5 Uncertainty

LEARNING OBJECTIVES

After studying this chapter you will:

- Understand the concept of uncertainty and the ways in which it arises from imperfect representation of geographic phenomena.
- Be aware of the uncertainties introduced in the three stages (conception, measurement and representation, and analysis) of database creation and use.
- Understand the concepts of vagueness and ambiguity and the uncertainties arising from the definition of key GI attributes.
- Understand how and why scale of geographic measurement and analysis can both create and propagate uncertainty.

Uncertainty in geographic representation arises because, of necessity, almost all representations of the world are incomplete. As a result, data in a geographic information (GI) system can be subject to measurement error, out of date, excessively generalized, or just plain wrong. This chapter identifies many of the sources of geographic uncertainty and the ways in which they operate in GI representations. Uncertainty arises from the way that GI users conceive of the world, how they measure and represent it, and how they analyze their representations of it. We investigate a number of conceptual issues in the creation and management of uncertainty, before reviewing the ways in which it may be measured using statistical and other methods. The propagation of uncertainty through geographical analysis is then considered. Uncertainty is an inevitable characteristic of GI usage, and one that users must learn to live with. In these circumstances, it becomes clear that all decisions based on GI systems are also subject to uncertainty.

5.1 Introduction

We can think of representations of the real world in GI databases as reconciling science with practice (how closely we can conform to the most appropriate scientific procedures: Section 2.4), concepts with applications (the normative versus the positive: Section 1.1.1), and analytical methods with the social context in which they are applied (see Section 1.7). Yet, almost always, such reconciliation is imperfect because, necessarily, representations of the world are incomplete (Section 3.4). In this chapter we will use *uncertainty* as an umbrella term to describe the problems that arise out of these imperfections. Occasionally, representations may approach perfect accuracy and precision (terms that we will define in Section 5.3.2.2)—as might be the case, for example, in the detailed site layout layer of a utility management system in which strenuous efforts are made to reconcile fine-scale multiple measurements of built environments. Yet perfect, or nearly perfect, representations of reality are the exception rather than the norm.

More often, the inherent complexity and detail of our world makes it virtually impossible to capture every single facet, at every possible scale, in

a digital representation. (Neither is this usually desirable; see the discussion of spatial sampling in Section 2.4.) Furthermore, different individuals see the world in different ways, and in practice no single view is likely to be seen universally as the best or to enjoy uncontested status. In this chapter we discuss how the processes and procedures of abstraction create differences between the contents of our (geographic and attribute) databases and the observable world that we purport them to represent. Such differences are almost inevitable, and understanding them can help us to manage uncertainty and to live with it.

It is impossible to make a perfect representation of the world, so uncertainty about it is inevitable.

Various terms are used to describe differences between the real world and how it appears in a GI database, depending on the context. The concept of error in statistics arises in part from omission of some relevant aspects of a phenomenon—as in the failure to fully specify all the predictor variables in a multiple regression model, for example. Similar problems arise when one or more variables are omitted from the calculation of a composite indicator—as, for example, in omitting road accessibility in an index of land value or omitting employment status from a measure of social deprivation (see Sections 3.8.2 and 15.2.1 for discussions of indicators). The established scientific notion of measurement *error* focuses on differences between observers or between measuring instruments. This raises issues of *accuracy,* which can be defined as the difference between reality and *our* representation of reality. Although such differences are often principally addressed in formal mathematical terms, the use of the word *our* acknowledges the varying perspectives that different observers may take upon a complex, multiscale, and inherently uncertain world.

Yet even this established framework is too simple for understanding quality or the defining standards of geographic data. The terms *ambiguity* and *vagueness* (defined in Section 5.2.2) identify further considerations that need to be taken into account in assessing the *quality* of a GI representation. Many geographic representations depend on inherently vague definitions and concepts. Quality is an important topic in GI systems, and many attempts have been made to identify its basic dimensions. The U.S. Federal Geographic Data Committee's (FGDC's) various standards list five components of quality: attribute accuracy, positional accuracy, logical consistency, completeness, and lineage. Definitions and other details on each of these and several more can be found on the FGDC's Web pages (www.fgdc.gov). Error, inaccuracy, ambiguity, and vagueness all contribute to the notion of uncertainty in the broadest sense, and uncertainty may thus be defined as a measure of the user's understanding of the difference between the contents of a dataset and the observable phenomena the data are believed to represent. This definition implies that phenomena are real, but includes the possibility that we are unable to describe them exactly. In GI systems, the term *uncertainty* has come to be used as the catchall term to describe situations in which the digital representation is simply incomplete and as a measure of the general quality of the representation.

Uncertainty accounts for the difference between the contents of a dataset and the phenomena that the data are supposed to represent.

The views outlined in the previous paragraph are themselves controversial and provide a rich ground for endless philosophical discussions. Some would argue that uncertainty can be inherent in phenomena themselves rather than just in their description. Others would argue for distinctions between *vagueness, uncertainty, fuzziness, imprecision, inaccuracy,* and many other terms that most people use as if they were essentially synonymous. Geographer Peter Fisher has provided a useful and wide-ranging discussion of these terms. We take the catchall view here and leave these arguments to further study.

In this chapter, we will discuss some of the principal sources of uncertainty and some of the ways in which uncertainty degrades the quality of a spatial representation. The way in which we conceive of a geographic phenomenon very much prescribes the way in which we are likely to set about representing (or measuring) it. Representation, in turn, heavily conditions the ways in which it may be analyzed within a GI system. This chain of events sequence, in which *conception* prescribes *representation*, which in turn prescribes *analysis*, is a succinct way of summarizing much of the content of this chapter and is summarized in Figure 5.1. In this diagram, U1, U2, and U3 each denote *filters* that can be thought of as selectively distorting or transforming the real world when it is stored and analyzed in a GI system. A later chapter (Section 12.2.1) introduces a fourth filter that mediates interpretation of analysis and the ways in which feedback may be accommodated through improvements in representation.

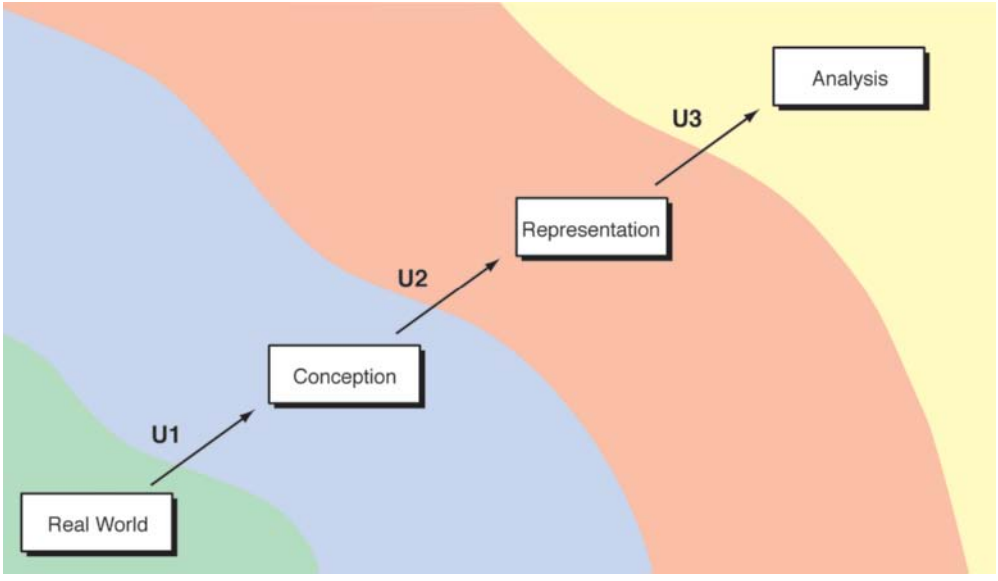


Figure 5.1 A conceptual view of uncertainty. The three filters, U1, U2, and U3, distort the way in which the complexity of the real world is conceived, represented, and analyzed in a cumulative way.

5.2 U1: Uncertainty in the Conception of Geographic Phenomena

In Chapter 3 we defined an atom of geographic information as linking a descriptive property or *attribute*, a *place*, and a *time* (Section 3.4). We have acknowledged that what is left out of a representation may be important, but have nonetheless assumed that what is included in a representation is founded on clear conceptions of places and attributes. In fact, it often turns out that this is not the case, and the working definitions that are used to represent places or attributes may or may not be fit for purpose.

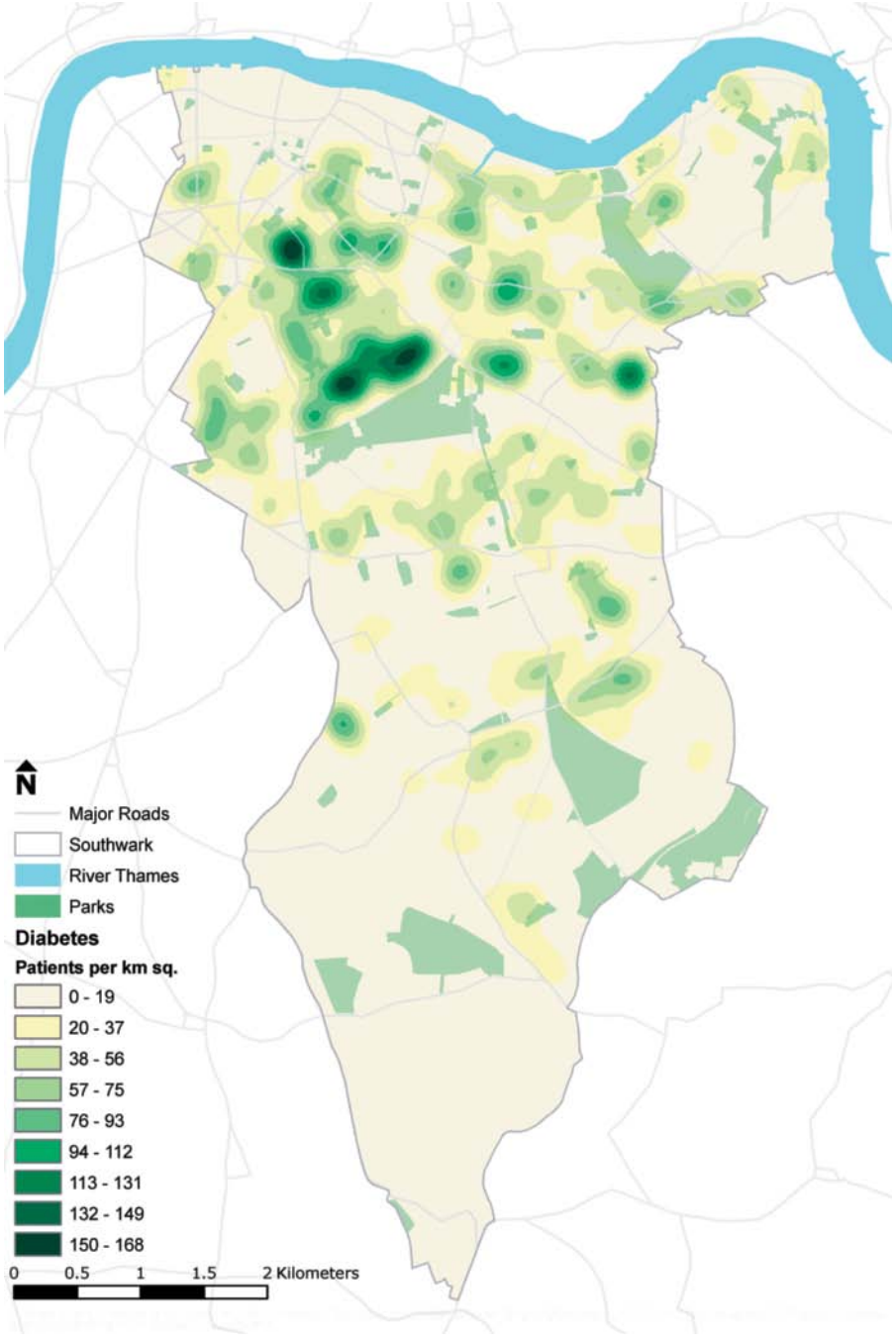
5.2.1 Conceptions of Place: Units of Analysis

The first component of an atom of geographic information is a *place*. The Tobler Law (Section 2.2), spatial autocorrelation (Section 2.7), and fractal geometry (Section 2.8) all testify to the special nature of spatial relationships between attributes, but this begs the important question of how these relationships may be best represented. In practice, this requires the creation of areal units of analysis with boundaries between them. Philosopher Barry Smith has argued that the basic typology of spatial boundaries involves distinguishing between *bona fide* (or physical) boundaries and *fiat* boundaries that are induced through human demarcation. Many geographic boundaries fall into the latter category and are said to demarcate areas that are *fiat objects*. Only rarely can geographic units of analysis be described as innate or *natural* to the purpose of geographic enquiry. What is the natural unit of measurement for a soil profile? What is the spatial extent of a *pocket* of high unemployment, or a *cluster* of cancer cases? How might we delimit the polluting effect of a coal-fired power station?

The questions become still more difficult in bivariate (two-variable) and multivariate (more than two variable) studies. At what scale is it appropriate to investigate any relationship between background radiation and the incidence of leukemia? Or to assess any relationship between labor-force qualifications and unemployment rates? Figure 5.2 shows some "hotspots" (local smoothed aggregations, designed to maintain the confidentiality of individual patient records) of the incidence of diabetes in the London Borough of Southwark. The raw data do indeed suggest local concentrations of the problem, but do not immediately suggest any areal basis to attempt interventions such as promoting healthier diets.

In many cases there are no natural units for geographic analysis.

The discrete object view of geographic phenomena relies far more on the idea of natural units of analysis than the field view. As such, this problem is more likely to be manifest in vector GI applications, such as those identified in Table 3.3. Things we manipulate, such as pencils, books, or screwdrivers, are obvious natural units. Biological organisms are almost always natural units of analysis, as are groupings such as households or families—though even here there are certainly difficult cases, such as the massive networks of fungal strands that are often claimed to be the largest living organisms on Earth, or extended families of human individuals. Most of the difficult cases fall into one of two categories—they are either instances of



Courtesy: Jakob Peterson and Maurizio Gibin

Figure 5.2 A map of local concentrations of diabetes in the London Borough of Southwark.

fields, where variation can be thought of as inherently continuous in space, or they are instances of poorly defined aggregations of discrete objects. In both of these cases it is up to the investigator to make the decisions about units of analysis, making the identification of the objects of analysis inherently subjective.

The absence of objective or uncontested definitions of place has not prevented geographers from attempting to classify them. The long-established regional geography tradition is fundamentally concerned with the quest to delineate zones that are internally homogeneous (with respect to climate, economic development, or agricultural land use, for example), set within a zonal scheme that maximizes between-zone heterogeneity—such as the map shown in Figure 5.3. Regional geography is fundamentally about delineating *uniform* zones, and many employ multivariate statistical techniques such as

Source: de Blij H. J. and Muller P. O. 2000. *Geography: Realms, Regions and Concepts* [9th ed.]. New York: Wiley, p. 113

Figure 5.3 The regional geography of Russia.

cluster analysis to supplement intuition—or sometimes to post-rationalize it.

Identification of homogeneous zones and spheres of influence lies at the heart of traditional regional geography as well as contemporary data analysis.

Other geographers have tried to develop *functional* zonal schemes in which zone boundaries delineate the breakpoints between the spheres of influence of adjacent facilities or features—as in the definition of travel-to-work areas, for example, or the definition of a river catchment. Zones may be defined such that there is maximal interaction within zones and minimal interaction between zones. Any functional zoning system is likely to prove contentious, if spending public funds results in different outcomes for individuals in different locations. Nowhere is this more apparent than in the contentious domain of defining and implementing community school catchment areas (Box 5.1).

5.2.2 Conceptions of Attributes: Vagueness and Ambiguity

5.2.2.1 Vagueness

The frequent absence of objective geographic individual units means that, in practice, the labels that we assign to zones and the ways in which we draw zone boundaries are often only vague best guesses. What absolute or relative incidence of oak trees in a forested zone qualifies it for the label *oak woodland* (Figure 5.5)? Or, in a developing country context in which aerial photography rather than ground enumeration is used to estimate population size, what rate of incidence of domestic properties indicates a zone of *dense* population? In each of these instances, it is expedient to transform point-like events (individual trees or individual properties) into area objects, and pragmatic decisions must then be taken in order to create a working definition of a spatial distribution. These decisions rarely have any absolute validity, and they raise two important questions:

- Is the defining boundary of a zone crisp and well defined?
- Is the assignment of a particular label to a given zone robust and defensible?

Uncertainty can exist both in the positions of the boundaries of a zone and in its attributes.

The idiographic tradition in geography (see Section 1.3) has a long-held preoccupation with defining

Applications Box 5.1

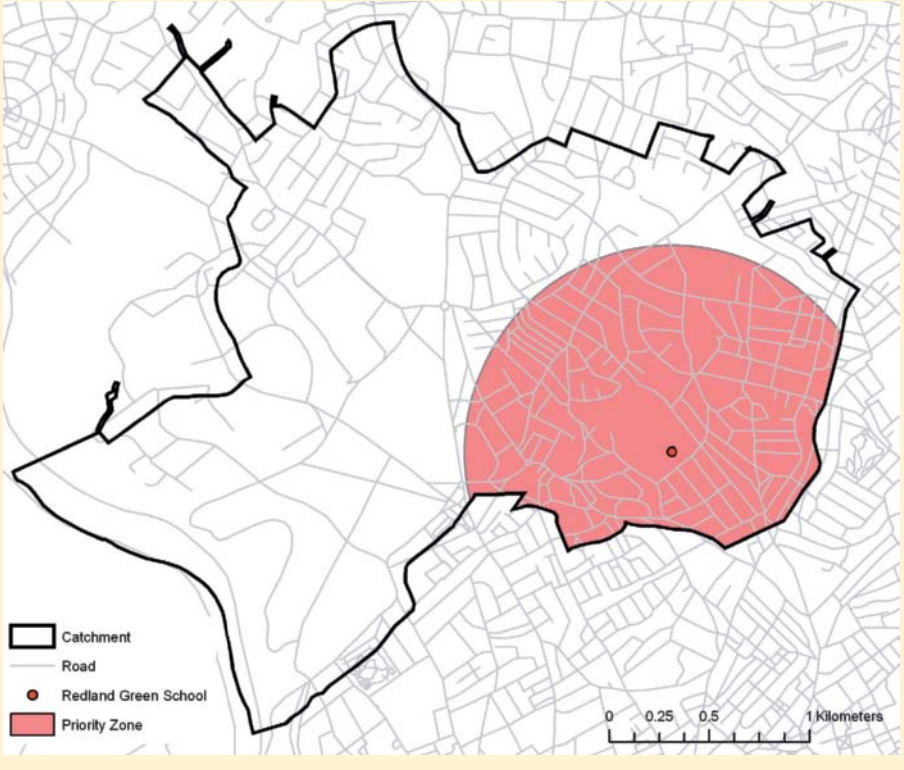
Functional Zones: Defining School Catchment Areas in Bristol, UK

In September 2007, the City of Bristol opened an attractive public-funded school at Redland Green, a wealthy area in which the inadequacies of provision had previously resulted in many parents seeking alternatives in the private sector. The location of the new school was itself the outcome of intense local lobbying, and the geography of the school's area of primary responsibility (in effect, its "catchment") was somewhat odd in appearance relative to the school's location (Figure 5.4).

When public (in the U.S., public-funded, sense) schools are oversubscribed in Britain, local authorities frequently use distance measures to allocate places to children who live nearer to the school. Bristol City Council almost totally failed to understand and anticipate demand for what was, in effect, a new public service, and in the first year that the new school buildings were open, almost all the places were allocated to prospective pupils living within 1.4 km of the school gates. Much less than half of the school's catchment area was served by the new school.

Responding to complaints from local parents, the UK Local Government Ombudsman ruled that the Council had given parents an "unrealistic expectation" of securing a place at the school and ordered the school to open its gates to appellants who had been denied access. While adhering to the Ombudsman's ruling, in the longer term the Council had to adapt to unforeseen circumstances by redefining the community for which the school was intended. It has considered a number of options, including redefining the catchment area to serve a much more geographically localized community of winners in its publicly funded schools lottery.

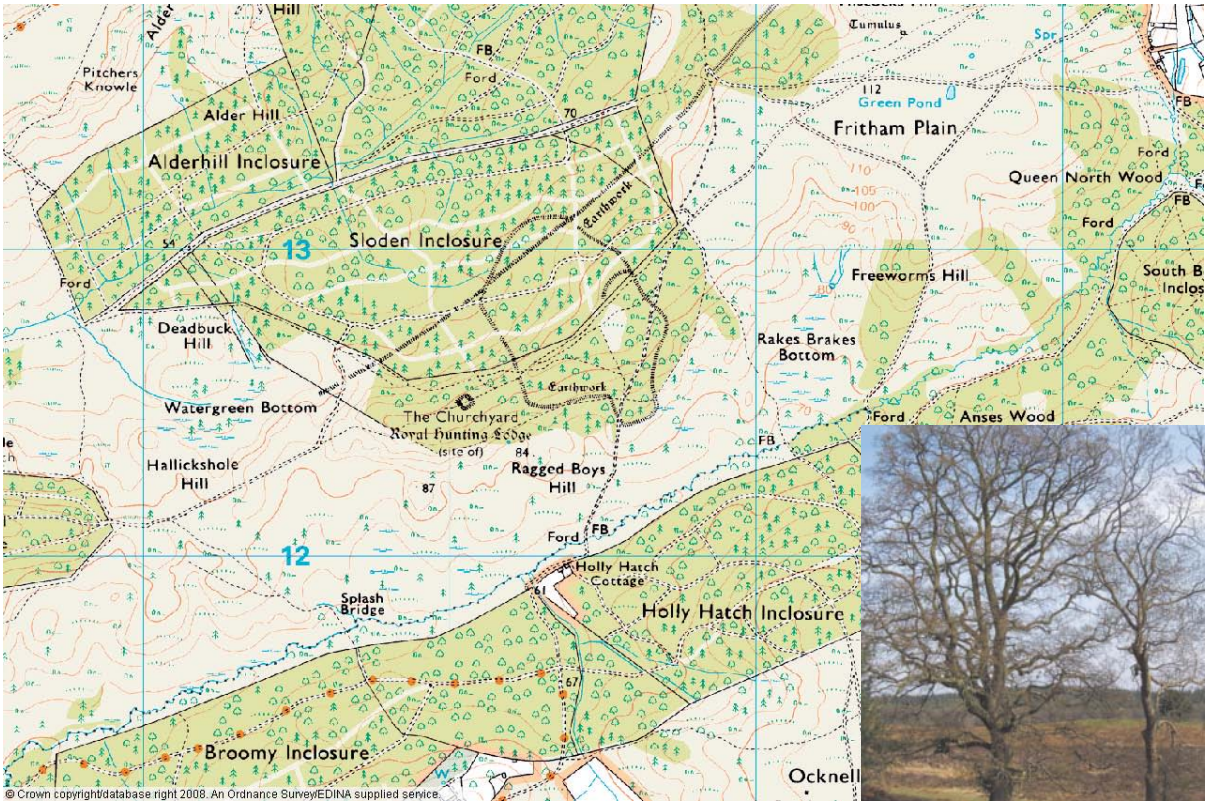
There is no such thing as a natural area for a school catchment. This being the case, GI systems can assist in defining a functional zone that is fit for purpose by deriving measures of demand and fair access to the facility—using socioeconomic data such as numbers of children of school age, travel time, and so on. Bristol City Council failed to do this. It was able to use ArcGIS measures of distance for the operational task of rationing places once the conceptual failure had become apparent, but this could not compensate for the failure to use any of the analytical functions of GI systems in any more strategic sense to anticipate demand for the new facility at the planning stage (see Chapter 17). The conception, and hence definition, of the school's catchment area was wholly inadequate for the community function envisaged for the school when public funds were initially committed to it.



Courtesy: Alex Singleton

Figure 5.4 The original catchment area of Bristol's Redland Green School, and the subarea within which offers of places were originally made.

Figure 5.5 A local assemblage of oak trees or part of an "oak woodland" on Ragged Boys Hill, in the New Forest, UK. The Ordnance Survey map of the area suggests that there are trees beyond the perimeter of the "woodland" area, whereas the tree symbology suggests that the woods are characterized by varying proportions of deciduous and coniferous trees.



Map: © Crown copyright/database right 2008 (an Ordnance Survey/EDINA supplied service).
Photo: courtesy Jim Champion

regions, and the vagaries inherent in tracing lines across maps can have important political or economic implications—as, for example, with the drawing of the Dayton Agreement lines to partition Bosnia and Herzegovina in 1995, or in defining the geographical areas to which political decision making should be devolved. The data processing power of GI systems may be harnessed through *geocomputation* (Section 15.1) to build regions from atoms of geographic information. Box 5.2 describes how the geography of Anglo-Saxon family names can be used to regionalize Great Britain.

Many English-language terms used to convey geographic information are inherently ambiguous.

Applications Box 5.2

Vagueness, Ambiguity, and the Geographies of Family Names

In Great Britain, family names ("surnames") entered common parlance in the Thirteenth Century. Most such names are toponyms (denoting landscape features or places, such as Castle or London), metonyms (denoting occupations, such as Smith or Baker), or diminutives (denoting family linkage, such as "William's son," Williamson or the abbreviated form Williams). Family names characteristic of unique places remain concentrated near the places where they were first coined—you are on average 53 times more likely to meet someone called Rossall in Blackpool, England, than in a randomly selected location, for example, because Blackpool is very close to the settlement of that name. Other broader types of names exhibit strong regional concentrations—as with the widespread use of the diminutive "-s" suffix in Wales (e.g., Jones or Williams).

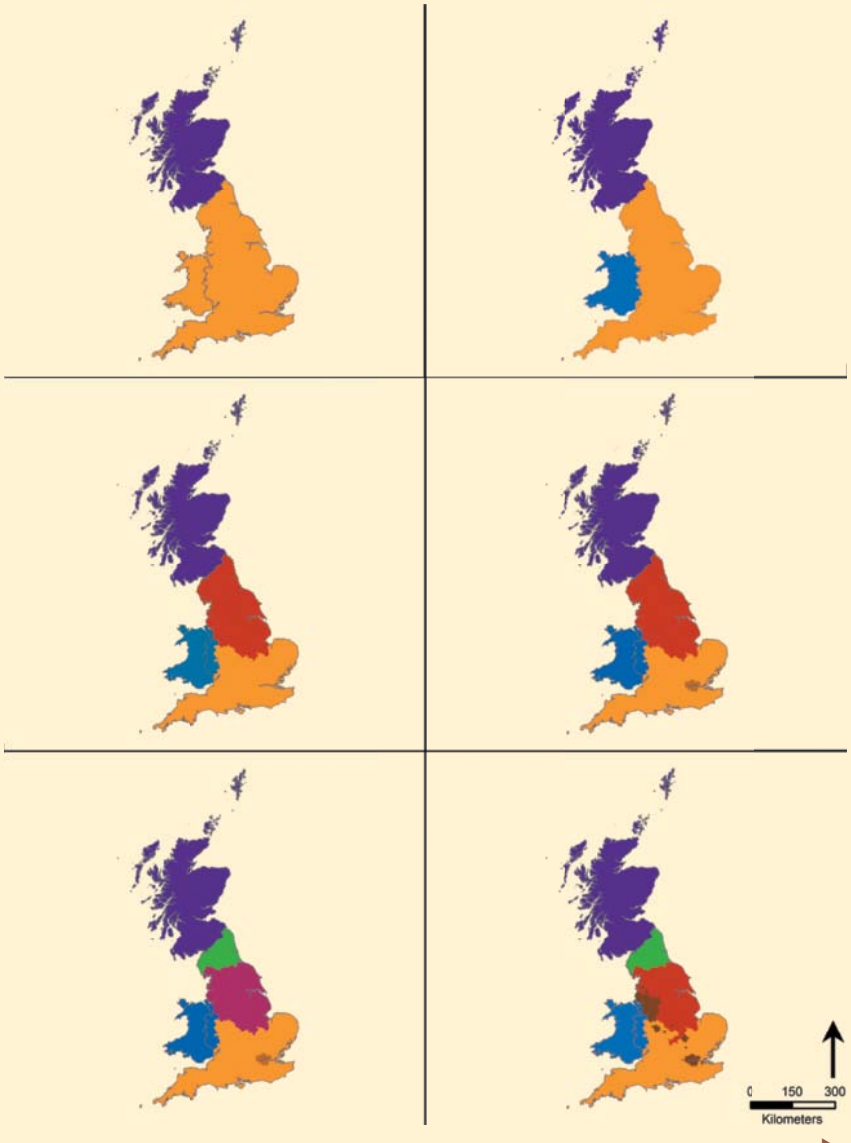
Analysis of family names tells us a lot about the enduring and unique human geographies of places, for the very good reason that throughout history most people have not moved very far from the places where they were brought up. The concentrations of individual names and types of names in places provide us with an interesting indicator of the distinctiveness of places and a basis to compare the shared characteristics of their populations. Using GI systems, we can do much more than map a single family name or a single family name type (see Box 1.2). Using geocomputational clustering techniques (see Section 15.1), we can identify the degree to which the mix of names in a particular place is distinctive.

Yet "distinctive" is a subjective, hence ambiguous, term. The lower the threshold that we adopt for defining a distinctive region, the more regions we will identify. Figure 5.6 shows how geographer James Cheshire has used names to partition Great Britain into between two and seven regions. Note how the distinctiveness of Scottish and Welsh names dominates the first two maps, successively followed by an emergent "north–south divide" in England, the separation of London from the rest of the south, further partitioning of the north, and finally the separation of the urban conurbations of northwest England.

These regions are much better than vague best guesses, as they are rooted in the naming conventions of a bygone age. But why stop at seven regions? Check out James Cheshire's regional geographies of Britain at www.spatialanalysis.co.uk/surnames.

Naming conventions can also help us resolve the ambiguities inherent in mapping the local geographies of ethnic minority populations, which is important for a wide range of policy applications. Data on ethnicity are collected in many population censuses but are vulnerable to the ambiguities arising from the ways in which individuals assign themselves to ethnic groups. (Would the third-generation descendant of a Polish immigrant to the U.S. describe himself or herself as "Polish," for example?). A different approach, which seeks neatly to circumvent these issues, begins with the adage that "a name is a statement." Research at University College London has used techniques of cluster analysis to classify names into more than 160 cultural, ethnic, and linguistic groups. The resulting geocomputational classification is the result of inductive classification of over 600 million names and enables the somewhat crude groups used in official statistics to be compared with a truly multicultural atlas of the UK. Figures 5.7A and B compare the official statistics classification with the names classification.

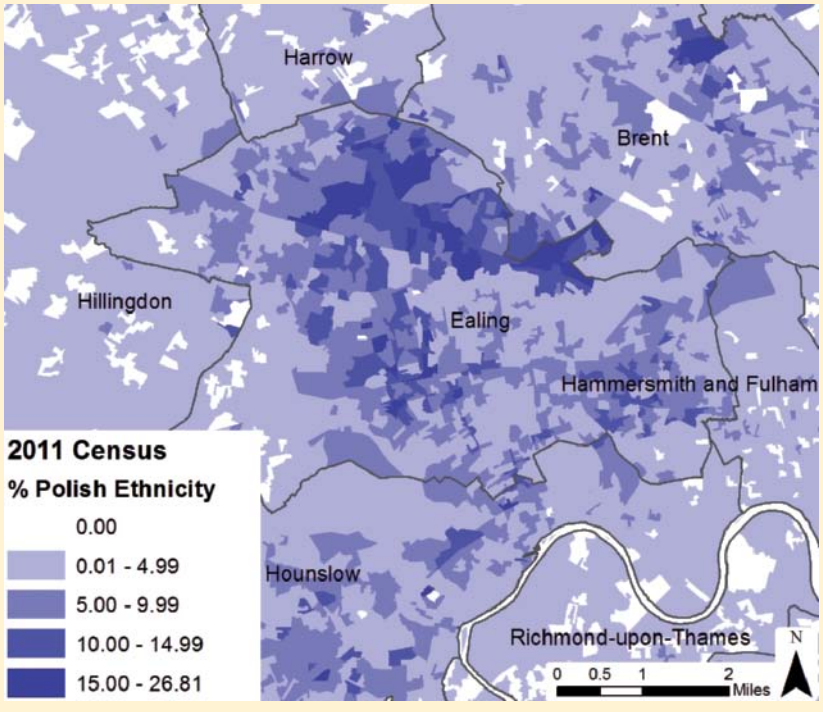
Check whether the classification assigns your name to the correct group at www.onomap.org and look at its global distribution at worldnames.publicprofiler.org.



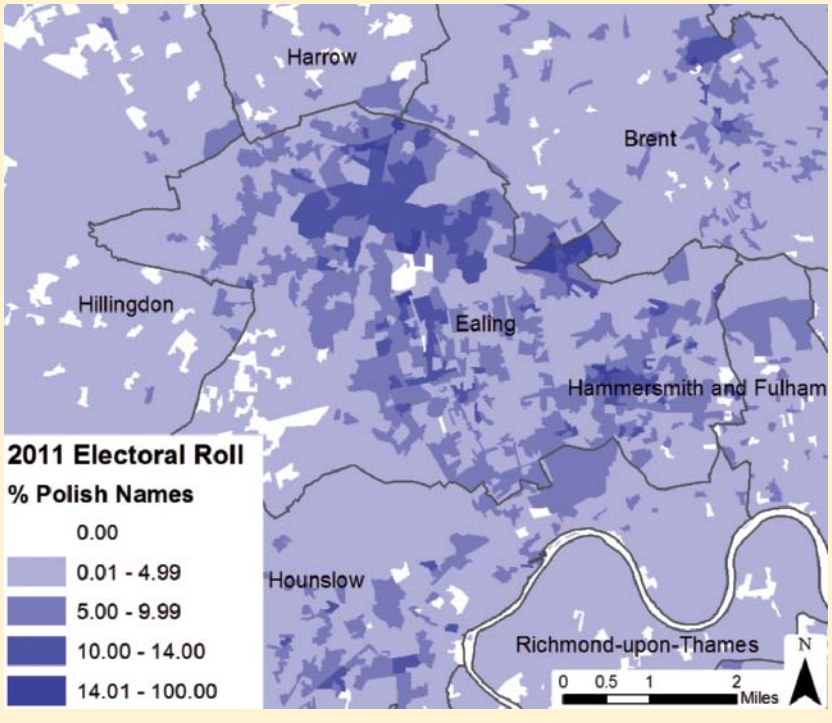
Courtesy: James Cheshire

Figure 5.6 The use of family names to regionalize Great Britain.

(A)



(B)



Courtesy: Guy Lansley

Figure 5.7 Differences between (A) percentages of the population describing themselves as of "Polish" ethnicity in 2011 UK Census for part of London, compared with (B) the percentages of people identified as having Polish names by applying names classification software to an enhanced version of the public Electoral Roll for the same year. Each definition and classification procedure has its own inherent uncertainties.

These questions have statistical implications (can we put numbers on the confidence associated with boundaries or labels?), cartographic implications (how can we convey the meaning of vague boundaries and labels through appropriate symbols on maps and displays of GI?) and cognitive implications (how do people subconsciously attempt to force things into categories and boundaries to satisfy a deep need to simplify the world?).

5.2.2.2 Ambiguity

Many objects are assigned different labels by different national or cultural groups, and such groups may share different spatial perceptions. Geographic prepositions in the English language such as *across*, *over*, and *in* do not have simple correspondences with terms in other languages. Object names and the topological relations between them may thus be inherently *ambiguous*. Perception, behavior, language, and cognition all play a part in the conception of real-world entities and the relationships between them. GI systems cannot provide the magic bullet of a value-neutral evidence base for decision making, and GI systems can be used to systematically privilege some worldviews over others. GI systems can also provide a formal framework for the reconciliation of different representations (see Section 3.3).

Ambiguity also arises in the conception and construction of *indicators* (see also Section 15.2.1). *Direct* indicators are deemed to bear a clear correspondence with a mapped phenomenon. Detailed household income figures, for example, can provide a direct indicator of the likely geography of expenditure and demand for goods and services; tree diameter at breast height can be used to estimate stand value; and field nutrient measures can be used to estimate agronomic yield. *Indirect* indicators are used when the best available measure is likely only to have surrogate link with the phenomenon of interest. Thus the incidence of central heating among households, or rates of multiple car ownership, might provide a surrogate for household income data if such data are not available, whereas local atmospheric measurements of nitrogen oxide can provide an indirect indicator of environmental health. Box 5.2 describes how people's names provide a direct indicator of their ethnicity and how this information may be used to improve on the detail, quality, and timeliness of ethnicity data collected in censuses of population.

Conception of the (direct or indirect) linkage between any indicator and the phenomenon of interest is subjective and hence ambiguous. Such measures will create errors of measurement if the correspondence between the two is imperfect, and these errors may be systematic. So, for example, differences in the conception of what hardship or deprivation entail can lead to specification of different composite indicators, whereas different geodemographic systems can include correspondingly varied cocktails of census variables. With regard to the natural environment, conception of critical defining properties of soils can lead to inherent ambiguity in their classification (see Section 5.2.3).

> **Ambiguity is introduced when imperfect indicators of phenomena are used instead of the phenomena themselves.**

Fundamentally, GI systems have upgraded our abilities to generalize about spatial distributions. Yet our abilities to do so may remain constrained by the different taxonomies that are conceived and used by data-collecting organizations within our overall study area. A study of wetland classification in the United States found no fewer than six agencies engaged in mapping the same phenomena over the same geographic areas, and each with its own definitions of wetland types (see Section 1.3). If wetland maps are to be used in regulating the use of land, as they are in many areas, then uncertainty in mapping clearly exposes regulatory agencies to potentially damaging and costly lawsuits. How might soils data classified according to the UK national classification be assimilated within a pan-European soils map, which uses a classification honed to the full range and diversity of soils found across the Europe rather than those just on an assemblage of offshore islands? How might different national geodemographic classifications be combined into a form suitable for a pan-European marketing exercise? These are all variants of the question:

How may mismatches between the categories of different classification schema be reconciled?

> **Differences in definitions are a major impediment to integration of geographic data over wide areas.**

Like the process of pinning down the different nomenclatures developed in different cultural settings, the process of reconciling the semantics of different classification schema is an inherently *ambiguous* procedure. Ambiguity arises in data concatenation when we are unsure regarding the *metacategory* to which a particular class should be assigned.

5.2.3 Fuzzy Approaches to Attribute Classification

One way of resolving the assignment process is to adopt a probabilistic interpretation. If we take a

statement like "the database indicates that this field contains wheat, but there is a 0.17 probability (or 17% chance) that it actually contains barley," there are at least two possible interpretations: (1) if 100 randomly chosen people were asked to make independent assessments of the field on the ground, 17 would determine that it contains barley, and 83 would decide it contains wheat; or (2) of 100 similar fields in the database, 17 actually contained barley when checked on the ground and 83 contained wheat. Of the two, we probably find the second more acceptable because the first implies that people cannot correctly determine the crop in the field.

But the important point is that, in conceptual terms, both of these interpretations are *frequentist* because they are based on the notion that the probability of a given outcome can be defined as the proportion of times the outcome occurs in some real or imagined experiment, when the number of tests is very large. Although this interpretation is reasonable for classic statistical experiments, like tossing coins or drawing balls from an urn, the geographic situation is different—there is only one field with precisely these characteristics, and one observer, and in order to imagine a number of tests we have to invent more than one observer, or more than one field. (The problems of imagining larger populations for some geographic samples are discussed further in Section 14.5.)

In part because of this problem, many people prefer the *subjectivist* conception of probability—that it represents a judgment about relative likelihood that is not the result of any frequentist experiment, real or imagined. Subjective probability is similar in many ways to the concept of fuzzy sets, and the latter framework will be used here to emphasize the contrast with frequentist probability.

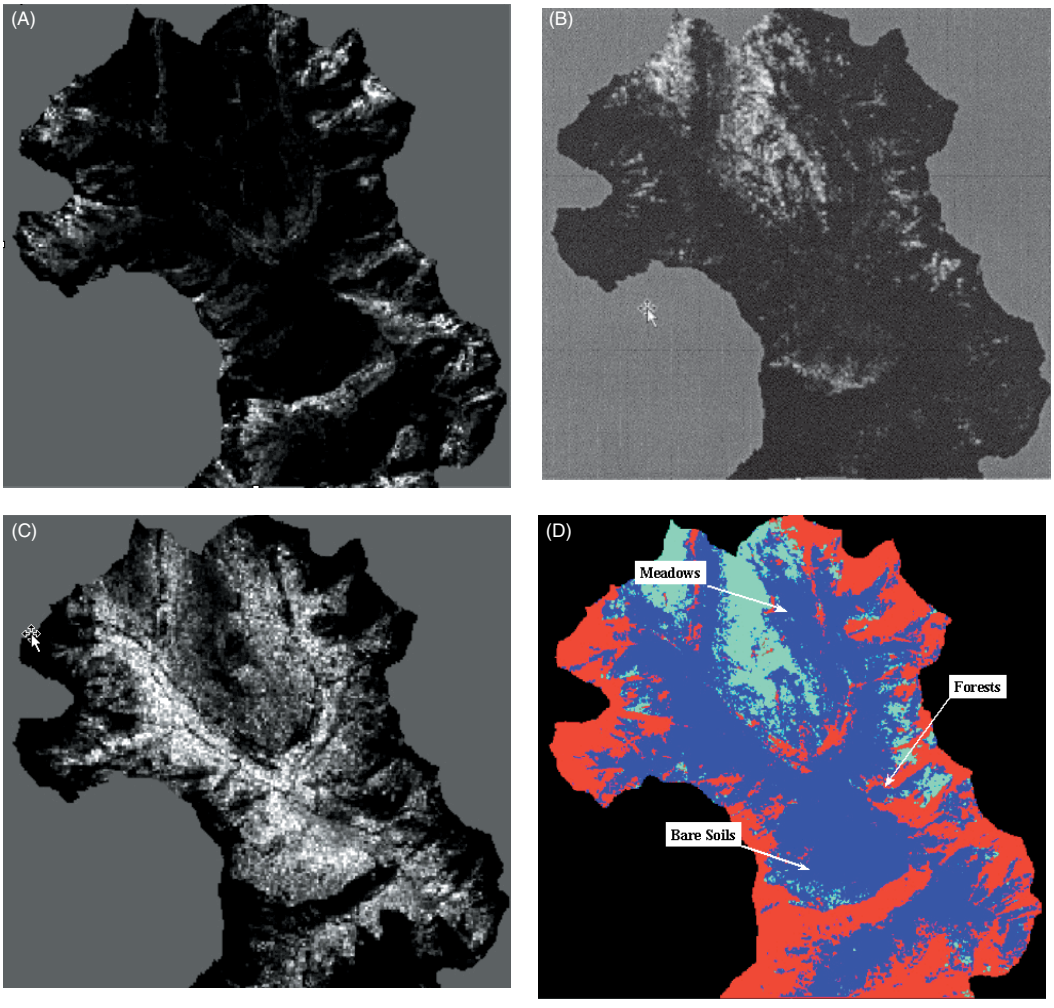
Suppose we are asked to examine an aerial photograph to determine whether a field contains wheat, and we decide that we are not sure. However, we are able to put a number on our degree of uncertainty by putting it on a scale from 0 to 1. The more certain we are, the higher the number. Thus we might say we are 0.90 sure it is wheat, and this would reflect a greater degree of certainty than 0.80. This degree of belonging to the class *wheat* is termed the *fuzzy membership*, and it is common, though not necessary, to limit memberships to the range 0 to 1. In effect, we have changed our view of membership in classes, and we have abandoned the notion that things must either belong to classes or not belong to them. In this new world, the boundaries of classes are no longer clean and crisp, and the set of things assigned to a set can be fuzzy.

In fuzzy logic, an object's degree of belonging to a class can be partial.

One of the major attractions of fuzzy sets is that they appear to let us deal with sets that are not precisely defined, and for which it is impossible to establish membership cleanly. Many such sets or classes are found in applications of GI, including land-use categories, neighborhood classifications, soil types, land cover classes, and vegetation types. Classes used for maps are often fuzzy, such that two people asked to classify the same location might disagree, not because of measurement error, but because the classes themselves are not perfectly defined and because opinions vary. As such, mapping is often forced to stretch the rules of scientific repeatability, which require that two observers will always agree.

Box 5.3 shows a typical extract from the legend of a soil map, and it is easy to see how two people might disagree, even though both are experts with years of experience in soil classification. Figure 5.8 shows an example of mapping classes using the fuzzy methods developed by A-Xing Zhu of the University of Wisconsin–Madison, which take both remote-sensing images and the opinions of experts as inputs. There are three classes, and each map shows the fuzzy membership values in one class, ranging from 0 (darkest) to 1 (lightest). This figure also shows the result of converting to *crisp* categories, or *hardening*—to obtain Figure 5.8D, each pixel is colored according to the class with the highest membership value.

Fuzzy approaches are attractive because they capture the uncertainty that many of us feel about the assignment of places on the ground to specific categories. But researchers have struggled with the question of whether they are more *accurate*. In a sense, if we are uncertain about which class to choose, then it is more accurate to say so, in the form of a fuzzy membership, than to be forced into assigning a class without qualification. But that does not address the question of whether the fuzzy membership value is accurate. If Class A is not well defined, it is hard to see how one person's assignment of a fuzzy membership of 0.83 in Class A can be meaningful to another person because there is no reason to believe that the two people share the same notions of what Class A means or of what 0.83 means, as distinct from 0.91 or 0.74. So although fuzzy approaches make sense at an intuitive level, it is more difficult to see how they could be helpful in the process of communication of geographic knowledge from one person to another.



Reproduced by permission: A-Xing Zhu

Figure 5.8 (A) Membership map for bare soils in the Upper Lake McDonald Basin, Glacier National Park. (B) Membership map for forest. (C) Membership map for alpine meadows. (D) Spatial distribution of the three cover types from hardening the membership maps.

Technical Box 5.3

Fuzziness in Classification: Description of a Soil Class

The following is the description of the Limerick series of soils from New England (the type location is in Chittenden County, Vermont), as defined by the US National Cooperative Soil Survey. Note the frequent use of vague terms such as *very*, *moderate*, *about*, *typically*, and *some*. Because the definition is so loose, it is possible for many distinct soils to be lumped together in this one class—and two observers may easily disagree over whether a given soil belongs to the class, even though both are experts. The definition illustrates the extreme problems of defining soil classes with sufficient rigor to satisfy the criterion of scientific repeatability.

"The Limerick series consists of very deep, poorly drained soils on flood plains. They formed in loamy alluvium. Permeability is moderate. Slope ranges from 0 to 3 percent. Mean annual precipitation is about 34 inches and mean annual temperature is about 45 degrees F. Depth to bedrock is more than 60 inches. Reaction ranges from strongly acid to neutral in the surface layer and moderately acid to neutral in the substratum. Textures are typically silt loam or very fine sandy loam, but lenses of loamy very fine sand or very fine sand are present in some pedons. The weighted average of fine and coarser sands, in the particle-size control section, is less than 15 percent."

5.3 U2: Further Uncertainty in the Representation of Geographic Phenomena

As with the conception of uncertainty, it is helpful to consider the representation of uncertainty with regard to the components of geographic information—measures of places (locations), attributes, and time period (although we do not consider time in detail here). We consider both the mode of representing place (Section 5.3.1) and the accuracy and precision with which it can be measured (Section 5.3.3). We consider the measurement of attributes at the nominal, ordinal, interval, and ratio scales (see Box 2.1).

5.3.1 Representation of Place/Location

The conceptual models (fields and objects) that were introduced in Chapter 3 impose very different filters upon reality, and as a result, their usual corresponding representational models (raster and vector) are characterized by different uncertainties. The vector model enables a range of powerful analytical operations to be performed (see Chapters 13 through 15), yet it also requires a priori conceptualization of the nature and extent of geographic individuals and the ways in which they nest together into higher-order zones. The raster model defines individual elements as square cells, with boundaries that bear no relationship at all to natural features, but nevertheless provides a convenient and (usually) efficient structure for data handling within a GI system. However, in the absence of effective automated pattern recognition techniques, human interpretation is usually required to discriminate between real-world spatial entities as they appear in a rasterized image.

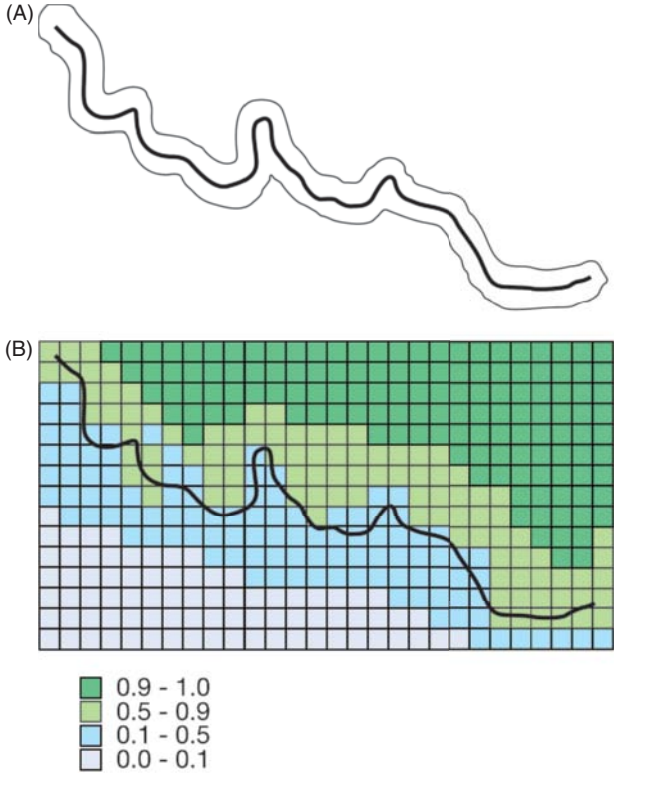
Although quite different representations of reality, both vector and raster data structures are attractive in their logical consistency, the ease with which they are able to handle spatial data, and (once the software is written) the ease with which they can be implemented in GI systems. But neither can provide any substitute for robust conception of geographic units of analysis (Section 5.2). This said, however, the conceptual distinction between fields and discrete objects is often useful in dealing with uncertainty. Figure 5.9 shows a coastline, which is often conceptualized as a discrete line object. But suppose we recognize that its position is uncertain. For example, the coastline shown on a 1:2,000,000 map is a gross generalization, in which major liberties are taken, particularly in areas where the coast is highly indented and irregular. Consequently, the 1:2,000,000 version leaves substantial uncertainty about the true location of the shoreline.

We might approach this by changing from a line to an area and mapping the area where the actual coastline lies, as shown in Figure 5.9A. But another approach would be to reconceptualize the coastline as a field by mapping a variable whose value represents the probability that a point is land. This is shown in Figure 5.9B as a raster representation. This would have far more information content and consequently much more value in many applications. But at the same time it would be difficult to find an appropriate data source for the representation—perhaps a fuzzy classification of an air photo, using one of an increasing number of techniques designed to produce representations of the uncertainty associated with objects discovered in images.

Uncertainty can be measured differently under field and discrete object views.

Indeed, far from offering quick fixes for eliminating or reducing uncertainty, the measurement process can actually increase it. Given that the vector and raster data models impose quite different filters on reality, it is unsurprising that they can each generate additional uncertainty in rather different ways. In field-based conceptualizations, such as those that underlie remotely sensed images expressed as rasters, spatial

Figure 5.9 The contrast between (A) discrete object and (B) field conceptualizations of an uncertain coastline.



objects are not defined a priori. Instead, the classification of each cell into one or other category builds together into a representation. In remote sensing, when resolution is insufficient to detect all the detail in geographic phenomena, the term *mixel* is often used to describe raster cells that contain more than one class of land—in other words, elements in which the outcome of statistical classification suggests the occurrence of multiple land-cover categories. The total area of cells classified as mixed should decrease as the resolution of the satellite sensor increases, assuming the number of categories remains constant, yet a completely mixel-free classification is very unlikely at any level of resolution. Even where the Earth's surface is covered with perfectly homogeneous areas, such as agricultural fields growing uniform crops, the failure of real-world crop boundaries to line up with pixel edges ensures the presence of at least some mixels. Neither does finer-resolution imagery solve all problems: medium-resolution data (defined as pixel size of between 30 m × 30 m and 1000 m × 1000 m) are typically classified using between 3 and 7 bands, whereas fine-resolution data (pixel sizes 10 m × 10 m or smaller) are typically classified using between 7 and 256 bands, and this can generate much greater heterogeneity of spectral values with attendant problems for classification algorithms.

A pixel whose area is divided among more than one class is termed a mixel.

The vector data structure, by contrast, defines spatial entities and specifies explicit topological relations (see Section 3.6) between them. Yet this often entails transformations of the inherent characteristics of spatial objects (Chapters 13 and 14). In conceptual terms, for example, although the true individual members of a population might each be defined as point-like objects, they will often appear in a GI database only as aggregate counts for apparently *uniform* zones. Such aggregation can be driven by the need to preserve the confidentiality of individual records, or simply by the need to limit data volume. Unlike the field conceptualization of spatial phenomena, this implies that there are good reasons for partitioning space in a particular way. In practice, partitioning is often made on grounds that are principally pragmatic, yet are rarely completely random (see Section 5.4). In most socioeconomic GI applications, for example, zones that are designed to preserve the anonymity of survey respondents may often be ad hoc containers. Larger aggregations are often used for the simple reason that they permit comparisons of measures over time (see Box 5.3). They may also reflect the way that a cartographer or GI software interpolates a boundary between sampled points, as in the creation of isopleth maps (see Box 2.5).

5.3.2 Statistical Models of Uncertainty in Attribute Measures

Scientists have developed many widely used methods for describing errors in observations and measurements, and these methods may be applicable to GI if we are willing to think of databases as collections of measurements. For example, a digital elevation model consists of a large number of measurements of the elevation of the Earth's surface. A map of land use is also in a sense a collection of measurements because observations of the land surface have resulted in the assignment of classes to locations. Both of these are examples of observed or measured attributes, but we can also think of location as a property that is measured.

A geographic database is a collection of measurements of phenomena on or near the Earth's surface.

Here we consider errors in nominal class assignment, such as of types of land use and errors in continuous (interval or ratio) scales, such as elevation (see Section 3.4).

5.3.2.1 Nominal Case

The values of nominal data serve only to distinguish an instance of one class from an instance of another or to identify an object uniquely (Section 3.4). If classes have an inherent ranking, they are described as ordinal data, but for purposes of simplicity the ordinal case will be treated here as if it were nominal. Consider a single observation of nominal data—for example, the observation that a single parcel of land is being used for agriculture (this might be designated by giving the parcel Class A as its value of the "Land-Use Class" attribute). For some reason, perhaps related to the quality of the aerial photography being used to build the database, the class may have been recorded falsely as Class G, Grassland. A certain proportion of parcels that are truly Agriculture might be similarly recorded as Grassland, and we can think of this in terms of a probability that parcels that are truly Agriculture are falsely recorded as Grassland.

Table 5.1 shows how this might work for all of the parcels in a database. Each parcel has a true class, defined by accurate observation in the field, and a recorded class as it appears in the database. The whole table is described as a *confusion matrix*, and instances of confusion matrices are commonly encountered in applications dominated by class data, such as classifications derived from remote sensing or aerial photography. The true class might be determined by ground check, which is inherently more accurate than classification of aerial photographs but much more expensive and time consuming. Ideally,

Table 5.1 Example of a misclassification or confusion matrix. A grand total of 304 parcels have been checked. The rows of the table correspond to the land-use class of each parcel as recorded in the database, and the columns to the class as recorded in the field. The numbers appearing on the principal diagonal of the table (from top left to bottom right) reflect correct classification.

	A	B	C	D	E	Total
A	80	4	0	15	7	106
B	2	17	0	9	2	30
C	12	5	9	4	8	38
D	7	8	0	65	0	80
E	3	2	1	6	38	50
Total	104	36	10	99	55	304

all the observations in the confusion matrix should lie along the principal diagonal, in the cells that correspond to agreement between true class and database class. But in practice certain classes are more easily confused than others, so certain cells off the diagonal will have substantial numbers of entries.

A useful way to think of the confusion matrix is as a set of rows, each defining a vector of values. The vector for any row i gives the proportions of cases in which what appears to be Class i is actually Class 1, 2, 3, and so on. Symbolically, this can be represented as a vector $\{p_1, p_2, \ldots, p_i, \ldots, p_n\}$, where n is the number of classes and p_i represents the proportion of cases for which what appears to be the class according to the database is actually Class i.

There are several ways of describing and summarizing the confusion matrix. If we focus on one row, then the table shows how a given class in the database falsely records what are actually different classes on the ground. For example, Row A shows that of 106 parcels recorded as Class A in the database, 80 were confirmed as Class A in the field, but 15 appeared to be truly Class D. The proportion of instances in the diagonal entries represents the proportion of correctly classified parcels, and the total of off-diagonal entries in the row is the proportion of entries in the database that appear to be of the row's class but are actually incorrectly classified. For example, there were only 9 instances of agreement between the database and the field in the case of Class D. If we look at the table's columns, the entries record the ways in which parcels that are truly of that class are actually recorded in the database. For example, of the 10 instances of Class C found in the field, 9 were recorded as such in the database and only 1 was misrecorded as Class E. The columns have been called the *producer's* perspective because the task of the producer of an accurate database is to minimize entries outside the diagonal cell in a given column; the rows have been called the *consumer's* perspective because they record what the contents of the database actually mean on the ground—in other words, the accuracy of the database's contents.

Users and producers of data look at misclassification in distinct ways.

For the table as a whole, the proportion of entries in diagonal cells is called the *percent correctly classified* (PCC) and is one possible way of summarizing the table. In this case 209/304 cases are on the diagonal, for a PCC of 68.8%. But this measure is misleading for at least two reasons. First, chance alone would produce some correct classifications, even in the worst circumstances, so it would be more meaningful if the scale were adjusted such that 0 represents chance. In this case, the number of chance hits on the diagonal in a random assignment is 76.2 (the sum of the row total times the column total divided by the grand total for each of the five diagonal cells). So the actual number of diagonal hits, 209, should be compared to this number, not 0. The more useful index of success is the *kappa index*, defined as

$$\kappa = \frac{\sum_{i=1}^{n} C_{ii} - \sum_{i=1}^{n} C_{i\cdot}C_{\cdot i}/C_{\cdot\cdot}}{C_{\cdot\cdot} - \sum_{i=1}^{n} C_{i\cdot}C_{\cdot i}/C_{\cdot\cdot}}$$

where c_{ij} denotes the entry in row i column j, the dots indicate summation (e.g., $c_{i\cdot}$ is the summation over all columns for row i, that is, the row i total, and $c_{\cdot\cdot}$ is the grand total), and n is the number of classes. The first term in the numerator is the sum of all the diagonal entries (entries for which the row number and the column number are the same). To compute PCC, we would simply divide this term by the grand total (the first term in the denominator). For kappa, both numerator and denominator are reduced by the same amount, an estimate of the number of hits (agreements between field and database) that would

occur by chance. This involves taking each diagonal cell, multiplying the row total by the column total, and dividing by the grand total. The result is summed for each diagonal cell. In this case kappa evaluates to 58.3%, a much less optimistic assessment than PCC.

The second issue with both of these measures concerns the relative abundance of different classes. In the table, Class C is much less common than Class A. The confusion matrix is a useful way of summarizing the characteristics of nominal data, but to build it there must be some source of more accurate data. Commonly, this is obtained by ground observation, and in practice the confusion matrix is created by taking samples of more accurate data, by sending observers into the field to conduct spot checks. Clearly, it makes no sense to visit every parcel, and instead a sample is taken. Because some classes are more common than others, a random sample that made every parcel equally likely to be chosen would be inefficient because too many data would be gathered on common classes, and not enough on the relatively rare ones. So, instead, samples are usually chosen such that a roughly equal number of parcels are selected in each class. Of course, these decisions must be based on the class as recorded in the database, rather than the true class. This is an instance of sampling that is *stratified* by class (see Section 2.4).

Sampling for accuracy assessment should pay greater attention to the classes that are rarer on the ground.

Parcels represent a relatively easy case, if it is reasonable to assume that the land-use class of a parcel is uniform over the parcel, and class is recorded as a single attribute of each parcel object. But as we noted in Sections 2.4 and 5.2.2.1, more difficult cases arise in sampling natural areas, for example, in the case of vegetation cover class, where parcel boundaries may not exist. Figure 5.10 shows a typical vegetation cover class map and is obviously highly generalized. If we were to apply the previous strategy, then we would test each area to see if its assigned vegetation cover class checks out on the ground. But unlike the parcel case, in this example the boundaries between areas are not fixed but are themselves part of the observation process, and we need to ask whether they are correctly located. Error in this case has two forms: misallocation of an area's class and mislocation of an area's boundaries. In some cases the boundary between two areas may be fixed because it coincides with a clearly defined line on the ground; but in other cases, the boundary's location is as much a matter of judgment as the allocation of an area's class.

Errors in land cover maps can occur in the locations of boundaries of areas, as well as in the classification of areas.

Figure 5.10 An example of a vegetation cover map.

In such cases we need a different strategy that captures the influence both of mislocated boundaries and of misallocated classes. One way to deal with this is to think of error not in terms of classes assigned to areas, but in terms of classes assigned to points. In a raster dataset, the cells of the raster are a reasonable substitute for individual points. Instead of asking whether area classes are confused and estimating errors by sampling areas, we ask whether the classes assigned to raster cells are confused, and we define the confusion matrix in terms of misclassified cells. This is often called *per-pixel* or *per-point* accuracy assessment, to distinguish it from the previous strategy of *per-polygon* accuracy assessment. As before, we would want to stratify by class, to make sure that relatively rare classes were sampled in the assessment.

5.3.2.2 Interval/Ratio Case

The second case addresses measurements that are made on interval or ratio scales. Here, error is best thought of not as a change of class but as a change of value, such that the observed value x' is equal to the true value x plus some distortion δx, where δx is hopefully small. δx might be either positive or negative because errors are possible in both directions. For example, the measured and recorded elevation at some point might be equal to the true elevation, distorted by some small amount. If the average distortion is zero, so that positive and negative errors balance

out, the observed values are said to be *unbiased*, and the average value will be true.

Error in measurement can produce a change of class, or a change of value, depending on the type of measurement.

Sometimes it is helpful to distinguish between *accuracy*, which has to do with the magnitude of δx, and *precision*. Unfortunately there are several ways of defining precision in this context, at least two of which are regularly encountered in the context of GI. Surveyors and others concerned with measuring instruments tend to define precision through the performance of an instrument in making repeated measurements of the same phenomenon. A measuring instrument is precise according to this definition if it repeatedly gives similar measurements, whether or not these are actually accurate. So a GPS receiver might make successive measurements of the same elevation, and if these are similar the instrument is said to be precise. Precision in this case can be measured by the variability among repeated measurements. But it is possible that all the measurements are approximately 5 m too high, in which case the measurements are said to be biased, even though they are precise, and the instrument is said to be inaccurate. Figure 5.11 illustrates this meaning of precision and its relationship to accuracy. The other definition of precision is more common in science generally. It defines precision as the number of digits

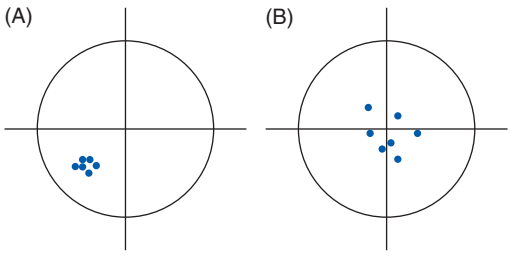


Figure 5.11 (A) Successive measurements have similar values (they are precise). (B) Precision is lower but accuracy is higher.

used to report a measurement, and again it is not necessarily related to accuracy. For example, a GPS receiver might measure elevation as 51.3456 m. But if the receiver is in reality only accurate to the nearest 10 cm, three of those digits are spurious, with no real meaning. So, although the precision is one ten-thousandth of a meter, the accuracy is only one-tenth of a meter. Box 5.4 summarizes the rules that are used to ensure that reported measurements do not mislead by appearing to have greater accuracy than they really do.

To most scientists, precision refers to the number of significant digits used to report a measurement, but it can also refer to a measurement's repeatability.

In the interval/ratio case, the magnitude of errors is described by the *root mean square error* (RMSE),

Technical Box 5.4

Good Practice in Reporting Measurements

Here are some simple rules that help to ensure that people receiving measurements from others are not misled by their apparently high precision.

1. The number of digits used to report a measurement should reflect the measurement's accuracy. For example, if a measurement is accurate to 1 m, then no decimal places should be reported. The measurement 14.4 m suggests accuracy to one-tenth of a meter, as does 14.0, but 14 suggests accuracy to 1 m.
2. Excess digits should be removed by rounding. Fractions above one-half should be rounded up, whereas fractions below one-half should be rounded down. The following examples reflect rounding to two decimal places:

   14.57803 rounds to 14.58

   14.57397 rounds to 14.57

   14.57999 rounds to 14.58

   14.57499 rounds to 14.57
3. These rules are not effective to the left of the decimal place; for example, they give no basis for knowing whether 1400 is accurate to the nearest unit or to the nearest hundred units.
4. If a number is known to be exactly an integer or whole number, then it is shown with no decimal point.

defined as the square root of the average squared error, or:

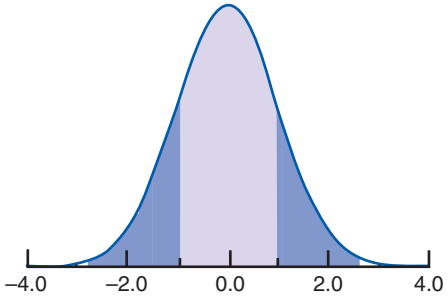
$$\left[\Sigma \delta x^2 / n\right]^{1/2}$$

where the summation is over the values of δx for all of the n observations. The RMSE is similar in a number of ways to the standard deviation of observations in a sample. Although RMSE involves taking the square root of the average squared error, it is convenient to think of it as approximately equal to the average error in each observation, whether the error is positive or negative. The U.S. Geological Survey uses RMSE as its primary measure of the accuracy of elevations in digital elevation models, and published values range up to 7 m.

Although the RMSE can be thought of as capturing the magnitude of the average error, many errors will be greater than the RMSE and many will be less. It is useful, therefore, to know how errors are *distributed* in magnitude—how many are large, how many are small. Statisticians have developed a series of models of error distributions, of which the most common and most important is the Gaussian distribution, otherwise known as the error function, the "bell curve," or the Normal distribution. Figure 5.12 shows the curve's shape. The height of the curve at any value of x gives the relative abundance of observations with that value of x. The area under the curve between any two values of x gives the probability that observations will fall in that range. If observations are unbiased, then the mean error is zero (positive and negative errors cancel each other out), and the RMSE is also the distance from the center of the distribution (zero) to the points of inflection on either side, as shown in the figure.

Let us take the example of a 7 m RMSE on elevations in a USGS digital elevation model; if error follows the Gaussian distribution, this means that some errors will be more than 7 m in magnitude, whereas some will be less, and also that the relative abundance of errors of any given size is described by the curve shown. 68% of errors will lie between +1.0 and –1.0 RMSEs, or +7 m and –7 m. In practice, many distributions of error do follow the Gaussian distribution, and there are good theoretical reasons why this should be so.

Figure 5.12 The Gaussian or Normal distribution. The lightly shaded area (between ±1 standard deviation) encloses 68% of the area under the curve, so 68% of observations will fall between these limits.



The Gaussian distribution is used to predict the relative abundances of different magnitudes of error.

To emphasize the mathematical formality of the Gaussian distribution, its equation is shown at the end of this paragraph. The symbol σ denotes the standard deviation, μ denotes the mean (in Figure 5.12 these values are 1 and 0, respectively), and exp is the exponential function, or "2.71828 to the power of" (also sometimes used to represent distance decay; see Section 2.5). Scientists believe that it applies very broadly and that many instances of measurement error adhere closely to the distribution because it is grounded in rigorous theory. It can be shown mathematically that the distribution arises whenever a large number of random factors contribute to error, and the effects of these factors combine additively—that is, a given effect makes the same additive contribution to error whatever the specific values of the other factors. For example, error might be introduced in the use of a steel tape measure over a large number of measurements because some observers consistently pull the tape very taut, or hold it very straight, or fastidiously keep it horizontal, or keep it cool, and others do not. If the combined effects of these considerations always contribute the same amount of error (e.g., +1 cm, or –2 cm), then this contribution to error is said to be additive.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

We can apply this idea to determine the inherent uncertainty in the locations of contours. The U.S. Geological Survey routinely evaluates the accuracies of its digital elevation models (DEMs) by comparing the elevations recorded in the database with those at the same locations in more accurate sources, for a sample of points. The differences are summarized in an RMSE, and in this example we will assume that errors have a Gaussian distribution with zero mean and a 7 m RMSE. Consider a measurement of 350 m. According to the error model, the truth might be as high as 360 m or as low as 340 m, and the relative frequencies of any particular error value are as predicted by the Gaussian distribution with a mean of zero and a standard

deviation of 7. If we take error into account, using the Gaussian distribution with an RMSE of 7 m, it is no longer clear that a measurement of 350 m lies exactly on the 350 m contour. Instead, the truth might be 340 m, or 360 m, or 355 m.

5.3.3 Statistical Models of Uncertainty in Location Measures

In the case of measurements of position, it is possible for every coordinate to be subject to error. In the two-dimensional case, a measured position (x', y') would be subject to errors in both x and y; specifically, we might write $x' = x + \delta x$, $y' = y + \delta y$, and similarly in the three-dimensional case where all three coordinates are measured, $z' = z + \delta z$. The *bivariate Gaussian distribution* describes errors in the two horizontal dimensions, and it can be generalized to the three-dimensional case. Normally, we would expect the RMSEs of x and y to be the same, but z is often subject to errors of quite different magnitude: for example, in the case of determinations of position using GPS. The bivariate Gaussian distribution also allows for correlation between the errors in x and y, but normally there is little reason to expect correlations.

Because it involves two variables, the bivariate Gaussian distribution has somewhat different properties from the simple (univariate) Gaussian distribution. As shown in Figure 5.12, 68% of cases lie within one standard deviation for the univariate case. But in the bivariate case with equal standard errors in x and y, only 39% of cases lie within a circle of this radius. Similarly, 95% of cases lie within two standard deviations for the univariate distribution, but it is necessary to go to a circle of radius equal to 2.15 times the x or y standard deviations to enclose 90% of the bivariate distribution and 2.45 times standard deviations for 95%.

National Map Accuracy Standards often prescribe the positional errors that are allowed in databases. For example, the 1947 U.S. National Map Accuracy Standard specified that 95% of errors should fall below 1/30 inch (0.85 mm) for maps at scales of 1:20,000 and finer (more detailed), and 1/50 inch (0.51 mm) for other maps (coarser, less detailed than 1:20,000). A convenient rule of thumb is that positions measured from maps are subject to errors of up to 0.5 mm at the scale of the map. Table 5.2 shows the distance on the ground corresponding to 0.5 mm for various common map scales.

A useful rule of thumb is that features on maps are positioned to an accuracy of about 0.5 mm.

Table 5.2 Positions measured from maps should be accurate to about 0.5 mm on the map. Multiplying this by the scale of the map gives the corresponding distance on the ground.

Map scale	Ground distance corresponding to 0.5-mm map distance
1:1250	52.5 cm
1:2500	1.25 m
1:5000	2.5 m
1:10 000	5 m
1:24 000	12 m
1:50 000	25 m
1:100 000	50 m
1:250 000	125 m
1:1 000 000	500 m
1:10 000 000	5 km

5.4 U3: Further Uncertainty in the Analysis of Geographic Phenomena

5.4.1 Internal and External Validation through Spatial Analysis

In Chapter 1 we defined a core remit of GI science as the resolution of scientific or decision-making problems through spatial analysis. Spatial analysis can be thought of as the process by which we turn raw spatial data into useful spatial information and thus far have thought of the creation of spatial information as adding value to attribute data through selectivity or preparation for purpose (see Chapters 13 and 14). A further defining characteristic is that the results of spatial analysis change when the frame or extent of the space under investigation changes. This also implies that the frame can be divided into units of analysis that are clearly defined, yet we have seen (Section 5.2.1) that there are likely to be few, if any, such units available to us. How can the outcome of spatial analysis be meaningful if it has such uncertain foundations?

Once again, this question has no easy answers, although we can begin by anticipating possible errors of positioning or the consequences of aggregating the subjects of analysis (such as individual people) into artificial geographic units of analysis (as when people are aggregated by census tracts, or disease incidences are aggregated by county). In so doing, we can illustrate how potential problems might arise, although we are unlikely to arrive at any definitive solutions—for

the simple reason that the truth is inherently uncertain. The ways in which we conceive and represent geographic phenomena may distort the outcome of spatial analysis by dampening or accentuating apparent variation across space or by restricting the nature and range of questions that can meaningfully be asked.

Good analysis cannot substitute for poor conceptions of geography or poor representation—but it can flag the likely consequences of both.

We can deal with this risk in three ways. First, although the analyst can only rarely tackle the *source* of uncertainty (analysts are rarely empowered to collect new, completely disaggregate data, for example), analysis using GI systems can help to pinpoint the ways in which uncertainty is likely to *operate* (or *propagate*) within the GI system and identify the likely degree of distortion arising from representational expedients.

Second, although we may have to work with areally aggregated data, GI systems allow us to model within-zone spatial distributions, and this can ameliorate the worst effects of artificial zonation. This allows us to gauge the effects of scale and aggregation through simulation of different possible outcomes. This is *internal validation* of the effects of scale, point placement, and spatial partitioning.

The third way that GI systems can address uncertainty is assessing the quality of a representation with reference to other data sources, thus providing a means of *external validation* of the effects of zonal averaging. In today's advanced GI service economy, there may be other data sources that can be used to gauge the effects of aggregation on our analysis. In Section 12.2.1 we formally refine the basic scheme presented in Figure 5.1 to consider the role of geovisualization in evaluating representations, although some of the validation principles set out in this chapter also entail visual approaches.

GI systems provide ways of validating representations, sometimes with and sometimes without reference to external data sources.

5.4.2 Validation through Autocorrelation: The Spatial Structure of Errors

Understanding the spatial structure of errors is key to accommodating their likely effects on the results of spatial analysis, and hence estimates or measures of spatial autocorrelation (Section 2.3) can provide an important validation measure. This is because strong positive spatial autocorrelation will reduce the effects of uncertainty on estimates of properties such as slope or area. The cumulative effects of error, termed *error propagation*, can nevertheless produce impacts that are surprisingly large. Some of the examples in this section have been chosen to illustrate the substantial uncertainties that can be produced by apparently innocuous data errors.

Error propagation measures the impacts of uncertainty in data on the results of GI system operations.

The confusion matrix, or more specifically a single row of the matrix, along with the Gaussian distribution, provide convenient ways of describing the error present in a single observation of a nominal or interval/ratio measurement, respectively. When a GI system is used to respond to a simple query, such as, "Tell me the class of soil at this point," or "What is the elevation here?" then these methods are good ways of describing the uncertainty inherent in the response. For example, a GI system might respond to the first query with the information "'Class A, with a 30% probability of Class C," and to the second query with the information "350 m, with an RMSE of 7 m." Notice how this makes it possible to describe nominal data as accurate to a percentage, but it makes no sense to describe a DEM, or any measurement on an interval/ratio scale, as accurate to a percentage. For example, we cannot meaningfully say that a DEM is "90% accurate."

However, many GI system operations involve more than the properties of single points, and this makes the analysis of error much more complex. For example, consider the query, "How far is it from this point to that point?" Suppose the two points are both subject to error of position because their positions have been measured using GPS units with mean distance errors of 50 m. If the two measurements were taken some time apart, with different combinations of satellites above the horizon, it is likely that the errors are independent of each other, such that one error might be 50 m in the direction of North, and the other 50 m in the direction of South. Depending on the locations of the two points, the error in distance might be as high as 100 m. On the other hand, if the two measurements were made close together in time, with the same satellites above the horizon, it is likely that the two errors would be similar, perhaps 50 m North and 40 m North, leading to an error of only 10 m in determining distance. The difference between these two situations can be measured in terms of the degree of *spatial autocorrelation*, or the interdependence of errors at different points in space (see Section 2.6).

The spatial autocorrelation of errors can be as important as their magnitude in many GI system operations.

Spatial autocorrelation is also important in analyzing probable errors in nominal data. Reconsider the agricultural field discussed in Section 5.2.3 that is known

to contain a single crop, perhaps barley. When seen from above, it is possible to confuse barley with other crops, so there may be error in the crop type assigned to points in the field. But because the field has only one crop, we know that such errors are likely to be strongly correlated. Spatial autocorrelation is almost always present in errors to some degree, but very few efforts have been made to measure it systematically. As a result, it is difficult to make good estimates of the uncertainties associated with many GI system operations.

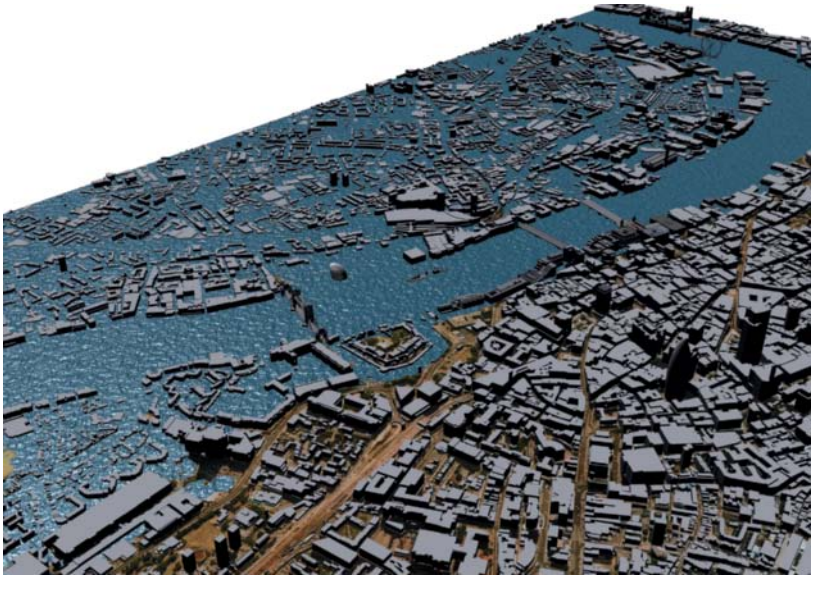
The spatial structure or autocorrelation of errors is important in many ways. DEM data are often used to estimate the slope of terrain, and this is done by comparing elevations at points a short distance apart. For example, if the elevations at two points 10 m apart are 30 m and 35 m, respectively, the slope along the line between them is 5/10, or 0.5. (A somewhat more complex method is used in practice, to estimate slope at a point in the *x*- and *y*-directions in a DEM raster, by analyzing the elevations of nine points—the point itself and its eight neighbors. The equations in Section 14.3.1 detail the procedure.) Now consider the effects of errors in these two elevation measurements on the estimate of slope. Suppose the first point (elevation 30 m) is subject to a RMSE of 2 m, and consider possible true elevations of 28 m and 32 m. Similarly, the second point might have true elevations of 33 m and 37 m. We now have four possible combinations of values, and the corresponding estimates of slope range from (33 − 32)/10 = 0.1 to (37 − 28)/10 = 0.9. In other words, a relatively small amount of error in elevation can produce wildly varying slope estimates.

The spatial autocorrelation between errors in geographic databases helps to minimize their impacts on many GI system operations.

What saves us in this situation, and makes estimation of slope from DEMs a practical proposition at all, is spatial autocorrelation among the errors. In reality, although DEMs are subject to substantial errors in absolute elevation, neighboring points nevertheless tend to have similar errors, and errors tend to persist over quite large areas. Most of the sources of error in the DEM production process tend to produce this kind of persistence of error over space, including errors due to misregistration of aerial photographs. In other words, errors in DEMs exhibit strong positive spatial autocorrelation.

Another important corollary of positive spatial autocorrelation can also be illustrated using DEMs. Suppose an area of low-lying land is predicted to be submerged by sea-level rise, and our task is to estimate the area of land affected (Figure 5.13). We are asked to do this using a DEM, which is known to have an RMSE of 2 m. Suppose the data points in the DEM are 30 m apart, and preliminary analysis shows that 100 points have elevations below the flood line. We might conclude that the area flooded is the area represented by these 100 points, or 900 × 100 sq m, or 9 hectares. But because of errors, it is possible that some of this area is actually above the flood line (we will ignore the possibility that other areas outside this may also be below the flood line, also because of errors), and it is possible that *all* the area is above. Suppose the recorded elevation for each of the 100 points is 2 m below the flood line. This is one RMSE (recall that the RMSE is equal to 2 m) below the flood line, and the Gaussian distribution tells us that the chance that the true elevation is actually above the flood line is approximately 16% (see Figure 5.12).

But what is the chance that *all* 100 points are actually above the flood line? Here again the answer



Courtesy: Andy Hudson-Smith

Figure 5.13 The hypothetical effects of a sea-level rise of 6 m on London, viewed in the Virtual London model.

depends on the degree of spatial autocorrelation among the errors. If there is none, in other words, if the error at each of the 100 points is independent of the errors at its neighbors, then the answer is $(0.16)^{100}$, or 1 chance in 1 followed by roughly 70 zeroes. But if there is strong positive spatial autocorrelation, so strong that all 100 points are subject to exactly the same error, then the answer is 0.16. One way to think about this is in terms of *degrees of freedom*. If the errors are independent, they can vary in 100 independent ways, depending on the error at each point. But if they are strongly spatially autocorrelated, the effective number of degrees of freedom is much less and may be as few as 1 if all errors behave in unison. Spatial autocorrelation has the effect of reducing the number of degrees of freedom in geographic data below what may be implied by the volume of information, in this case the number of points in the DEM.

Spatial autocorrelation acts to reduce the effective number of degrees of freedom in geographic data.

A further case concerns the accommodation of positional uncertainties in the location of the boundaries of a discrete object. Figure 5.14 shows a square approximately 100 m on each side. Suppose the square has been surveyed by determining the locations of its four corner points using GPS, and suppose the circumstances of the measurements are such that there is an RMSE of 1 m in both coordinates of all four points and that errors are independent.

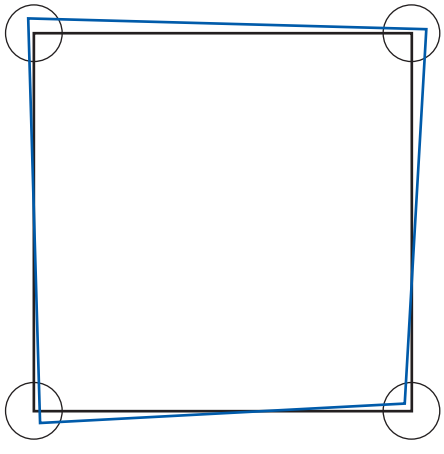
Suppose our task is to determine the area of the square. A GI system can do this easily, using a standard algorithm (see Figure 14.1). Computers are precise (in the sense of Box 5.4) and capable of working to many significant digits, so the calculation might be reported by printing out a number to eight digits, such as 10,014.603 sq m, or even more. But the number of significant digits will have been determined by the precision of the machine, and not by the accuracy of the determination. Box 5.4 summarized some simple rules for ensuring that the precision used to report a measurement reflects as far as possible its accuracy, and clearly those rules will have been violated if the area is reported to eight digits. But what is the appropriate precision?

In this case we can determine exactly how positional accuracy affects the estimate of area. It turns out that area has an error distribution that is Gaussian, with a standard deviation (RMSE), which in our particular case is 200 sq m. In other words, each attempt to measure the area will give a different result, the variation between them having a standard deviation of 200 sq m. This means that the five rightmost digits in the estimate are spurious, including two digits to the left of the decimal point. So if we were to follow the rules of Box 5.4, we would print 10,000 rather than 10,014.603. (Note the problem with standard notation here, which does not let us omit digits to the left of the decimal point even if they are spurious and so leaves some uncertainty about whether or not the tens and units digits are certain—and note also the danger that if the number is printed as an integer it may be interpreted as exactly the whole number.) We can also turn the question around and ask how accurately the points would have to be measured to justify eight digits, and the answer is approximately 0.01 mm, far beyond the capabilities of normal surveying practice.

A useful way of visualizing spatial autocorrelation and interdependence is through animation. Each frame in the animation is a single possible map, or *realization* of the error process. If a point is subject to uncertainty, each realization will show the point in a different possible location, and a sequence of images will show the point shaking around its mean position. If two points have perfectly correlated positional errors, then they will appear to shake in unison, as if they were at the ends of a stiff rod. If errors are only partially correlated, then the system behaves as if the connecting rod were somewhat elastic.

The inherent difficulties in accommodating spatially autocorrelated errors have led many researchers to explore a more general strategy of simulation to evaluate the impacts of uncertainty on the results of spatial analysis. In essence, simulation requires the generation of a series of realizations, as defined earlier. This is often called Monte Carlo simulation in reference to the realizations that occur when dice are tossed or cards are dealt in various games of chance. For example, we could simulate error in a single measurement from a DEM by generating a series of numbers with a mean equal to the measured elevation, and a standard deviation equal to the known RMSE and a Gaussian distribution. Simulation uses everything that is known

Figure 5.14 Error in the measurement of the area of a square 100 m on each side.

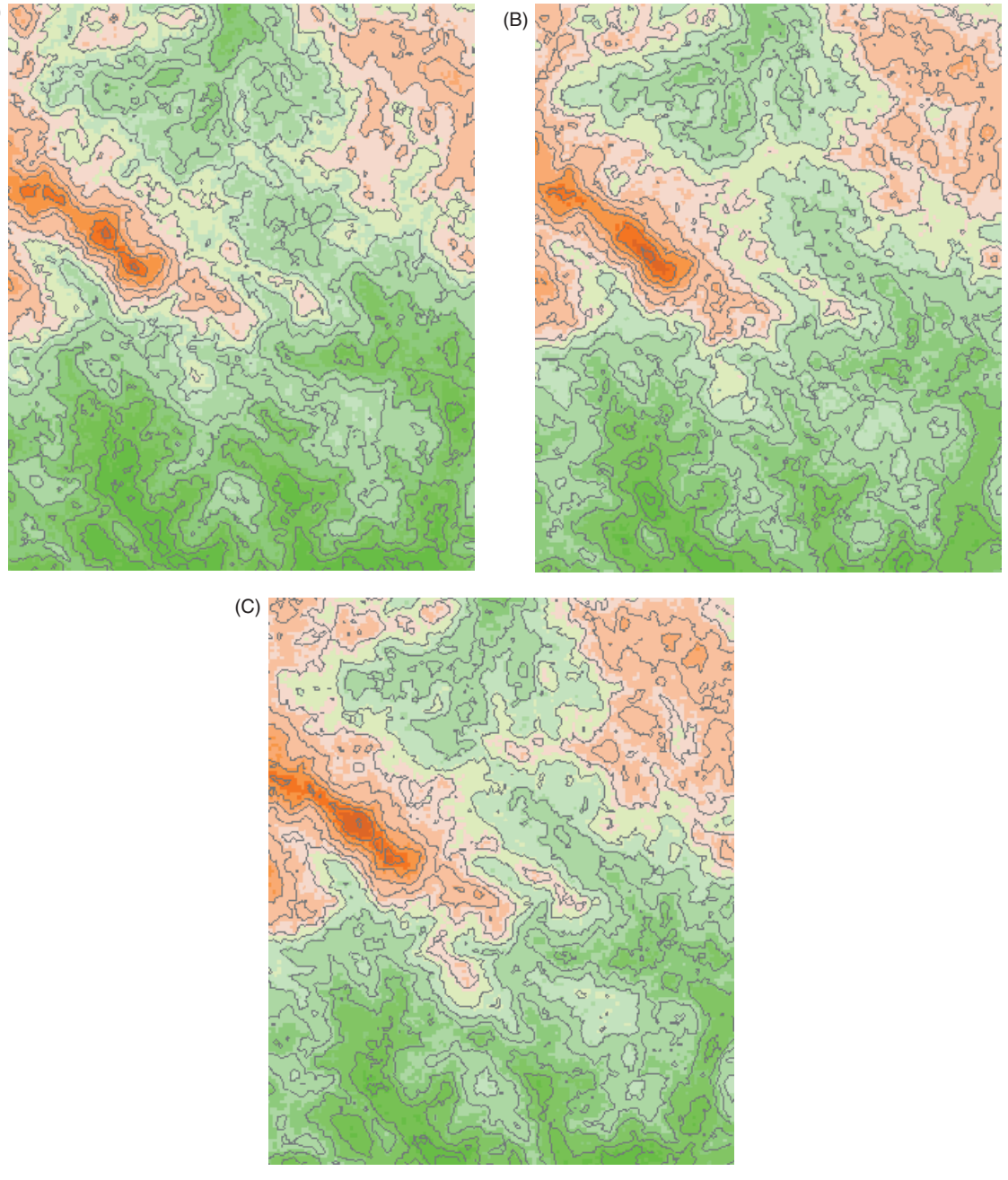


about a situation, so if any additional information is available we would incorporate it in the simulation. We might assume that elevations must be whole numbers of meters, and we would simulate this by rounding the numbers obtained from the Gaussian distribution. With a mean of 350 m and an RMSE of 7 m the results of the simulation might, for example, be 341, 352, 356, 339, 349, 348, 355, 350, . . .

Simulation is an intuitively simple way of getting the uncertainty message across.

Because of spatial autocorrelation, it is impossible in most circumstances to think of databases as decomposable into component parts, each of which can be independently disturbed to create alternative realizations, as in the previous example. Instead, we have to think of the entire database as a realization and create alternative realizations of the database's contents that preserve spatial autocorrelation. Figure 5.15 shows an example, simulating the effects of uncertainty on a digital elevation model. Each of the three realizations is a complete map, and the

Figure 5.15 Three realizations of a model simulating the effects of error on a digital elevation model.



simulation process has faithfully replicated the strong correlations present in errors across the DEM.

5.4.3 Validation through Investigating the Effects of Aggregation and Scale

We have already seen that a fundamental difference between geography and other scientific disciplines is that the definition of its objects of study is only rarely unambiguous and, in practice, rarely precedes our attempts to measure their characteristics. In socio-economic analysis these objects of study (geographic individuals) are usually aggregations because the spaces that human individuals occupy are geographically unique, and confidentiality restrictions usually dictate that uniquely attributable information must be anonymized in some way. Even in natural-environment applications, the nature of sampling in the data collection process (Section 2.4) often makes it expedient to collect data pertaining to aggregations of one kind or another. Thus geographic individuals are likely to be defined as areal aggregations of the units of study. Moreover, in cases where data are collected to serve a range of end uses (as with general population surveys, or natural-resource inventories), the zonal systems are unlikely to be determined with the end point of particular spatial analysis applications in mind.

As a consequence, we cannot be certain in ascribing even dominant characteristics *of* areas to true individuals or point locations *in* those areas. This source of uncertainty is known as the *ecological fallacy* and has long bedeviled the analysis of spatial distributions. (The opposite of ecological fallacy is atomistic fallacy, in which the individual is considered in isolation from his or her environment. This use of the term *ecology* has nothing to do with the way that the term is commonly used today.) The ecological fallacy problem is a consequence of aggregation into the basic units of analysis and is illustrated in Figure 5.16.

Inappropriate inference from aggregate data about the characteristics of individuals is termed the ecological fallacy.

The likelihood of committing ecological fallacy in GI analysis depends on the nature of the aggregation being used. If the members of a set of zones are all perfectly uniform and homogeneous (Section 5.2.1), then the only geographic variation (heterogeneity) that occurs will be between zones. However, nearly all zones are internally heterogeneous to some degree, and greater heterogeneity increases the likelihood and severity of the ecological fallacy problem. That said, it is important to be aware that there are few documented case studies that demonstrate the occurrence of ecological fallacy in practice, and many

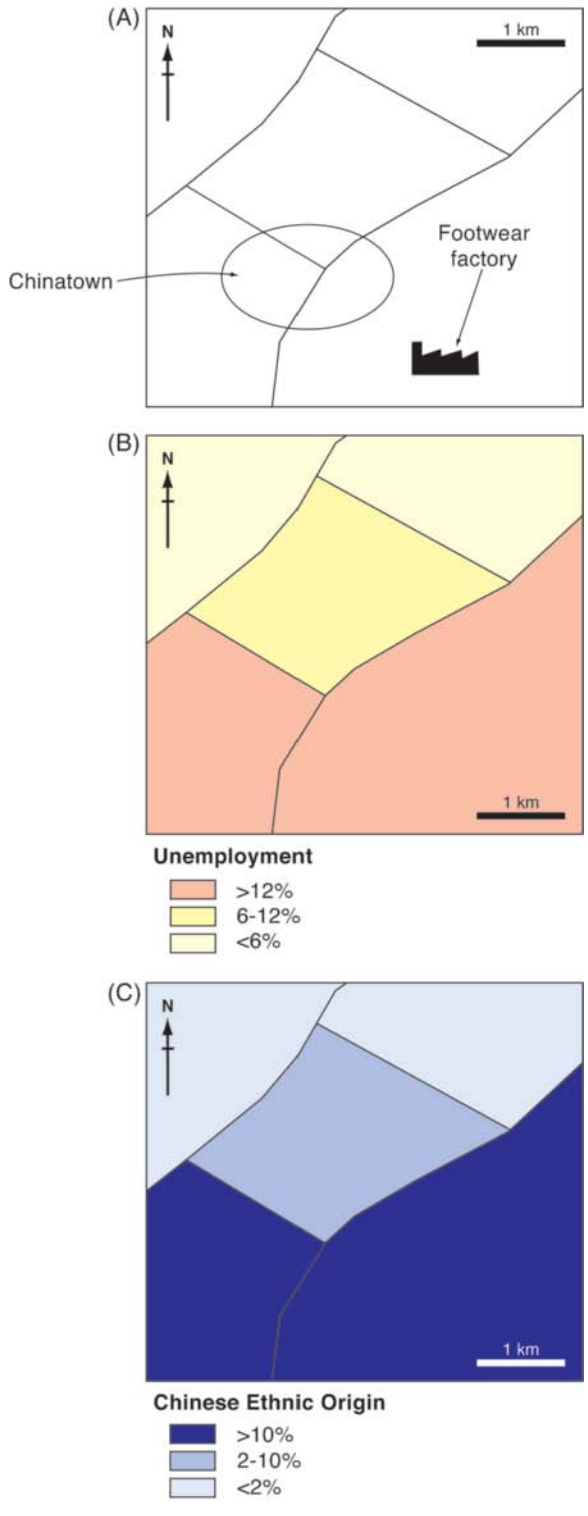


Figure 5.16 A hypothetical illustration of the problem of ecological fallacy. (A) Before it closed down, the footwear factory drew its labor from its local neighborhood and a jurisdiction to the west. The closure caused high unemployment, but not among the service sector workers of Chinatown. Yet comparison of choropleth maps (B) and (C) indicates a spurious relationship between Chinese ethnicity and unemployment.

general-purpose zoning systems in socioeconomic GI are designed to maximize within-zone homogeneity in the most salient population characteristics.

The potential of *aggregation* to create problems in GI analysis is compounded by problems arising from the different *scales* at which zonal systems may be defined. This was demonstrated more than half a century ago in a classic paper by Yule and Kendall, who used data for wheat and potato yields from the (then) 48 counties of England to demonstrate that correlation coefficients tend to increase with scale. They aggregated the 48-county data into zones so that there were first 24, then 12, then 6, and finally just 3 zones. Table 5.3 presents the range of their results, from near zero (no correlation) to over 0.99 (almost perfect positive correlation), although subsequent research has suggested that this range of values is atypical.

Relationships typically grow stronger when based on larger geographic units.

Scale turns out to be important because of the property of spatial autocorrelation outlined in Section 4.3. A succession of subsequent research papers has reaffirmed the existence of similar effects in multivariate analysis. However, rather discouragingly, scale effects in multivariate cases do not follow any consistent or predictable trends.

Further *aggregation* effects can arise in GI analysis because there is a multitude of ways in which the mosaic of basic areal units (the geographic individuals) can be assembled together into zones, and the requirement that zones be made up of spatially contiguous elements presents only a weak constraint on the huge combinatorial range. This gives rise to the related *aggregation* or *zonation* problem, in which different combinations of a given number of geographic individuals into coarser-scale areal units can yield widely different results.

In a classic 1984 study, geographer Stan Openshaw applied correlation and regression analysis to the attributes of a succession of zoning schemes. He demonstrated that the constellation of elemental zones within aggregated areal units could be used to manipulate the results of spatial analysis to a wide range of quite different prespecified outcomes. These numerical experiments have some sinister counterparts in the real world, the most notorious example of which is the political gerrymander of 1812 (see Section 14.1.2). Spatial distributions can be designed (with or without GI systems) that are unrepresentative of the scale and configuration of real-world geographic phenomena; such outcomes may even emerge by chance. The outcome of multivariate spatial analysis is also similarly sensitive to the particular zonal scheme that is used.

Taken together, the effects of scale and aggregation are generally known as the *Modifiable Areal Unit Problem* (MAUP). The ecological fallacy and the MAUP have long been recognized as problems in applied spatial analysis, and through the concept of spatial autocorrelation (Section 2.3), they are also understood to be related problems. Increased technical capacity for numerical processing and innovations in scientific visualization have refined the quantification and mapping of these measurement effects, and have also focused interest on the effects of within-area spatial distributions upon analysis.

Table 5.3 1950 Yule and Kendall's data for wheat and potato yields from 48 English counties.

No. of geographic areas	Correlation
48	0.2189
24	0.2963
12	0.5757
6	0.7649
3	0.9902

5.4.4 Validation with Reference to External Sources: Data Integration and Shared Lineage

The term concatenation is used to describe the integration of two or more different data sources, such that the contents of each are accessible in the product. The polygon overlay operation that will be discussed in Section 13.2.4, and its field-view counterpart, is one simple form of concatenation. The term concatenation is used to describe the range of functions that attempt to overcome differences between datasets, or to merge their contents (as with rubber-sheeting: see Section 8.3.2.2). Conflation thus attempts to replace two or more versions of the same information with a single version that reflects the pooling, or weighted averaging, of the sources. The individual items of information in a single geographic dataset often share lineage, in the sense that more than one item is affected by the same error. This happens, for example, when a map or photograph is registered poorly because all the data derived from it will have the same error. One indicator of shared lineage is the persistence of error—because all points derived from the same misregistration will be displaced by the same, or a similar, amount. Because neighboring points are more likely to share lineage than distant points, errors tend to exhibit strong positive spatial autocorrelation.

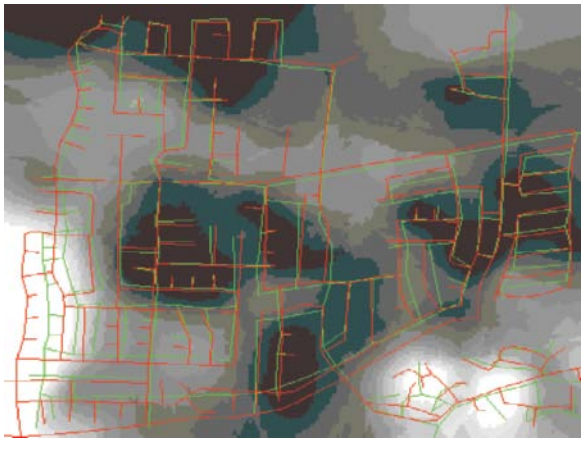
Conflation combines the information from two data sources into a single source.

When two datasets that share no common lineage are concatenated (for example, they have not been subject to the same misregistration), then the relative positions of objects inherit the absolute positional errors of both, even over the shortest distances. Although the shapes of objects in each dataset may be accurate, the relative locations of pairs of neighboring objects may be wildly inaccurate when drawn from different datasets. The anecdotal history of GI is full of examples of datasets that were perfectly adequate for one application, but failed completely when an application required that they be merged with some new dataset that had no common lineage. For example, merging GPS measurements of point positions with streets derived from the U.S. Bureau of the Census TIGER files may lead to surprises where points appear on the wrong sides of streets. If the absolute positional accuracy of a dataset is 50 m, as it was with parts of earlier versions of the TIGER database, points located less than 50 m from the nearest street will frequently appear to be misregistered.

Datasets with different lineages often reveal unsuspected errors when overlaid.

Figure 5.17 shows an example of the consequences of overlaying data with different lineages. In this case, two datasets of streets (shown in green and red) produced by different commercial vendors using their own process fail to match in position by amounts of up to 100 m; they also fail to match in the names of many streets, and even the existence of streets. The integrative functionality of GI systems makes it an attractive possibility to generate multivariate indicators from diverse sources. Such data are likely to have been collected at a range of different scales, and for a range of areal units as diverse as census tracts, river catchments, land ownership parcels, travel-to-work areas, and market research surveys. The uncertainties arising out of overlay, as well as the sources and operation of many other forms of uncertainty, have been investigated by GI scientist Nicholas Chrisman (Box 5.5).

Figure 5.17 Overlay of two street databases for part of Goleta, California (horizontal extent of image c. 5km).



Established procedures of statistical inference can only be used to reason from representative samples to the populations from which they were drawn. Yet these procedures do not regulate the assignment of inferred values to other (usually smaller) zones, or their apportionment to ad hoc regions. There are emergent tensions within socioeconomic analysis, for there is a limit to the usefulness of inferences drawn from conventional, scientifically valid data sources that may be out-of-date, zonally coarse, and irrelevant to the problem under investigation.

Yet the alternative of using new rich sources of crowd-sourced VGI (see Sections 3.8.3 and 10.2), social media, or marketing data may be profoundly unscientific in its inferential procedures.

5.4.5 Internal and External Validation; Induction and Deduction

Reformulation of the MAUP into a *geocomputational* (Section 15.1) approach to zone design amounts to inductive use of GI systems to seek patterns through repeated scaling and aggregation experiments, alongside much better deductive *external* validation using any of the multitude of new datasets that are a hallmark of the information age. Neither of these approaches, used in isolation, is likely to resolve the uncertainties inherent in spatial analysis. Zone-design experiments are merely playing with the MAUP, and most of the new sources of external validation are unlikely to sustain full scientific scrutiny, particularly if they were assembled through nonrigorous survey designs or self-selection of respondents supplying observations.

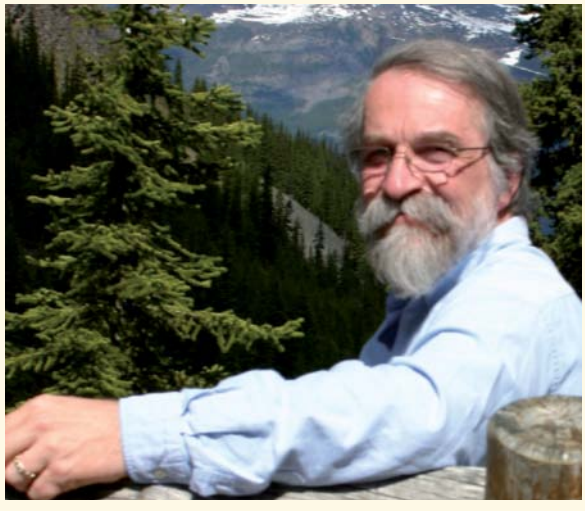
There is no solution to the Modifiable Areal Unit Problem, but simulation of large numbers of alternative zoning schemes can gauge its likely effects.

The conception and measurement of elemental zones, the geographic individuals, may be ad hoc, but it is rarely wholly random either. Can our recognition and understanding of the empirical effects of the MAUP help us to neutralize its effects? Not really. In measuring the distribution of all possible zonally averaged outcomes ("simple random zoning" in analogy to simple random sampling in Section 2.4), there is no tenable analogy with the established procedures of statistical inference and its concepts of precision and error. Even if there were, as we will see

Biographical Box 5.5

Nicholas Chrisman, Globe-Trotting GI Scientist

Nicholas Chrisman (Figure 5.18) joined the geography discipline at what might later be seen as the high tide of the quantitative movement, graduating from the University of Massachusetts in 1972. Following graduation, he took an offer of a one-year position at the Laboratory for Computer Graphics and Spatial Analysis at Harvard University, and in the event stayed there for ten years. His early work entailed programming the topological relations between spatial objects and developing fuzzy tolerance procedures for overlay analysis (Section 13.2.4). He then moved away from software development to attend Bristol University in the UK for a PhD focused on error and data quality, studying there at the same time as two of the authors of this book (Paul Longley and David Maguire). Returning to the United States, he next joined an interdisciplinary team at the University of Wisconsin–Madison, working on the development of GI applications to understand soil erosion. This process extended his understanding of the social and institutional components of GI systems, work that continued after he moved to become Professor of Geography at the University of Washington from 1987–2004.



Courtesy: Linda Chrisman

Figure 5.18 Nicholas Chrisman, international GI scientist.

Nick's next move was to Canada as Scientific Director of the GEOIDE Network, linking researchers and students at 34 universities in a dozen disciplines. During this period, he taught *sciences géomatiques* in French at Université Laval. Between 2013–14 he worked as Discipline Head of Geospatial Sciences at RMIT University in Melbourne, Australia, before moving back to Washington for semi-retirement.

Within this close connection to France over the decades, Chrisman has worked in five countries in three continents. Reflecting on such a variegated career and the development and application of GI systems in his diverse career destinations, Nick remains very mindful that each has shared common scientific endeavors, yet each also has specific local requirements. Or, as he wrote in his classic textbook: "Remember that many marvelous technical advances will not save a system that was not supported by a community that really wanted it" (Chrisman, 2001, *Exploring GIS,* Wiley, p. 243). Look out for a thoroughly revised version of this classic book, infused with these personal experiences, in 2015.

in Section 14.5, there are limits to the application of classic statistical inference to spatial data.

Zoning seems similar to sampling, but its effects are very different.

The way forward seems to entail a threefold response: first, to conduct analysis in response to specific, clearly formulated hypotheses; second, to use GI systems to customize zoning schemes to correspond with these hypotheses; and third, to undertake validation of data with respect to external sources, particularly those that might confirm or refute our assumptions about the likely level of within-zone heterogeneity within our aggregated data. In this way, the MAUP will dissipate if analysts understand the particular areal units that they wish to study.

There is also a sense here that resolution of the MAUP requires acknowledgment of the uniqueness of places. The time dimension is also important: the areal objects of study are ever-changing, and our perceptions of what constitutes an appropriate areal schema should be subject to change. Indeed, infusing the time dimension into GI arguably creates the need for new overarching conceptions and paradigms (meta theories). And finally, within the socioeconomic realm, the act of defining zones can also be self-validating if the allocation of individuals affects the interventions that are subsequently made, be they a mail-shot about a shopping opportunity or deployment of aid under policies to alleviate hardship or deprivation. Spatial discrimination affects spatial behavior, and so the principles of zone design are of much more than academic interest.

5.5 Consolidation

Uncertainty is much more than error. Just as the amount of available digital data and our abilities to process them have developed, so our understanding of the quality of digital depictions of reality has broadened. It is one of the supreme ironies of analysis using contemporary GI systems that as we accrue more and better data and have more computational power at our disposal, we seem to become more uncertain about the quality of our digital representations and the adequacy of our areal units of analysis. Richness of representation provided by Big Data and greater computational power serve to make us more aware of the range and variety of established uncertainties and challenge us to integrate new ones—not least understanding the sources and operation of bias that arises in the assembly of social media datasets.

The only route beyond this impasse is to continue to advance hypotheses about the likely generalized structure of spatial data, albeit in a spirit of humility rather than of conviction. Hypothesis generation requires more than the brute force of high-power computing, and so progress requires greater a priori understanding about the structure in spatial as well as attribute data. There are some general rules to guide us here, and spatial autocorrelation measures provide further structural clues (Section 2.3). In Section 13.2.1 we discuss how context-sensitive spatial analysis techniques, such as geographically weighted regression, provide a bridge between general statistics and the case-study approach.

Geocomputation helps too, by allowing us to gauge the sensitivity of outputs to inputs but, unaided, is unlikely to present unequivocal best solutions. The fathoming of uncertainty requires a combination of the cumulative development of a priori knowledge (we should expect scientific research to be cumulative in its findings), external validation of data sources, and inductive generalization in the fluid, eclectic data-handling environment that is contemporary GI science.

What does all this mean in practice? Here are some rules for how to live with uncertainty. First, because there can be no such thing as perfectly accurate analysis, it is essential to acknowledge that uncertainty is inevitable. It is better to take a positive approach by learning what one can about uncertainty, rather than to pretend that it does not exist. To behave otherwise is unconscionable and can also be very expensive in terms of lawsuits, bad decisions, and the unintended consequences of actions (see Chapter 17).

Second, most spatial analysts often have to rely on others to provide data. Historically, spatial framework data were supplied through government-sponsored mapping programs (e.g., those of the U.S. Geological Survey or Great Britain's Ordnance Survey), but commercial sources and volunteered geographic information (see Section 10.2) have come increasingly to the fore. Many Big Data, such as those from social media sources, are also sometimes georeferenced, but little is known about the degree to which they are representative of any clearly defined population (see Section 2.4). Data should never be taken as the truth; instead, it is essential to assemble all that is known about their quality and to use this knowledge to assess whether the data are fit for use. Metadata (Section 10.2) are designed specifically for this purpose and will often include assessments of quality. When these are not present, it is worth spending the extra effort to contact the creators of the data, or other people who have tried to use them, for advice on quality. Never trust data that have not been assessed for quality, or data from sources that do not have good reputations for quality.

Third, the uncertainties in the outputs of GI analysis are often much greater than one might expect, given knowledge of input uncertainties, because many GI system processes are highly nonlinear. Yet some spatial processes dampen uncertainty rather than amplify it. Given this condition, it is important to gain some impression of the likely impacts of uncertain inputs to GI systems upon outputs.

Fourth, rely on multiple sources of data whenever possible in order to facilitate external validation. It may be possible to obtain maps of an area at several different scales, for example, or to conflate several different open-source databases. Raster and vector datasets are often complementary (e.g., when combining a remotely sensed image with a topographic map). Digital elevation models can often be augmented with spot elevations, or GPS measurements.

> **38.74376% of all statistics are made up (including this one). Avoid spurious precision, and evaluate the provenance of all of the data that you use.**

Finally, be honest and informative in reporting the results of GI analysis. Recognize that uncertainty is never likely to be eliminated, but that it can be managed as part of good scientific practice. Input data sources may be presented with more apparent precision than is justified by their actual accuracy on the ground, and lines may have been drawn on maps with widths that reflect relative importance, rather than uncertainty of position. It is up to the users to redress this imbalance by finding ways of communicating what they know about accuracy, rather than relying on the GI system to do so. It is wise to include plenty of caveats into reported results, so that they reflect what we believe to be true, rather than a narrow and literal interpretation of what the system appears to be saying.

Questions for Further Study

1. What tools do GI system designers build into their products to help users deal with uncertainty? Take a look at your favorite GI system from this perspective. Does it allow you to associate metadata about data quality with datasets? Is there any support to accommodate propagation of uncertainty? How does it determine the number of significant digits when it prints numbers? What are the pros and cons of including such tools?
2. Using aggregate data for Iowa counties, Stan Openshaw found a strong positive correlation between the proportion of people over 65 and the proportion who were registered voters for the Republican Party. What, if anything, does this tell us about the tendency for older people to register as Republicans?
3. Where is the flattest place on Earth? Search the Web for appropriate documents and give reasons for your answer (Figure 5.19).
4. You are a senior retail analyst for Safemart, which is contemplating expansion from its home state to three others in the United States. Assess the relative merits of your own company's store loyalty card data (which you can assume are similar to those collected by any retail chain with which you are familiar) and of data from the most recent Census in planning this strategic initiative. Pay particular attention to issues of survey content, the representativeness of population characteristics, and problems of scale and aggregation. Suggest ways in which the two data sources might complement one another in an integrated analysis.

Courtesy: NASA

Figure 5.19 Salt "flats"?

Further Reading

Burrough, P. A. and Frank, A. U. (eds.). 1996. *Geographic Objects with Indeterminate Boundaries*. London: Taylor and Francis.

Fisher, P. F. 2005. Models of uncertainty in spatial data. In Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. (eds.). *Geographical Information Systems: Principles, Techniques, Management and Applications (Abridged Edition)*. New York: Wiley, pp. 191–205.

Heuvelink, G. B. M. 1998. *Error Propagation in Environmental Modelling with GIS*. London: Taylor and Francis.

Openshaw, S. and Alvanides, S. 2005. Applying geocomputation to the analysis of spatial distributions. In Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. (eds.). *Geographical Information Systems: Principles, Techniques, Management and Applications (Abridged Edition)*. New York: Wiley, pp. 267–282.

Zhang, J. X. and Goodchild, M. F. 2002. *Uncertainty in Geographical Information*. New York: Taylor and Francis.

Smith, B. and Varzi, A. C. 2000. Fiat and bona fide boundaries. *Philosophy and Phenomenological Research* 60(2), pp. 401–420.