1

Working with Spatial Data

1.1 Introduction

The last decades of the twentieth century witnessed the introduction of revolutionary new tools for collecting and analyzing ecological data at the landscape level. The best known of these are the global positioning system (GPS) and the various remote sensing technologies, such as multispectral, hyperspectral, thermal, radar, and LiDAR imaging. Remote sensing permits data to be gathered over a wide area at relatively low cost, and the GPS permits locations to be easily determined on the surface of the earth with sub-meter accuracy. Other new technologies have had a similar impact on some areas of agricultural, ecological, or environmental research. For example, measurement of apparent soil electrical conductivity (EC_a) permits the rapid estimation of soil properties, such as salinity and clay content, for use in agricultural (Corwin and Lesch, 2005) and environmental (Kaya and Fang, 1997) research. The "electronic nose" is in its infancy, but in principle it can be used to detect the presence of organic pollutants (Capelli et al., 2014) and plant diseases (Wilson et al., 2004). Sensor technologies such as these are beginning to find their way into fundamental ecological research.

The impact of these technologies in agricultural crop production has been strengthened by another device of strictly agricultural utility: the yield monitor. This device permits "on the go" measurement of the flow of harvested material. If one has spatially precise measurements of yield and of the factors that affect yield, it is natural to attempt to adjust management practices in a spatially precise manner to improve yield or reduce costs. This practice is called *site-specific crop management*, which has been defined as the management of a crop at a spatial and temporal scale appropriate to its natural variation (Lowenberg-DeBoer and Erickson, 2000).

Technologies such as remote sensing, soil electrical conductivity measurement, the electronic nose, and the yield monitor are precision measurement instruments at the landscape scale, and they allow scientists to obtain landscape-level data at a previously unheard-of spatial resolution. These data, however, differ in their statistical properties from the data collected in traditional plot-based ecological and agronomic experiments. The data often consist of thousands of individual values, each obtained from a spatial location either contiguous with or very near to those of neighboring data. The advent of these massive spatial data sets has been part of the motivation for the further development of the theory of spatial statistics, particularly as it applies to ecology and agriculture.

Although some of the development of the theory of spatial statistics has been motivated by large data sets produced by modern measurement technologies, there are many older data sets to which the theory may be profitably applied. Figure 1.1 shows the locations of a subset of about one-tenth of a set of 4,101 sites taken from the Wieslander survey

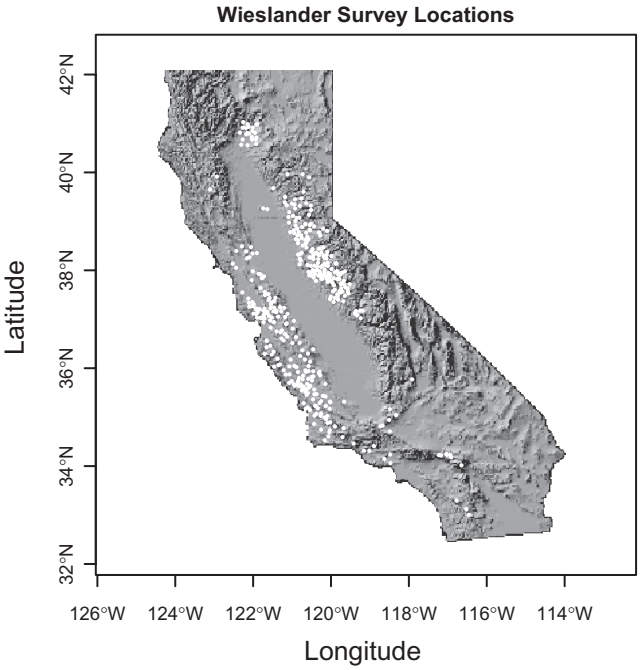

FIGURE 1.1
Locations of survey sites taken from the Wieslander vegetation survey. The white dots represent one-tenth of the 4,101 survey locations used to characterize the ecological relationships of oak species in California oak woodlands. The hillshade map was downloaded from the University of California, Santa Barbara Biogeography Lab, http://www.biogeog.ucsb.edu/projects/gap/gap_data_state.html, and reprojected in ArcGIS.

(Wieslander, 1935). This was a project led by Albert Wieslander of the US Forest Service that documented the vegetation in about 155,000 km^2 of California, or roughly 35% of its area. Allen et al. (1991) used a subset of the Wieslander survey to develop a classification system for California's oak woodlands, and subsequently Evett (1994) and Vayssières et al. (2000) used these data to compare methods for predicting oak woodland species distributions based on environmental properties.

Large spatial data sets like those arising from modern precision landscape measuring devices or from large-scale surveys such as the Wieslander survey present the analyst with two major problems. The first is that, even in cases where the data satisfy the assumptions of traditional statistics, traditional statistical inference is often meaningless with data sets having thousands of records because the null hypothesis is almost always rejected (Matloff, 1991). Real data never have *exactly* the hypothesized property, and the large size of the data set makes it impossible to be close enough. This would seem to make these data sets ideal for modern nonparametric methods associated with "Big Data." This approach, however, is often defeated by the second problem, which is that the data records from geographically nearby sites are often more closely related to each other than they are to data records from sites farther away. Moreover, if one were to record some data attribute at a location in between two neighboring sample points, the resulting attribute data value would likely be similar to those of both of its neighbors, and thus would be limited in the amount of information it added. The data in this case are, to use the statistical term, *spatially autocorrelated*. We will define this term more precisely and discuss its consequences in greater detail in Chapter 3, but it is clear that these data

violate the assumption of both traditional statistical analysis and modern nonparametric analysis that data values or errors are mutually independent. The theory of spatial statistics concerns itself with these sorts of data.

Data sets generated by modern precision landscape measurement tools frequently have a second, more ecological, characteristic that we will call "low ecological resolution." This is illustrated in the maps in Figure 1.2, which are based on data from an agricultural field. These three thematic maps represent measurements associated respectively with soil properties, vegetation density, and reproductive material. The figures are arranged in a rough sequence of influence. That is, one can roughly say that soil properties influence vegetative growth, and vegetative growth influences reproductive growth. These influences,

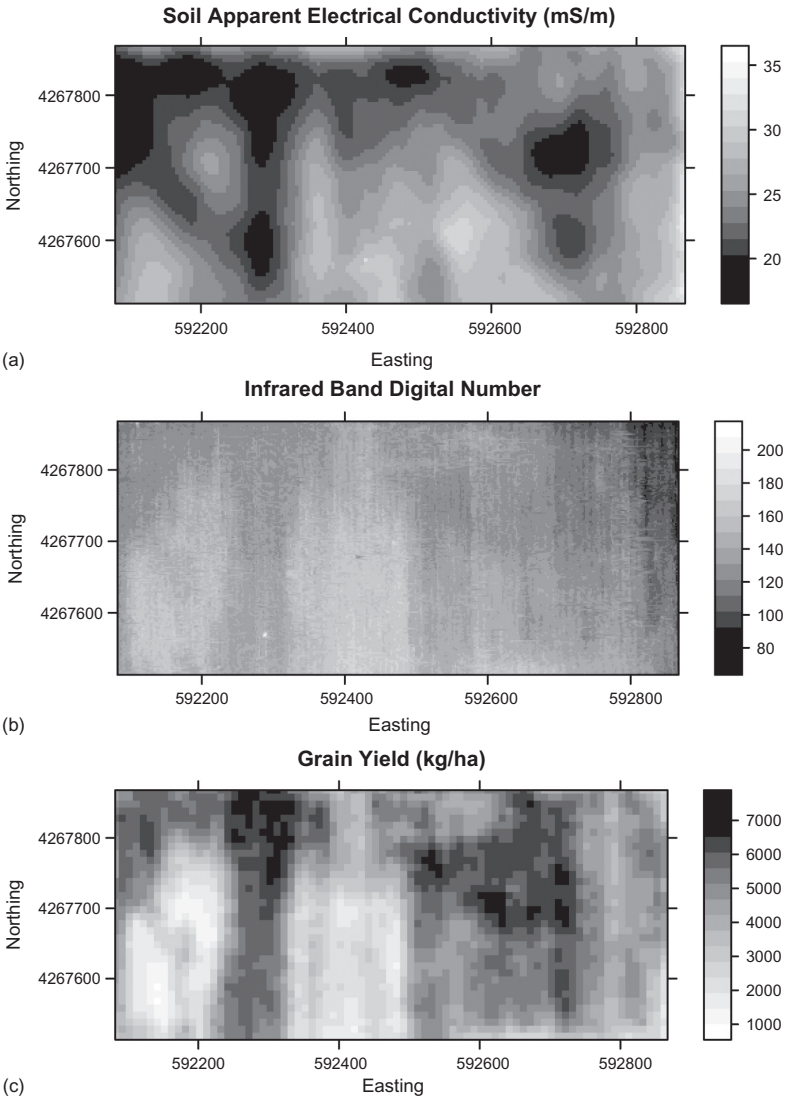

FIGURE 1.2
Spatial plots of three data sets from Field 2 of Data Set 4, a wheat field in Central California: (a) gray-scale representation of apparent soil electrical conductivity (mS m^{-1}); (b) digital number value of the infrared band of an image of the field taken in May, 1996; and (c) yield map (kg ha^{-1}).

however, are very complex and nonlinear. Three quantities are displayed in Figure 1.2: soil apparent electrical conductivity (EC_a) in Figure 1.2a, infrared reflectance (IR digital number) in Figure 1.2b, and grain yield in Figure 1.2c. The complexity of the relationship between soil, vegetation, and reproductive material is reflected in the relationship between the patterns in Figures 1.2a, 1.2b, and 1.2c, which is clearly not a simple one (Ball and Frazier, 1993; Benedetti and Rossini, 1993). In addition to the complexity of the relationships among these three quantities, each quantity itself has a complex relationship with more fundamental properties like soil clay content, leaf nitrogen content, and so forth. To say that the three data sets in Figure 1.2 have a very high spatial resolution but a low "ecological resolution" means that there is a complex relationship between these data and the fundamental quantities that are used to gain an understanding of the ecological processes in play. The low ecological resolution often makes it necessary to supplement these high spatial resolution data with other, more traditional data sets that result from the direct collection of soil and plant data.

The three data sets in Figure 1.2, like many spatial data sets, are filled with suggestive visual patterns. The two areas of low grain yield in the southern half of the field in Figure 1.2c clearly seem to align with patterns of low infrared reflectance in Figure 1.2b. However, there are other, more subtle patterns that may or may not be related between the maps. These include an area of high EC_a on the west side of the field that may correspond to a somewhat similarly shaped area of high infrared reflectance, an area of high apparent EC_a in the southern half of the field centered near 592,600E that may possibly correspond to an area of slightly lower infrared reflectance in the same location, and a possible correspondence between low apparent EC_a, low infrared reflectance, and low grain yield at the east end of the field. Scrutinizing the maps carefully might reveal other possibilities. The problem with searching for matching patterns like these, however, is that the human eye is famously capable of seeing patterns that are not really there. Also, the eye may miss subtle patterns that do exist. For this reason, a more objective means of detecting true patterns must be used, and statistical analysis can often provide this means.

Data sets like the ones discussed in this book are often unreplicated, relying on observations rather than controlled experiments. Every student in every introductory applied statistics class is taught that the three fundamental concepts of experimental design, introduced by Sir Ronald Fisher (1935), are randomization, replication, and blocking. The primary purpose of replication is to provide an estimate of the variance (Kempthorne, 1952, p. 177). In addition, replication helps to ensure that experimental treatments are interspersed, which may reduce the impact of confounding factors (Hurlbert, 1984). Some statisticians might go so far as to say that data that are not randomized and replicated are not worthy of statistical analysis.

In summary, we are dealing with data sets that may not satisfy the traditional assumptions of statistical analysis, that often involve relations between measured quantities for which there is no easy interpretation, that tempt the analyst with suggestive but potentially misleading visual patterns, and that frequently result from unreplicated observations. In the face of all these issues, one might ask why anyone would even consider working with data like these, when traditional replicated small plot experiments have been used so effectively for over a hundred years. A short answer is that replicated small plot experiments are indeed valuable and can tell us many things, but they cannot tell us everything. Often phenomena that occur at one spatial scale are not observable at another, and sometimes complex interactions that are important at the landscape scale lose their importance at the plot scale. Richter et al. (2009) argue that laboratory experiments conducted under highly standardized conditions can produce results with little validity beyond the specific

environment in which the experiment is conducted, and this same argument can also be applied to small plot field experiments. Moreover, econometricians and other scientists for whom replicated experiments are often impossible have developed powerful statistical methods for analyzing and drawing inferences from observational data. These methods have much to recommend them for ecological and agricultural use as well. The fundamental theme of this book is that methods of spatial statistics, many of them originally developed with econometric and other applications in mind, can prove useful to ecologists and crop and soil scientists. These methods are certainly not intended to replace traditional replicated experiments, but when properly applied they can serve as a useful complement, providing additional insights, increasing the scope, and serving as a bridge from the small plot to the landscape. Even in cases where the analysis of data from an observational study does not by itself lead to publishable results, such an analysis may provide a much-improved focus to the research project and enable the investigator to pose much more productive hypotheses.

1.2 Analysis of Spatial Data

1.2.1 Types of Spatial Data

Cressie (1991, p. 8) provides a very convenient classification of spatial data into three categories, which we will call (1) geostatistical data, (2) areal data, and (3) point pattern data. These can be explained as follows. Suppose we denote by D a site (the *domain*) in which the data are collected. Let the locations in D be specified by position coordinates (x, y), and let $Y(x, y)$ be a quantity or vector of quantities measured at location (x, y). Suppose, for example, the domain D is the field in Figure 1.2. The components of Y would then be the three quantities mapped in the figure together with any other measured quantities. The first of Cressie's (1991) categories, *geostatistical data*, consists of data in which the components of $Y(x, y)$ vary continuously in the spatial variables (x, y). The value of Y is measured at a set of points with coordinates (x_i, y_i), $i = 1, \ldots, n$. A common goal in the analysis of geostatistical data is to interpolate Y at points where it is not measured.

The second of Cressie's (1991) categories consists of data that are defined *only* at a set of locations, which may be points or polygons, in the domain D. If the locations are polygons, then the data are assumed to be uniform within each polygon. If the data are measured at a set of points then one can arrange a mosaic or lattice of polygons that each contain one or more measurement points and whose areal aggregate covers the entire domain D. For this reason, these data are called *areal data*. In referring to the spatial elements of areal data, we will use the terms *cell* and *polygon* synonymously. A *mosaic* is an irregular arrangement of cells, and a *lattice* is a regular rectangular arrangement.

It is often the case that a biophysical process represented by areal data varies continuously over the landscape (Schabenberger and Gotway, 2005, p. 9). Thus, the same data set may be treated in one analysis as geostatistical data and in another as areal data. Figure 1.3 illustrates this concept. The figure shows the predicted values of clay content in the same field as that shown in Figure 1.2. Clay content was measured in this field by taking soil cores on a rectangular grid of square cells 61 m apart. Figure 1.3a shows an interpolation of the values obtained by kriging (Section 6.3.2). This is an implementation of the geostatistical model. Figure 1.3b shows the same data modeled as a lattice of discrete

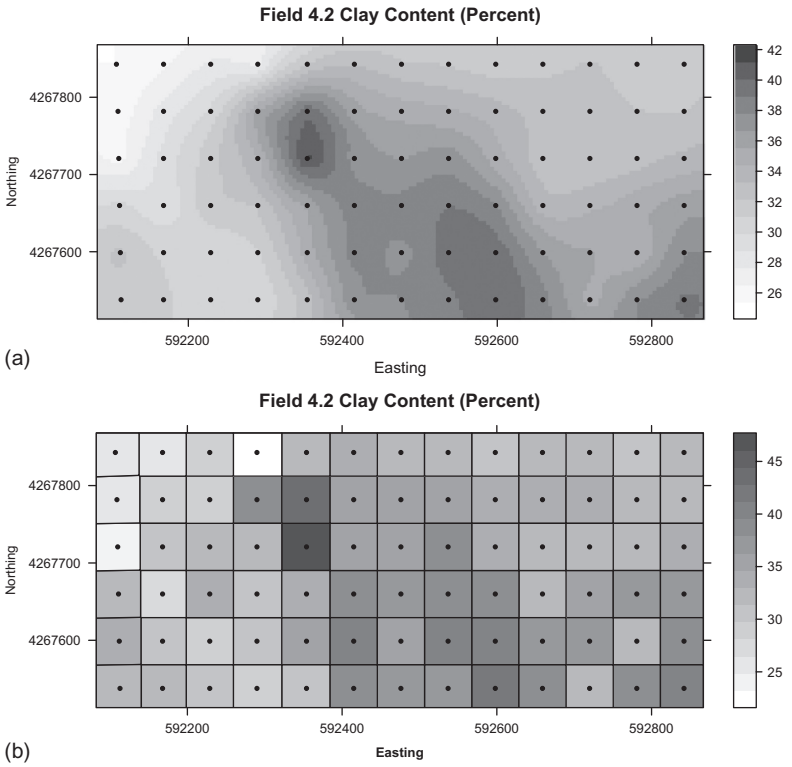

FIGURE 1.3
Predicted values of soil clay content in Field 2 of Data Set 4 using two different spatial models: (a) interpolated values obtained using kriging based on the geostatistical model and (b) polygons based on the areal model. The soil sample locations are shown as black dots.

polygons. In both figures, the actual data collection locations are shown as black dots. It happens that each polygon in Figure 1.3b contains exactly one data location, but in other applications, data from more than one location may be aggregated into a single polygon, and the polygons may not form a regular lattice.

The appropriate data model, geostatistical or areal, may depend on the objective of the analysis. If the objective is to accurately predict the value of clay content at unmeasured locations, then generally the geostatistical model would be used. If the objective is to determine how clay content and other quantities influence some response variable such as yield, then the areal model may be more appropriate. In many cases, it is wisest not to focus on the classification of the data into one or another category but rather on the objective of the analysis and the statistical tools that may be best used to achieve this objective.

The third category of spatial data consists of points in the domain D that are defined by their location, and for which the primary questions of interest involve the pattern of these locations. Data falling in this category, such as the locations of members of a species of tree, are called *point patterns*. For example, Figure 1.6 below shows a photograph of oak trees in a California oak woodland. If each tree is considered as a point, then point pattern analysis is concerned with the pattern of the spatial distribution of the trees. When spatial data are collected at irregularly located points, as is the case with the Wieslander data, knowledge of their pattern may be helpful in determining whether the sampling points have introduced a bias into the data. The primary focus of this book is on geostatistical

and areal data, which we will jointly refer to as *georeferenced* data. We do, however, occasionally discuss point pattern data analysis, primarily as a means of better understanding georeferenced data.

1.2.2 The Components of Spatial Data

Georeferenced data have two components. The first is the *attribute component*, which is the set of all the measurements describing the phenomenon being observed. The second is the *spatial component*, which consists of that part of the data that at a minimum describes the location at which the data were measured and may also include other properties such as spatial relationships with other data. The central problem addressed in this book is the incorporation into the analysis of attribute data of the effect of spatial relationships. As a simple example, consider the problem of developing a linear regression model relating a response variable to a single environmental explanatory variable. Suppose that there are two measured quantities. The first is leaf area index, or LAI, the ratio between total leaf surface area and ground surface area, at a set of locations. We treat this as a response variable and represent it with the symbol Y. Let Y_i denote the LAI measured at location i. Suppose that there is one explanatory variable, say, soil clay content. Explanatory variables will be represented using the symbol X, possibly with subscripts if there are more than one of them. We can postulate a simple linear regression relation of the form

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \; i = 1, \ldots, n. \tag{1.1}$$

There are several ways in which spatial effects can influence this model. One very commonly occurring one is a reduction of the *effective sample size*. Intuition for this concept can be gained as follows. Suppose that the n data values are collected on a square grid and that $n = m^2$ where m is the number of measurement locations on a side. Suppose now that we consider sampling on a $2m$ by $2m$ grid subtending the same geographic area. This would give us $4n$ data values and thus, *if these values were independent*, provide a much larger sample and greater statistical power. One could consider sampling on a $4m$ by $4m$ grid for a still larger sample size, and so on *ad infinitum*. Thus, it would seem that there is no limit to the power attainable if only we are willing to sample enough. The problem with this reasoning is that the data values are *not* independent: values sampled very close to each other will tend to be similar. If the values are similar, then the errors ε_i will also be similar. This property of similarity among nearby error values violates a fundamental assumption on which the statistical analysis of regression data is based, namely, the independence of errors. This loss of independence of errors is one of the most important properties of spatial data, and will play a major role in the development of methods described in this book.

1.2.3 Spatial Data Models

In the previous two sections, we have defined georeferenced data as data that contains an attribute component and a spatial component and for which a value can be assigned to any location in the domain D. Section 1.2.1 described one model for a geographic system that might generate such data, namely, the areal data model in which every point in the domain D is contained in a polygon, and the mosaic or lattice of these polygons covers the entire domain. This representation is called the *vector data model* (Lo and Yeung, 2007, p. 87).

It works well for systems in which large contiguous area have uniform or nearly uniform values (such as data defined by state or province) and less well for systems in which the attribute values vary continuously in space.

The second commonly used model in the analysis of georeferenced data is the *raster model* (Lo and Yeung, 2007, p. 83). In this model, the data are represented by a grid of raster cells, each of which contains one attribute value. The data represented in Figure 1.2 may be considered as exemplary of raster data. Much of the analysis in this book will be carried out using the vector model, but we will also on numerous occasions use the raster model, which more accurately represents georeferenced data that vary continuously in space. It is wise to develop a facility with both models in order to be able to select the one most appropriate for the given situation and analytical objective.

1.2.4 Topics Covered in the Text

The topics covered in this book are ordered as follows. The statistical software used for almost all of the analysis in the book is R (R Development Core Team, 2017). No familiarity with R is assumed, and Chapter 2 contains an introduction to the R package. The property of spatial data described in the previous paragraph, that attribute values measured at nearby locations tend to be similar, is called *positive spatial autocorrelation*. In order to develop a statistical theory that allows for spatial autocorrelation, one must define it precisely. This is the primary topic of Chapter 3. That chapter also describes the principal effects of spatial autocorrelation on the statistical properties of data. Having developed a precise definition of autocorrelation, one can measure its magnitude in a particular data set. Chapter 4 describes some of the most widely used measures of spatial autocorrelation and introduces some of their uses in data analysis.

Beginning with Chapter 5, the chapters of this book are organized in the same general order as that in which the phases of a study are carried out. The first phase is the actual collection of the data. Chapter 5 contains a discussion of sampling plans for spatial data. The second phase of a study is the preparation of the data for analysis, which is covered in Chapter 6. This is a much more involved process with spatial data than with non-spatial data, first because the data sets must often be converted to a file format compatible with the analytical software; second, because the data sets must be georegistered (i.e., their locations must be aligned with the points on the earth at which they were measured); third, because the data are often not measured at the same location or on the same scale (this is the problem of *misalignment*); and fourth, because in many cases the high volume and data collection process make it necessary to remove outliers and other "discordant" data records. After data preparation is complete, the analyses can begin.

Statistical analysis of a particular data set is often divided into exploratory and confirmatory components. Chapters 7 through 9 deal with the exploratory component. Chapter 7 discusses some of the common graphical methods for exploring spatial data. Chapter 8 describes the use of linear methods, including multiple regression and the general linear model. These methods are here used in an exploratory context only since spatial autocorrelation is not taken into account. The discussion of regression analysis also serves as a review on which to base much of the material in subsequent chapters.

Chapter 9 continues the application of non-spatial methods to spatial data, in this case focusing on nonparametric methods, that is, methods that do not make assumptions about the structure of the random component of the data. These include generalized additive models, recursive partitioning, and random forest. Chapters 8 and 9 serve as a link between the exploratory and confirmatory components of the analysis. The focus of Chapter 10 is

the effect of spatial autocorrelation on the sample variance. The classical estimates based on independence of the random variables are no longer appropriate, and in many cases an analytical estimate is impossible. Although approximations can be developed, in many cases bootstrap estimation provides a simple and powerful tool for generating variance estimates. Chapter 11 enters fully into confirmatory data analysis. One of the first things one often does at this stage is to quantify, and test hypotheses about, association between measured quantities. It is at this point that the effects of spatial autocorrelation of the data begin to make themselves strongly felt. Chapter 12 describes these effects and discusses some methods for addressing them.

In Chapter 12, we address the incorporation of spatial effects into models for the relationship between a response variable and its environment through the use of mixed model theory. Chapter 13 describes autoregressive models, which have become one of the most popular tools for statistical modeling. Chapter 14 provides an introduction to Bayesian methods of spatial data analysis, which are rapidly gaining in popularity as their advantages become better exploited. Chapter 15 provides a very brief introduction to the study of spatiotemporal data. Chapter 16 describes methods for the explicit incorporation of spatial autocorrelation into the analysis of replicated experiments. Although the primary focus of this book is on data from observational studies, if one has the opportunity to incorporate replicated data, one certainly should not pass this up. Chapter 17 describes the summarization of the results of data analysis into conclusions about the processes that the data describe. This is the goal of every ecological or agronomic study, and the methods for carrying out the summarization are very much case dependent. All we can hope to do is to try to offer some examples.

1.3 The Data Sets Analyzed in This Book

Every example in this book is based on one of four data sets. There are a few reasons for this arrangement. Probably the most important is that one of the objectives of the book is to give the reader a complete picture, from beginning to end, of the collection and analysis of spatial data characterizing an ecological process. A second is that these data sets are not ideal "textbook" data. They include a few warts. Real data sets almost always do, and one should know how to confront them. Also, sometimes they yield ambiguous results. Again, this can happen with real data sets and one must know what to do when it does happen.

Two of the data sets involve unmanaged ecosystems and two involve cultivated ecosystems (ecologists should also study the cultivated systems and agronomists should also study the natural systems). All four data sets come from studies in which I have participated. This is not because I feel that these studies are somehow superior to others that I might have chosen, but rather because at one time or another my students and collaborators and I have lived with these data sets for an extended period, and I believe one must live with a data set for a long time to really come to know it and analyze it effectively. The book is not a research monograph, and none of the analytical objectives given here is an actual current research objective. Their purpose is strictly expository. In that context, although the discussion of each method has the ostensible purpose of furthering the stated research objective, the true purpose is to expose the methods to the reader. This does have one beneficial side effect. In real life, the application of a particular method to a particular data set sometimes does not lead to any worthwhile results. That sometimes happens with the analyses in this book, and I hope this gives the reader a more realistic impression of the process of data analysis.

The present section contains a brief description of each of the four data sets. A more thorough, "Materials and Methods" style description is given in Appendix B. The data sets, along with the R code to execute the analyses described in this book, are available on the book's website http://psfaculty.plantsciences.ucdavis.edu/plant/sda2.htm. The data are housed in four subfolders of the folder *data*, called *Set1*, *Set2*, *Set3*, and *Set4*. In addition to a description of the data sets, Appendix B also contains the R statements used to load them into the computer. These statements are therefore not generally repeated in the body of the text.

Two other data folders, which are also provided on the website, appear in the R code: *auxiliary* and *created*. Readers should be able to generate these files for themselves, but they are made available just in case. The folder *created* contains files that are created as a result of doing the analyses and exercises in the book. The folder *auxiliary* contains data downloaded from the Internet. Large, publicly available data sets play a very important role in the analysis of ecological data, and one of the objectives of this book is to introduce the reader to some of these data sets and how they can be used.

1.3.1 Data Set 1: Yellow-Billed Cuckoo Habitat

This data set was collected as a part of a study of the change over a 50 year period in the extent of suitable habitat for the western yellow-billed cuckoo (*Coccyzus americanus occidentalis*), a California state listed endangered species that is one of California's rarest birds (Gaines and Laymon, 1984). The yellow-billed cuckoo, whose habitat is the interior of riparian forests, once bred throughout the Pacific Coast states. Land use conversion to agriculture, urbanization, and flood control have, however, restricted its current habitat to the upper reaches of the Sacramento River. This originally meandering river has over most of its length been constrained by the construction of flood control levees. One of the few remaining areas in which the river is unrestrained is in its north central portion (Figure 1.4).

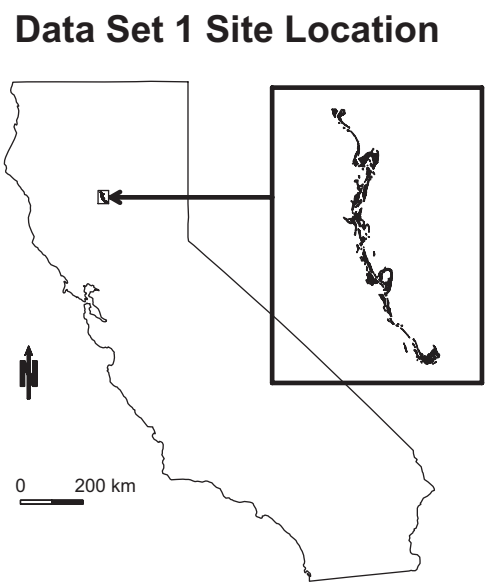

FIGURE 1.4
Location of study area of yellow-billed cuckoo habitat in California, in the north-central portion of the Sacramento River.

The data set consists of five explanatory variables and a response variable. The explanatory variables are defined over a polygonal map of the riparian area of the Sacramento River between river-mile 196, located at Pine Creek Bend, and river-mile 219, located near the Woodson Bridge State Recreation Area (although metric units are used exclusively in this book, the river-mile is a formal measurement system used to define locations along a river). This map was created by a series of geographic information system (GIS) operations as described in Appendix B.1. Figure 1.5 shows the land cover mosaic in the northern end of the study area as measured in 1997. The white areas are open water and the light gray areas are either managed (lightest gray) or unmanaged (medium gray) areas of habitat unsuitable for the yellow-billed cuckoo. Only the darkest areas contain suitable vegetation cover. These areas are primarily riparian forest, which consists of dense stands of willow and cottonwood. The response variable is based on cuckoo encounters at each of the 21 locations situated along the river.

Our objective in the analysis of Data Set 1 is to test a model for habitat suitability for the yellow-billed cuckoo. The model is based on a modification of the California Wildlife Habitat Relationships (CWHR) classification system (Mayer and Laudenslayer, 1988). This is a classification system that "contains life history, geographic range, habitat relationships, and management information on 694 species of amphibians, reptiles, birds, and mammals known to occur in the state" (https://www.wildlife.ca.gov/Data/CWHR). The modified CWHR model is expressed in terms of habitat patches, where a *patch* is defined as "a geographic area of contiguous land cover." Photographs of examples of the habitat types used in the model are available on the CWHR website given above. Our analysis will test a habitat suitability index based on a CWHR model published by Laymon and Halterman (1989). The habitat suitability model incorporates the following explanatory variables: (1) patch area, (2) patch width, (3) patch distance to water, (4) within-patch ratio of high vegetation to medium and low vegetation, and (5) patch vegetation species. The original discussion of this analysis is given by Greco et al. (2002).

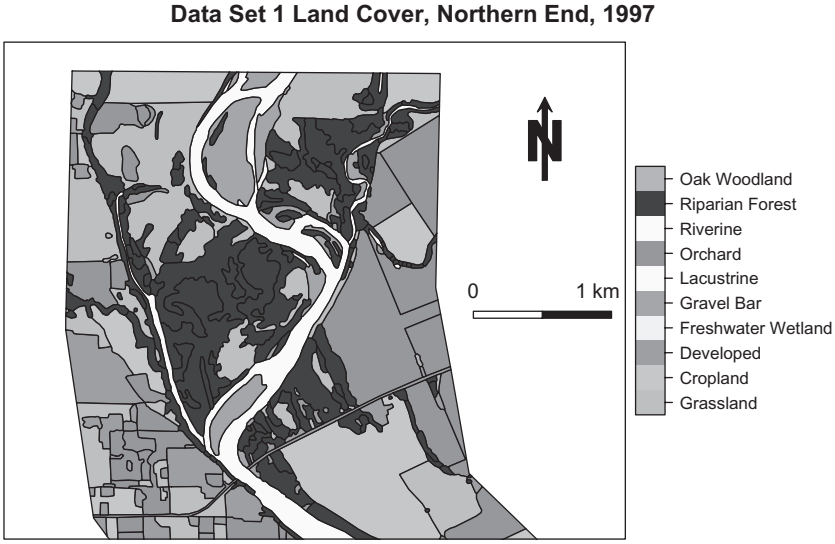

FIGURE 1.5
Land use in 1997 at the northern end of the region along the Sacramento River from which Data Set 1 was collected. The white areas are open water (riverine = flowing, lacustrine = standing). The light gray areas are unsuitable; only the charcoal colored area is suitable riparian forest.

The scope of inference of this study is restricted to the upper Sacramento River, since this is the only remaining cuckoo habitat. Data Set 1 is the only one of the four data sets for which the analysis objective is predictive rather than explanatory. Moreover, the primary objective is not to develop the best (by some measure) predictive model but rather to test a specific model that has already been proposed. As a part of the analysis, however, we will try to determine whether alternative models have greater predicative power or, conversely, whether any simpler subsets of the proposed model are its equal in predictive power.

1.3.2 Data Set 2: Environmental Characteristics of Oak Woodlands

Hardwood rangelands cover almost four million hectares in California, or over 10% of the state's total land area (Allen et al., 1991; Waddell and Barrett, 2005). About three-quarters of hardwood rangeland is classified as oak woodland, defined as areas dominated by oak species and not productive enough to be considered timberland (Waddel and Barret, 2005, p. 12). Figure 1.6 shows typical oak woodland located in the foothills of the Coast Range. Blue oak (*Quercus douglasii*) woodlands are the most common hardwood rangeland type, occupying about 500,000 ha. These woodlands occupy an intermediate elevation between the grasslands of the valleys and the mixed coniferous forests of the higher elevations. Their climate is Mediterranean, with hot, dry summers and cool, wet winters. Rangeland ecologists have expressed increasing concern over the apparent failure of certain oak species, particularly blue oak, to reproduce at a rate sufficient to maintain their population (Standiford et al., 1997; Tyler et al., 2006; Zavaleta et al., 2007; López-Sánchez et al., 2014). The appearance in California in the mid-1990s of the pathogen *Phytophthora ramorum*, associated with the condition known as sudden oak death, has served to heighten this concern (Waddel and Barrett, 2005).

The reason or reasons for the decline in oak recruitment rate are not well understood. Among the possibilities are climate change, habitat fragmentation, increased seedling predation resulting from changes in herbivore populations, changes in fire regimes, exotic plant and animal invasions, grazing, and changes in soil conditions (Tyler et al., 2006). One component of an increased understanding of blue oak ecology is an improved characterization of the ecological factors associated with the presence of mature blue oaks. Although changes in blue oak recruitment probably began occurring coincidentally with

FIGURE 1.6
Oak woodlands located in the foothills of the Coast Range of California. (Photograph by the author).

the arrival of the first European settlers in the early nineteenth century (Messing, 1992), it is possible that an analysis of areas populated by blue oaks during the first half of the twentieth century would provide some indication of environmental conditions favorable to mature oaks. That is the general objective of the analysis of Data Set 2. This data set consists of records from 4,101 locations surveyed as a part of the Vegetation Type Map (VTM) survey in California, carried out during the 1920s and 1930s (Wieslander, 1935; Jensen, 1947; Allen et al., 1991a; Allen-Diaz and Holzman, 1991), depicted schematically in Figure 1.1. Allen-Diaz and co-workers (Allen-Diaz and Holzman, 1991) entered these data into a database, and Evett (1994) added geographic coordinates and climatic data to this database. The primary goal of the analysis of Data Set 2 is to characterize the environmental factors associated with the presence of blue oaks in the region sampled by the VTM survey. The original analysis was carried out by Vayssières et al. (2000).

The intended scope of inference of the analysis is that portion of the state of California where blue oak recruitment is possible. As a practical matter, this area can be divided roughly into four regions (Appendix B.2 and Figure 1.1): the Coast Range on the west side of the Great Central Valley, the Klamath Range to the north, the Sierra Nevada on the east side of the valley, and the Traverse Ranges to the south. The analysis in this book will be restricted to the Coast Range and the Sierra Nevada, with the Klamath Range and the Traverse Ranges used as validation data.

1.3.3 Data Set 3: Uruguayan Rice Farmers

Rice (*Oriza sativa* L.) is a major Uruguayan export crop. It is grown under flooded conditions. Low dykes called *taipas* are often used to promote even water levels (Figure 1.7). Rice is generally grown in rotation with pasture, with a common rotation being 2 years of rice followed by 3 years of pasture. Data Set 3 contains measurements taken in 16 rice fields in eastern Uruguay (Figure 1.8) farmed by twelve different farmers over a period of three growing seasons. Several locations in each field were flagged and measurements were made

FIGURE 1.7
A typical Uruguayan rice field. The low dykes, called *taipas*, are used to promote more even flooding. (Courtesy of Alvaro Roel).

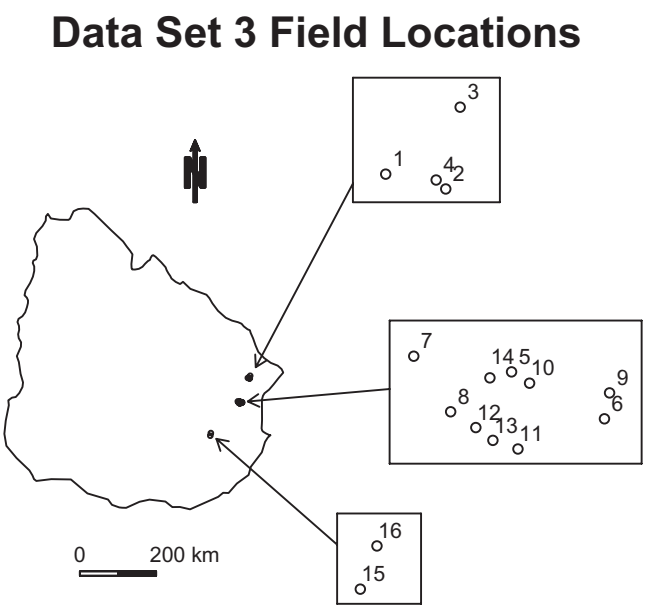

FIGURE 1.8
Map of Uruguay, showing locations of the 16 rice fields measured in Data Set 3. Insets are not drawn to scale.

at each location of soil and management factors that might influence yield. Hand harvests of rice yield were collected at these locations. Regional climatic data were also recorded.

Some of the farmers in the study obtained consistently higher yields than others. The overall goal of the analysis is to try to understand the management factors primarily responsible for variation in yields across the agricultural landscape. A key component of the analysis is to distinguish between management factors and environmental factors with respect to their influence on crop yield. We must take into account the fact that a skillful farmer, or one with access to more resources, might manage a field not having conditions conducive to high yield in a different way from that in which the same farmer would manage a field with conditions favorable to high yield. For example, a skillful farmer growing a crop on infertile soil might apply fertilizer at an appropriate rate, while a less skillful farmer growing a crop on a fertile soil might apply fertilizer at an inappropriate rate. The less skillful farmer might obtain a better yield, but this would be due to the environmental conditions rather than the farmer's skill. Conversely, a farmer, skillful or otherwise, who could not afford to purchase fertilizer might not apply it at a rate sufficient to maximize yield. The original analysis was carried out by Roel et al. (2007). The scope of the study is intended to include those rice farmers in eastern Uruguay who employ management practices similar to those of the study participants. For simplicity, we assume that the farmers' objective is to maximize yield rather than profit.

1.3.4 Data Set 4: Factors Underlying Yield in Two Fields

One of the greatest challenges in site-specific crop management is to determine the factors underlying observed yield variability. Data Set 4 was collected as a part of a four-year study initiated in two fields in Winters, California, which is located in the Central Valley.

This was the first major precision agriculture study in California, and the original objective was simply to determine whether or not the same sort of spatial variability in yield existed in California as had been observed in other states. We will pursue a different objective in this book, namely, to determine whether we can establish which of the measured explanatory variables influenced the observed yield variability. Data from one of the fields (designated Field 2) is shown in Figure 1.2.

In the first year of the experiment, both fields were planted to spring wheat (*Triticum aestivum* L.). In the second year, both fields were planted to tomato (*Solanum lycopersicum* L.). In the third year, Field 1 was planted to bean (*Phaseolus vulgaris* L.) and Field 2 was planted to sunflower (*Helianthus annuus* L.), and in the fourth year Field 1 was planted to sunflower and Field 2 was planted to corn (*Zea mays* L.). Each crop was harvested with a yield monitor equipped harvester so that yield maps of the fields were available. False color infrared aerial photographs (i.e., images where infrared is displayed as red, red as green, and green as blue) were taken of each field at various points in the growing season. Only those from the first season are used in this book. During the first year, soil and plant data were collected by sampling on a square grid 61 m in size, or about two sample points per hectare. In addition, a more densely sampled soil electrical conductivity map of Field 2 was made during the first year of the experiment.

As was mentioned in the discussion of Data Set 2, the climate in California's Central Valley is Mediterranean, with hot, essentially rain-free summers and cool, wet winters. Spring wheat is planted in the late fall and grown in a supplementally irrigated cropping system, while the other crops, all of which are grown in the summer, are fully irrigated. Both fields were managed by the same cooperator, a highly skilled farmer who used best practices as recommended by the University of California. Thus, while the objective of the analysis of Data Set 3 is to distinguish the most effective management practices among a group of farmers, the objective of the analysis of Data Set 4 is to determine the factors underlying yield variability in a pair of fields in which there is no known variation in management practices.

The original analysis of the data set was carried out by Plant et al. (1999). The scope of the study, strictly speaking, extends only to the two fields involved. Obviously, however, the real interest lies in extending the results to other fields. Success in achieving the objective would not be to establish that the same factors that influence yield in these fields were also influential in other fields, which is not the case anyway. The objective is rather to develop a methodology that would permit the influential factors in other fields to be identified.

1.3.5 Comparing the Data Sets

The four data sets present a variety of spatial data structures, geographical structures, attribute data structures, and study objectives. Let's take a moment to compare them, with the idea that by doing so you might get an idea as to how they relate to your own data set and objectives. Data Sets 2 and 3 have the simplest spatial structure. Both data sets consist of a single set of data records, each of which is the aggregation of a collection of measurements made at a location modeled as a point in space. Data Set 2 contains quantities that are measured at different spatial scales. This data set covers the largest area and presents the greatest opportunity to incorporate publicly available data into the analysis. The spatial structure of Data Set 1 is more complex in that it consists of data records that are modeled as properties of a mosaic of polygons. Data Set 1 is the only data set that includes this polygonal data structure in the raw data. Data Set 4 does not include any polygons, but it does include raster data sets. In addition, Data Set 4 includes point data sets measured

at different locations. Moreover, the spatial points that represent the data locations in different point data sets actually represent the aggregation of data collected over different areas. This combination of types of data, different locations of measurement, and different aggregation areas leads to the problem of *misalignment* of the data, which will be discussed frequently in this book, but primarily in Chapter 6.

Moving on to geographic structure, Data Set 4 is the simplest. The data are recorded in two fields, each of size less than 100 ha. We will use Universal Transverse Mercator (UTM) coordinates (Lo and Yeung, 2007, p. 43) to represent spatial location. If you are not familiar with UTM coordinates, you need to learn about them. In brief, they divide the surface of the earth into 120 *Zones*, based on location in the northern or southern hemisphere. Within each hemisphere, UTM coordinates are based on location within 60 bands running north and south, each of width six degrees of longitude. Within each zone, location is represented by an *Easting* (x coordinate) and a *Northing* (y coordinate), each measured in meters. Locations in Data Sets 1 and 3 also are represented in UTM coordinates, although both of these data sets cover a larger geographic area than that of Data Set 4. Depending on the application, locations in Data Set 2 may be represented in either degrees of longitude and latitude or UTM coordinates. This data set covers the largest geographic area. Moreover, the locations in Data Set 2 are in two different UTM Zones.

1.4 Further Reading

Spatial statistics is a rapidly growing field, and the body of literature associated with it is growing rapidly as well. The seminal books in the field are Cliff and Ord (1973) and, even more importantly, Cliff and Ord (1981). This latter monograph set the stage for much that was to follow, providing a firm mathematical foundation for measures of spatial autocorrelation and for spatial regression models. Ripley (1981) contains coverage of many topics not found in Cliff and Ord (1981). It is set at a slightly higher level mathematically but is nevertheless a valuable text even for the practitioner. Besides Cliff and Ord (1981), the other seminal reference is Cressie (1991). This practically encyclopedic treatment of spatial statistics is at a mathematical level that some may find difficult, but working one's way through it is well worth the effort. Griffith (1987) provides an excellent introductory discussion of the phenomenon of spatial autocorrelation. Upton and Fingleton (1985) provide a wide array of examples for applications-oriented readers. Their book is highly readable while still providing considerable mathematical detail.

Among more recent works, the two monographs by Haining (1990, 2003) provide a broad range of coverage and are very accessible to readers with a wide range of mathematical sophistication. Legendre and Legendre (1998) touch on many aspects of spatial data analysis. Griffith and Lane (1999) present an in-depth discussion of spatial data analysis through the study of a collection of example data sets. Waller and Gotway (2004) discuss many aspects of spatial data analysis directly applicable to ecological problems. Lloyd (2010) gives an introduction to spatial statistics for those coming from a GIS background. Chun and Griffith (2013) provide a very nice, brief undergraduate-level introduction to spatial statistics. Oyana and Margai (2015) provide a data-centric treatment of the subject.

There are also a number of collections of papers on the subject of spatial data analysis. Two early volumes that contain a great deal of relevant information are those edited by Bartels and Ketellapper (1979) and Griffith (1989). The collection of papers edited by

Arlinghaus (1995) has a very applications oriented focus. The volume edited by Longley and Batty (1996), although it is primarily concerned with GIS modeling, contains a number of papers with a statistical emphasis. Similarly, the collection of papers edited by Stein et al. (1999), although it focuses on remote sensing, is broadly applicable. The volume edited by Scheiner and Gurevitch (2001) contains many very useful chapters on spatial data analysis. Fotheringham and Rogerson (2009) have assembled papers by primary contributors on a number of topics in spatial analysis into a single handbook.

The standard introductory reference for geostatistics is Isaaks and Srivastava (1989). Webster and Oliver (1990, 2001) provide coverage of geostatistical data analysis that is remarkable for its ability to combine accessibility with depth of mathematical treatment. Another excellent source is Goovaerts (1997). Journel and Huijbregts (1978) is especially valuable for its historical and applied perspective. Pielou (1977) is an excellent early reference on point pattern analysis. Fortin and Dale (2005) devote a chapter to the topic. Diggle (1983) provides a mathematically more advanced treatment.

A good discussion of observational versus replicated data is provided by Kempthorne (1952). Sprent (1998) discusses hypothesis testing for observational data. Gelman and Hill (2007, Chapters 9 through 10) discuss the use of observational data in the development of explanatory models. Bolker (2008) and Zuur et al. (2007) are good sources for analysis of ecological data.

For those unfamiliar with GIS concepts, Lo and Yeung (2007) provide an excellent introduction. In addition to Lo and Yeung (2007), which we use as the primary GIS reference, there are many others who provide excellent discussions of issues such as map projections, Thiessen polygons, spatial resampling, and so forth. Prominent among these are de Smith et al. (2007), Bonham-Carter (1994), Burrough and McDonnell (1998), Clarke (1990), and Johnston (1998). Good GIS texts at a more introductory level include Chang (2008), Demers (2009), and Longley et al. (2001).