grounded in the theory of regionalized variables and falls within the field of *geostatistics*. These methods are discussed at greater length in de Smith, Goodchild, and Longley's *Geospatial Analysis* reader and at www.spatialanalysisonline.com.

13.3.6.1 Thiessen Polygons

Thiessen polygons were suggested by Thiessen as a way of interpolating rainfall estimates from a few rain gauges to obtain estimates at other locations where rainfall had not been measured. The method is very simple: to estimate rainfall at any point, take the rainfall measured at the closest gauge. This leads to a map in which rainfall is constant within polygons surrounding each gauge and changes sharply as polygon boundaries are crossed. Although many users associate polygons defined in this way with Thiessen, they are also known as Voronoi and Dirichlet polygons. They have many other uses besides spatial interpolation:

- Thiessen polygons can be used to estimate the trade areas of each of a set of retail stores or shopping centers.

- They are used internally in a GI system as a means of speeding up certain geometric operations, such as search for a point's nearest neighbor.
- They are the basis of some of the more powerful methods for generalizing vector databases (Section 3.8).
- As a method of spatial interpolation they leave something to be desired, however, because the sharp change in interpolated values at polygon boundaries is often implausible.

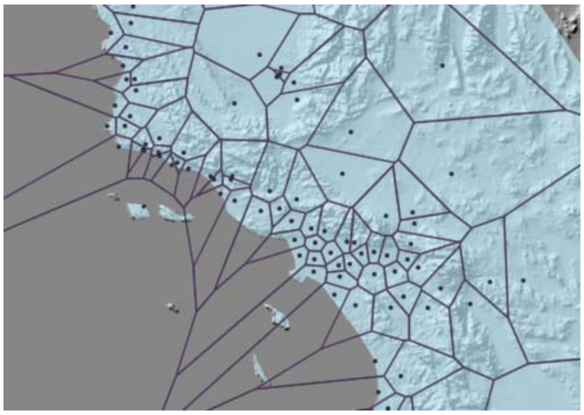
Figure 13.24 shows a typical set of Thiessen polygons. If each pair of points that share a Thiessen polygon boundary is connected, the result is the network of irregular triangles. These are named after Delaunay and are frequently used as the basis for the triangles of a triangulated irregular network (TIN) representation of terrain (Section 7.2.3.4).

13.3.6.2 Inverse-Distance Weighting

IDW is the workhorse of spatial interpolation, the method that is most often used in GI analysis. It employs the Tobler Law by estimating unknown measurements as weighted averages over the known measurements at nearby points, giving the greatest weight to the nearest points.

More specifically, denote the point of interest as **x**, and the points where measurements were taken as $\mathbf{x}_i$, where *i* runs from 1 to *n*, if there are *n* data points. Denote the unknown value as $z(\mathbf{x})$ and the known measurements as z_i. Give each of these points a weight w_i, which will be evaluated based on the distance d_i from $\mathbf{x}_i$ to **x**. Figure 13.25 explains this notation with a diagram. Then the weighted average computed at **x** is

Figure 13.24 Thiessen polygons drawn around each station in part of the Southern California ozone-monitoring network. Note how the polygons, which enclose the area closest to each point, in theory extend off the map to infinity and so must be truncated by the system at the edge of the map.



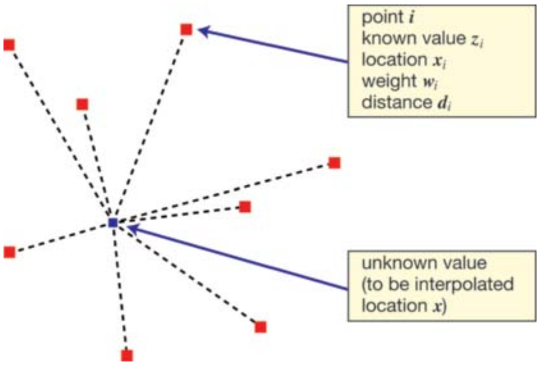


Figure 13.25 Notation used in the equations defining spatial interpolation.

$$z(\mathbf{x}) = \sum_i w_i z_i / \sum_i w_i$$

In other words, the interpolated value is an average over the observed values, weighted by the *w*'s. Notice the similarity between this equation and that used to define potential $P(\mathbf{x})$ in the previous subsection. The only difference is the presence here of a denominator, reflecting the nature of IDW as an averaging process rather than a summation.

There are various ways of defining the weights, but the option most often employed is to compute them as the inverse squares of distances—in other words (compare the options discussed in Section 2.5):

$$w_i = 1/d_i^2$$

This means that the weight given to a point drops by a factor of 4 when the distance to the point doubles (or by a factor of 9 when the distance triples). In addition, most software gives the user the option of ignoring altogether points that are further than some specified distance away, or of limiting the average to a specified number of nearest points, or of averaging over the closest points in each of a number of direction sectors. But if these values are not specified the software will assign default values to them.

IDW provides a simple way of guessing the values of a continuous field at locations where no measurement is available.

IDW achieves the desired objective of creating a smooth surface whose value at any point is more like the values at nearby points than the values at distant points. If it is used to determine *z* at a location where *z* has already been measured, it will return the measured value because the weight assigned to a point at zero distance is infinite. For this reason IDW is described as an *exact* method of interpolation because its interpolated results honor the data points exactly. (An *approximate* method is allowed to deviate from the measured values in the interests of greater smoothness, a property that is often useful if deviations are interpreted as indicating possible errors of measurement, or local deviations that are to be separated from the general trend of the surface.)

But because IDW is an average it suffers from certain specific characteristics that are generally undesirable. A weighted average that uses weights that are never negative must always return a value that is between the limits of the measured values; no point on the interpolated surface can have an interpolated *z* that is more than the largest measured *z*, or less than the smallest measured *z*. Imagine an elevation surface with some peaks and pits, but suppose that the peaks and pits have not actually been measured, but are merely indicated by the values of the measured points. Figure 13.26 shows a cross section of such a surface. Instead of interpolating peaks and pits as one might expect, IDW produces the kind of result shown in the figure—small pits where there should be peaks, and small peaks where there should be pits. This behavior is often obvious in GI system output that has been generated using IDW. A related problem concerns extrapolation: if a trend is indicated by the data, as shown in Figure 13.26, IDW will inappropriately indicate a regression to the mean outside the area of the data points.

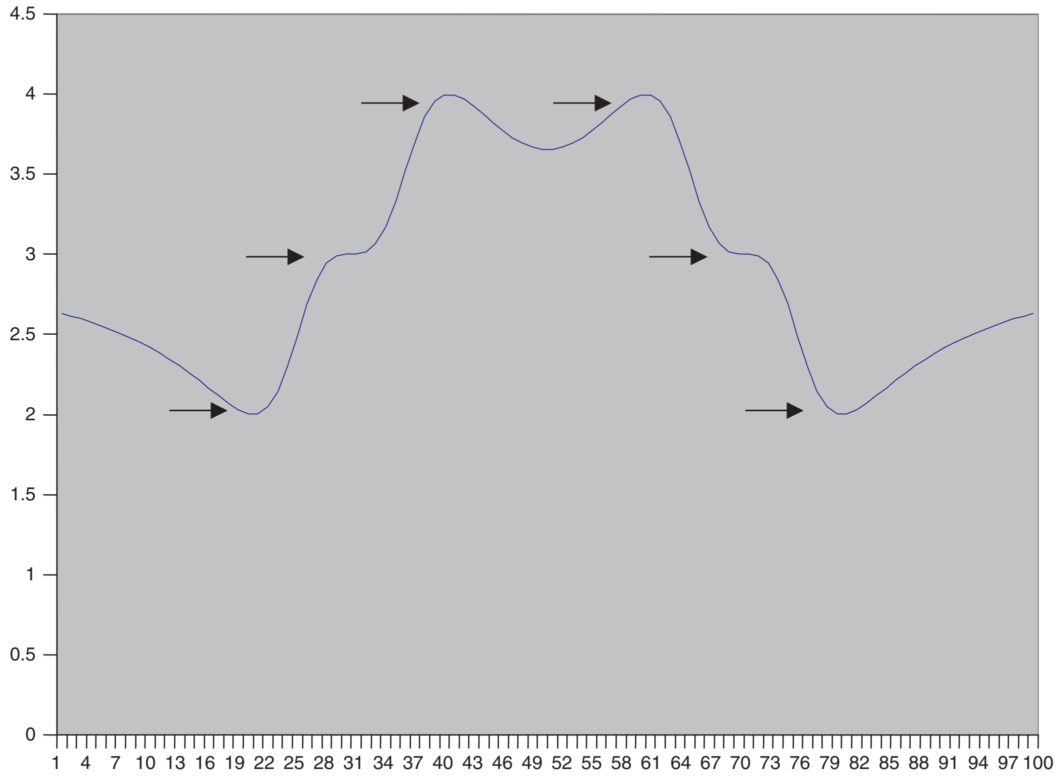
IDW interpolation may produce counter-intuitive results in areas of peaks and pits and outside the area covered by the data points.

In short, the results of IDW are not always what one would want. There are many better methods of spatial interpolation that address the problems that were just identified, but IDW's ease of programming and its conceptual simplicity make it among the most popular. Users should simply beware and take care to examine the results of interpolation to ensure that they make good sense.

13.3.6.3 Kriging

Of all the common methods of spatial interpolation, Kriging makes the most convincing claim to be grounded in good theoretical principles. The basic idea is to discover something about the general properties of the surface, as revealed by the measured values, and then to apply these properties in

Figure 13.26 Potentially undesirable characteristics of IDW interpolation. Data points located at 20, 30, 40, 60, 70, and 80 have measured values of 2, 3, 4, 4, 3, and 2, respectively. The interpolated profile shows a pit between the two highest values and regression to the overall mean value of 3 outside the area covered by the data.



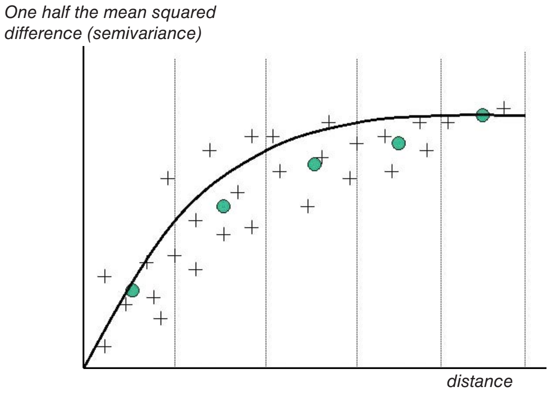
estimating the missing parts of the surface. Smoothness is the most important property (note the inherent conflict between this and the properties of fractals, Section 2.8), and it is operationalized in Kriging in a statistically meaningful way. There are many forms of Kriging, and the overview provided here is very brief. Further reading is identified at the end of the chapter.

There are many forms of Kriging, but all are firmly grounded in theory.

Suppose we take a point **x** as a reference and start comparing the values of the field there with the values at other locations at increasing distances from the reference point. If the field is smooth (if the Tobler Law is true, i.e., if there is positive spatial autocorrelation), the values nearby will not be very different—$z(\mathbf{x})$ will not be very different from $z(\mathbf{x}_i)$. To measure the amount, we take the difference and square it because the sign of the difference is not important: $(z(\mathbf{x}) - z(\mathbf{x}_i))^2$. We could do this with any pair of points in the area.

As distance increases, this measure will likely increase also, and in general a monotonic (consistent) increase in squared difference with distance is observed for most geographic fields. (Note that, as noted earlier, z must be measured on a scale that is at least interval, though *indicator Kriging* has been developed to deal with the analysis of nominal fields.) In Figure 13.27, each point represents one pair of values drawn from the total set of data points at which measurements have been taken. The vertical axis represents one-half of the squared difference (one-half is taken for mathematical reasons), and the graph is known as the *semivariogram* (or *variogram* for short—the difference of a factor of two is often overlooked in practice, though it is important mathematically). To express its contents in summary form, the distance axis is divided into a number of ranges or *bins*, as shown, and points within each range are averaged to define the heavy points shown in the figure. This semivariogram has been drawn without regard to the *directions* between points in a pair. As such, it is said to be an *isotropic* variogram. Sometimes there is sharp variation in the behavior in different directions, and *anisotropic* semivariograms are created for different ranges of direction (e.g., for pairs in each 90 degree sector).

Figure 13.27 A semivariogram. Each cross represents a pair of points. The solid circles are obtained by averaging within the ranges or buckets of the distance axis. The solid line is the best fit to these five points, using one of a small number of standard mathematical functions.



An anisotropic variogram asks how spatial dependence changes in different directions.

Note how the points of this typical variogram show a steady increase in squared difference up to a certain limit and how that increase then slackens off and virtually ceases. Again, this pattern is widely observed for continuous fields, and it indicates that difference in value tends to increase up to a certain limit, but then to increase no further. In effect, there is a distance beyond which there are no more geographic surprises. This distance is known as the *range* and the value of difference at this distance as the *sill*.

Note also what happens at the other, lower end of the distance range. As distance shrinks, corresponding to pairs of points that are closer and closer together, the semivariance falls, but there is a suggestion that it never quite falls to zero, even at zero distance. In other words, if two points were sampled a vanishingly small distance apart, they would give different values. This is known as the *nugget* of the semivariogram. A nonzero nugget occurs when there is substantial error in the measuring instrument, such that measurements taken a very small distance apart would be different due to error, or when there is some other source of local noise that prevents the surface from being truly smooth. Accurate estimation of a nugget depends on whether there are pairs of data points sufficiently close together. In practice, the sample points may have been located at some time in the past, outside the user's control, or may have been spread out to capture the overall variation in the surface, so it is often difficult to make a good estimate of the nugget.

The nugget can be interpreted as the variation among repeated measurements at the same point.

To make estimates using Kriging, we need to reduce the semivariogram to a mathematical function, so that semivariance can be evaluated at any distance, not just at the midpoints of buckets as shown in Figure 13.27. In practice, this means selecting one from a set of standard functional forms and fitting that form to the observed data points to get the best possible fit. This is shown in the figure. The user of a Kriging function in a GI system will have control over the selection of distance ranges and functional forms and whether a nugget is allowed.

Finally, the fitted semivariogram is used to estimate the values of the field at points of interest. As with IDW, the estimate is obtained as a weighted combination of neighboring values, but the estimate is designed to be the best possible given the evidence of the semivariogram. In general, nearby values are given greater weight, but unlike IDW direction is also important—a point can be *shielded* from influence if it lies behind another point because the latter's greater proximity suggests greater importance in determining the estimated value, whereas relative direction is unimportant in an IDW estimate. The process of maximizing the quality of the estimate is carried out mathematically, using the precise measures available in the semivariogram.

Kriging responds both to the proximity of sample points and to their directions.

Unlike IDW, Kriging has a solid theoretical foundation, but it also includes a number of options (e.g., the choice of the mathematical function for the semivariogram) that require attention from the user. In that sense it is definitely not a *black box* that can be executed blindly and automatically, but instead forces the user to become directly involved in the estimation process. For that reason GI software designers will likely continue to offer several different methods, depending on whether the user wants something that is quick, despite its obvious faults, or better but more demanding of the user.

13.3.6.4 A Final Word of Caution

Spatial interpolation and density estimation are in many ways logical twins: both begin with points and end with surfaces. Moreover, we already noted the similarity in the equations for density estimation and IDW. But conceptually the two approaches could not be more different because spatial interpolation seeks to estimate the missing parts of a continuous field from samples of the field taken at data points, whereas density estimation creates a continuous field from discrete objects. The values interpolated by spatial interpolation have the same measurement scale as the input values, but in density estimation the result is simply a count per unit area.

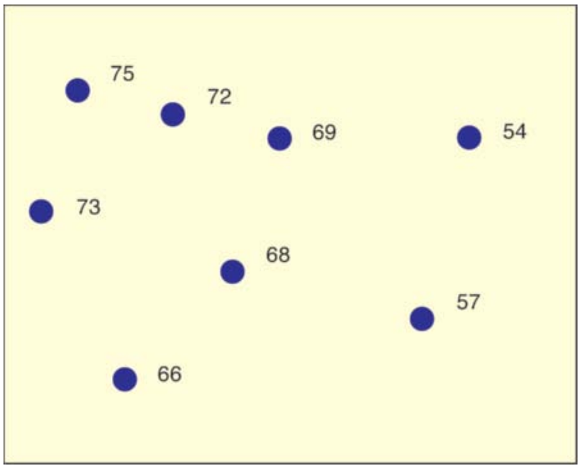


Figure 13.28 A dataset with two possible interpretations: first, a continuous field of atmospheric temperature measured at eight irregularly spaced sample points, and second, eight discrete objects representing cities, with associated populations in thousands. Spatial interpolation makes sense only for the former and density estimation only for the latter.

Figure 13.28 illustrates this difference. The dataset can be interpreted in two sharply different ways. In the first, it is interpreted as sample measurements from a continuous field, and in the second as a collection of discrete objects. In the discrete-object view there is nothing between the objects but empty space—no missing field to be filled in through spatial interpolation. It would make no sense at all to apply spatial interpolation to a collection of discrete objects and no sense at all to apply density estimation to samples of a field.

Density estimation makes sense only from the discrete-object perspective and spatial interpolation only from the field perspective.

13.4 Conclusion

This chapter has discussed some basic methods of spatial analysis based on two concepts: location and distance. Chapter 14 continues with techniques based on more advanced concepts, and Chapter 15 examines spatial modeling. Several general issues have been raised throughout the discussion: issues of scale and resolution and accuracy and uncertainty, which are discussed in greater detail in Chapter 5.

Questions for Further Study

1. Did Dr. John Snow actually make his inference strictly from looking at his map? What information can you find on the Web on this issue (try www.jsi.com)?
2. You are given a map showing the home locations of the customers of an insurance agent and are asked to construct a map showing the agent's market area. Would spatial interpolation or density estimation be more appropriate, and why?
3. What is conditional simulation, and how does it differ from Kriging? Under what circumstances might it be useful?
4. What are the most important characteristics of the three methods of spatial interpolation discussed in this chapter? Using a test dataset of your own choosing, compute and describe the major features of the surfaces interpolated by each.

Further Reading

De Smith, M. J., Goodchild, M. F., and Longley, P. A. 2009. *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools* (3rd ed.). Winchelsea: Winchelsea Press. www.spatialanalysisonline.com.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.* Hoboken, NJ: Wiley.

McLafferty, S. and Cromley, E. 2011. *GIS and Public Health* (2nd ed.). New York: Guilford.

O'Sullivan, D. and Unwin, D. J. 2010. *Geographic Information Analysis* (2nd ed.). Hoboken, NJ: Wiley.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis.* New York: Chapman and Hall.