7

Preliminary Exploration of Spatial Data

7.1 Introduction

"Exploratory analysis is detective work." This is how Tukey (1977, p. 1) opens the book *Exploratory Data Analysis*, and it would be impossible to write a better introductory sentence. Tukey (1977) expands the analogy between data analysis and the criminal justice system. The initial phase involves the search for evidence and is generally carried out by the police. Once the evidence has been gathered, the evaluation of the strength of that evidence is the responsibility of the judge and jury. Data analysis can also be divided into two phases (which are often carried out by the same person or team). The exploratory phase examines the properties of the data and attempts to uncover patterns and indications. The confirmatory phase tests the strength of these patterns and indications and in so doing addresses questions about the population from which the sample data were drawn.

One must be careful, however, not to press the analogy between data analysis and criminal justice too far. Ordinarily, the phases of the criminal justice process are more or less sequential, and once the case has gone before the judge and jury the investigative phase ends. With data analysis, particularly when dealing with spatial data, the process is more flexible. After an initial exploratory phase, the exploratory and confirmatory processes to some extent co-occur, with results of the confirmatory phase often motivating further exploration. The main danger that must be avoided is that of "data snooping," that is, of letting the confirmatory process be influenced inappropriately by the exploratory process through an introduced bias. Such an inappropriate influence would occur, for example, if the investigator selected for a hypothesis test subsets of the data that appeared in the exploratory phase to satisfy the hypothesis.

Exploratory data analysis (EDA) consists of a collection of tools, and the question of which tool to employ depends on the data and on the objective of the study. EDA is best introduced by means of examples. This also provides a good way to introduce the data sets that we will be studying throughout the rest of the book. In the contexts of the simulated research projects that we carry through the book, the objective in this chapter is to initiate data exploration by graphical means. In a real data analysis program, one would likely cycle through applications of graphical analysis and applications of the tools in the chapters to follow. To the extent that this is feasible we will do so in the book.

Each of the four data sets consists of a mixture of several different kinds of data, and in order to guide the initial exploration it will be useful to establish a set of categories into which these kinds fall. These categories can be arranged into a rough sequence of presumed influence (Figure 7.1). The use of association to infer influence has a long and inglorious history (Sprent, 1969, p. 29; Box et al., p. 487), and so we first emphasize that when we discuss influence, we must do so from the perspective of the ecological scientist

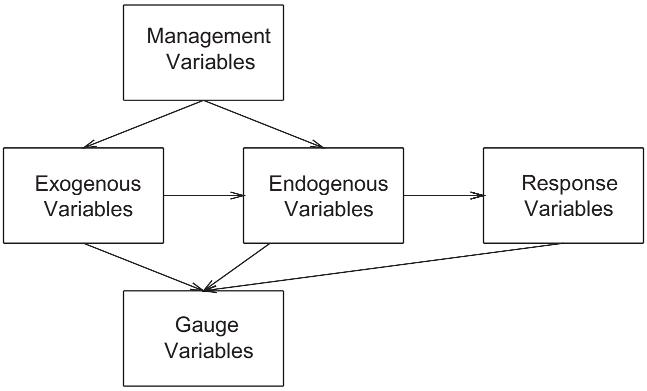

FIGURE 7.1
Classification of types of variable for use in guiding exploratory analysis. This classification provides assistance in avoiding the confounding of variables that play different roles in the ecological processes governing the behavior of the system.

rather than the data analyst. Assumptions about influence relationships must be made based on the ecologist's knowledge of process and mechanism and used to inform the analysis of the data, rather than using a blind analysis of the data to inform conclusions about influence. Failure to consider process and mechanism in the data analysis and to, for example, carry out a multiple regression in which all possible predictors are simply lumped together with no thought to their ecological relationships, will lead to conclusions that are at best confusing and at worst misleading.

I find it useful in thinking about process and influence to schematically divide the variables of the model into five categories: exogenous, endogenous, management, response, and gauge variables. To avoid unnecessary confusion, I must point out that my use of the terms *endogenous* and *exogenous*, although motivated by econometrics (Wooldridge, 2013, p. 87), is not the same as their use in this discipline. To illustrate the five categories, we will use as an example a data set describing a crop in a farmer's field, such as Data Set 3 or 4. At one end of the presumed chain of influence are *exogenous variables*, those variables that measure the influence of the environment on the processes that directly influence the values of the response variables. Soil clay content and daily high temperature are examples of exogenous variables. At the other end of the influence chain are the *response variables* (or variable; often there is only one). Crop yield is an example in Data Sets 3 and 4. In between the exogenous variables and the response variables are *endogenous variables*, which are influenced by the exogenous variables and, in turn, influence the response variables. Leaf nitrogen content in Data Set 4 is an example of an endogenous variable. In some data sets, the values of one or more of the exogenous, endogenous, or response variables may be represented by *gauge variables*. These are variables that do not directly play a role in the chain of influence, but whose values represent variables that do play such a role. The aerial images and soil electrical conductivity measurements in Data Set 4 are examples. Finally, human influence on the system may enter in the form of *management variables*. These are exogenous variables that are implemented through human intervention. Fertilizer rates and irrigation effectiveness in Data Set 3 are examples. This subdivision into categories is intended to help guide thinking, not to constrain or confuse it. The selection of categories is often a matter of taste, and, to paraphrase Cressie (1991, p. 114), one person's endogenous variable may be another person's exogenous variable.

Systems such as Data Sets 1 and 2, which do not involve human intervention to produce an end product, offer a different challenge in assessing interrelationships among variables. Although some "natural" systems may include management variables, they are not present in either of these data sets. If some form of resource were provided, such as artificial bird habitat in Data Set 1 or planting and maintenance of oak seedlings in Data Set 2, then the intensity of these interventions would be a management variable. An example of a gauge variable in either data set would be a remotely sensed quantity such as a vegetation index in aerial or satellite images. With cuckoo presence or count as the response variable in Data Set 1, there are arguably no true endogenous variables. If bird's nests could be identified, these would serve as an endogenous variable in Data Set 1. In Data Set 2, we can consider variables such as *Elevation* and *CoastDist*, the shortest distance to the Pacific coast, as exogenous variables since they do not influence blue oak presence directly.

7.2 Data Set 1

For convenience, we will repeat in our discussion of each data set some of the descriptive material that was presented in Chapter 1. Data Set 1 was collected as a part of a study of the properties of habitat suitable for the western yellow-billed cuckoo (*Coccyzus americanus occidentalis*), a California state-listed endangered species. The bird, whose habitat is the interior of riparian forests, once bred throughout the Pacific Coast states of the USA. Land use conversion to agriculture, urbanization, and flood control have, however, restricted its current habitat to the upper reaches of the Sacramento River (Figure 1.3). Greco et al. (2002) provide the following descriptive characterization of suitable habitat: "a mosaic of riparian forest vegetation consisting of willow (*Salix* spp.) and cottonwood (*Populus fremontii*) forests in combination with open-water habitats such as an oxbow lake or backwater channel. Dense vegetation less than 20 m in height is especially important for nesting while both low and high vegetation are used for foraging." Cuckoos nest in large forest patches (greater than 5–20 ha), and they are not found in parts of the patch less than 100 m wide (Laymon and Halterman, 1989), although they can cross gaps in the forest of less than 100 m (Gaines and Laymon 1984).

The Sacramento River is a naturally meandering river that has over most of its length been constrained by levees; the upper reaches are the only remaining unconstrained portion. There has been discussion of the possibility of instituting a levee setback system in which the levees would be relocated to permit some movement of the lower reaches of the river, and there is interest in estimating the effect of this policy on yellow-billed cuckoo habitat. This requires a habitat suitability model. Such a model has been developed, based on years of observation, by Laymon and Halterman (1989). The objective of the analysis of Data Set 1 is to test this model based on the combination of habitat variables and bird observation data. The data set consists of five explanatory variables (the habitat data) and a response variable (the bird observation data). A detailed description of the processes used to create the data set is given in Appendix B.1. For convenience, some of that information is summarized here.

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. This area can be visualized in the USGS Earth Explorer (https://earthexplorer.usgs.gov/) by searching for "Woodson Bridge State Recreation Area."

The explanatory variables are 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” (http://www.dfg.ca.gov/biogeodata/cwhr). The modified CWHR model is expressed in terms of habitat patches, where a *patch* is defined for the purpose of the cuckoo habitat model as “a geographic area of the following contiguous land cover categories: riparian, freshwater emergent wetland, or lacustrine.” These definitions correspond to the CWHR classifications VRI, FEW, and LAC, respectively. Photographs of examples of these habitat types are available on the CWHR website. The shapefile containing the habitat patches is included in the data as the file *habitatpatches.shp*. Code for loading this and the other data files into R is given in Appendix B.1.

The original habitat suitability model of Laymon and Halterman (1989) incorporates the following predictor 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. Variables 1–4 were incorporated directly into the model tested here. Variable 5, patch vegetation species, was interpreted as follows (Greco et al., 2002, p. 187). As a river like the Sacramento meanders, it alters the characteristics of the land surrounding it. After the meandering of the river has exposed a formerly submerged site, the land goes through a series of successional stages (Conard et al., 1977). These generally pass from an open gravel bar to a shrub-dominated thicket to a willow/cottonwood forest to a mixed willow/cottonwood oak woodland to a mature oak woodland. From the perspective of habitat quality for the yellow-billed cuckoo, the most important aspect of this succession is the fraction of the patch dominated by the willow/cottonwood mixture. It was impossible to distinguish willow/cottonwood from oak in the aerial photographs used to create the land cover maps. Therefore, floodplain age was used as a surrogate variable for stand composition. Floodplain age was estimated by a method described by Greco et al. (2007). Portions of a patch with a high floodplain age were taken to indicate land that passed to oak dominance. Portions of a patch with a low floodplain age were taken to indicate land that was still dominated by gravel bar and shrubs. A calibration of this model indicated that habitat quality of a patch could be well represented by the ratio of the area of the patch of age less than 60 years to the total area of the patch. The optimal ratio was found to be between 0.67 and 0.8.

The habitat suitability model of Laymon and Halterman (1989) was implemented using the scheme shown in Table 7.1. Each patch was assigned a suitability score for each of the five variables, and the overall patch suitability was computed as the geometric mean of the individual scores (Greco et al., 2002). The geometric mean was used because it emphasizes patches that satisfy all of the criteria (Cooperrider, 1986). The data set in the file *set1data.shp* contains the explanatory variables. This file was constructed as described in Appendix B.1.

TABLE 7.1

Habitat Patch Suitability Model for the Yellow-Billed Cuckoo.

Suitability Class	Suitability Score	Patch Area (ha)	Patch Width (m)	Patch Distance to Water (m)	Height Class Ratio (H:H+L+M)	Floodplain Age Ratio <60
Optimum	1.0	>80	>600	<100	0.45–0.55	0.67–0.8
Suitable	0.66	40–80	200–600	<100	0.2–0.45, 0.55–0.67	0.5–0.67, 0.8–0.875
Marginal	0.33	17–40	100–200	<100	<0.2	0.375–0.5, >0.875
Unsuitable	0.0	<17	<100	>100	>0.67	<0.375

Since the classification variable *distance to water* has only one separation value, which classifies land greater than 100 m from water as unsuitable (Table 7.1), this variable is not explicitly included in the attribute data. Instead, a buffer of 100 m around water polygons was created, and land use polygons that fell outside this buffer and did not touch a polygon that overlapped the buffer were eliminated.

Since the objective of the study is to analyze a habitat suitability model for the western yellow-billed cuckoo, it is evident that the response variable must be some measure of bird population or presence. A difficulty in this regard is that the bird is, to quote Gaines (1974, p. 204), "furtive, and thus easily overlooked." The commonly used manner of surveying for the presence or absence of the yellow-billed cuckoo is to visit a site, play a tape recording of the bird's *kowlp* call, and observe whether or not there is a response (Gaines, 1974). The standard procedure (Greco, 1999, p. 232) is to play the call five to ten times at each observation point. The call is audible for a distance of about 300 m (Laymon and Halterman, 1987). Failure to elicit a response is taken to imply that no cuckoos are present at that location. The response may be simply a bird call and may not include an actual sighting. The data set used in this study is an aggregation of observations collected over a period between 1978 and 1992 (Gaines and Laymon, 1984; Laymon and Halterman, 1987; Halterman, 1991; Greco, 2002). The survey includes a total of 21 locations within the study area (Figure 7.2). There are two forms of observation data that can be used as a response variable. The first is presence/absence, that is, whether or not a cuckoo observation occurs at a location. The second is the number of observations. Although they are suggestive and should not be completely ignored, the count data are fraught with practical difficulties. One is that it is impossible to determine with certainty the number of individual birds represented by these observations. Moreover, even if this relationship could be determined, it is not clear how one would interpret the difference between a single observation of one bird, multiple observations of the same bird, and multiple observations of different birds.

A second serious problem with the interpretation of the bird count data is that it is an *extensive* property. The number of birds counted in a particular area may depend on the size of the area. An extensive property is one that depends on the size of the medium containing it, while an *intensive* property does not (Lewis and Randall, 1961). For example, in the context of ecology, population is an extensive property while population density is an intensive property. Generally speaking, an intensive property can be derived from an extensive property by dividing the extensive property by an appropriate denominator. For example, population density is derived from population by dividing by area of the region containing the population. It is easy to create misleading thematic maps by either displaying an extensive property when an intensive property would be more appropriate, or by using an inappropriate denominator to derive an intensive property from an extensive one (Kronmal, 1993; Henning, 2003, p. 195).

Although there is no guarantee that, if everything else is fixed, bird count in a patch will be proportional to patch area, this quantity is nevertheless a logical choice. Patch area is a predictor variable in the CWHR model, so choosing this as the denominator for count data would have the effect of putting this quantity on both sides of the equals sign, or, to put it another way, to enforce an assumption that population density depends on patch area. In our initial exploration and analysis of the data, we will consider only presence-absence data. We will use the count data, without normalization by patch area, in Section 8.4.4 as an example of the construction of regression models for count data, but we will have to keep in mind the weaknesses of the data set.

Following the standard procedure of Appendix B.1, data are loaded as spatial features (`sf`) objects and converted to spatial (`sp`) objects for plotting the figures in the book. As discussed

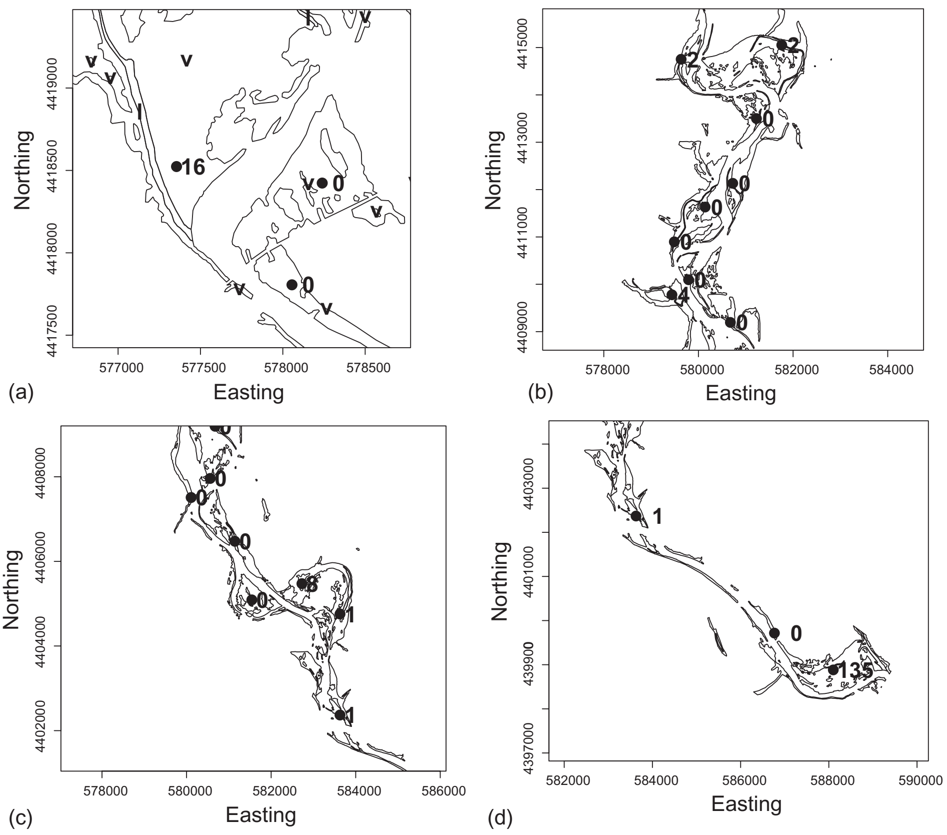

FIGURE 7.2
Habitat patches and locations of cuckoo observation points along the Sacramento River in the far northern portion of Data Set 1. The number next to each location signifies the total number of cuckoo observations in the pooled survey data. Parts (a) through (d) show regions from north to south. The letters in the Figure 7.2a are the vegetation type codes: v = riparian forest, and l = lacustrine. These are not shown in the other figures to avoid clutter.

in Section 2.6, however, we will also get a quick overview of the data using the `sf` function `plot()` as well as calls to the function `ggplot()`. The code for this section, as well as the other sections, includes calls to these functions, although the results are not shown.

The quick call to the `sf` function `plot()` shows that the study site is extremely long and narrow (Figure 1.3). Figure 7.2 shows the site divided into four zones of roughly equal length that include all observation points. The locations of the observation points are indicated, along with the number of birds observed at each location. The code to create Figure 7.2a is as follows. The `SpatialPolygonsDataFrame` `data.Set1.patches` contains the contents of the shapefile *habitatpatches.shp*, and the `SpatialPointsDataFrame` `data.Set1.obs` is created from the contents of the file *obspts.csv* (Appendix B.1).

```
> plot(data.Set1.patches.sp, axes = TRUE, # Fig. 7.2a
+    xlim = c(577000,578500), ylim = c(4417500,4419400))
> title(xlab = "Easting", ylab = "Northing", cex.lab = 1.5)
> points(data.Set1.obs, pch = 19, cex = 2)
> text(coordinates(data.Set1.obs)[,1] + 100,
+    coordinates(data.Set1.obs)[,2],
+    labels=as.character(data.Set1.obs$Abund), cex = 2, font = 2)
```

```
> y <- lapply(data.Set1.patches.sp@polygons, slot, "labpt")
> patches.loc <- matrix(0, length(y), 2)
> for (i in 1:length(y)) patches.loc[i,] <- unlist(y[[i]])
> text(patches.loc,
+   labels=as.character(data.Set1.patches.sp$VegType), cex=2, font=2)
```

The code using `lapply()` described in Section 2.4.3 is used to extract the attribute data values so that they can be plotted as character values in the appropriate locations. This is only done in Figure 7.2a; the other figures contain so many polygons that displaying all of their values would create too much clutter. Inspection of Figure 7.2d indicates that the southernmost observation point has an anomalously large number of sightings. Either there is something unusual about this location that makes the cuckoos really love it or there is a problem with the data.

We can get an idea of the relationships among the data by examining the small region at the northern end of the site. Figure 7.3 shows thematic maps of height class, cover class, and age class in the northern end, corresponding to Figure 7.2a. The code to create Figure 7.3a is the following. The `SpatialPolygonsDataFrame` `data.Set1.sp` contains the shapefile *set1.data.shp*.

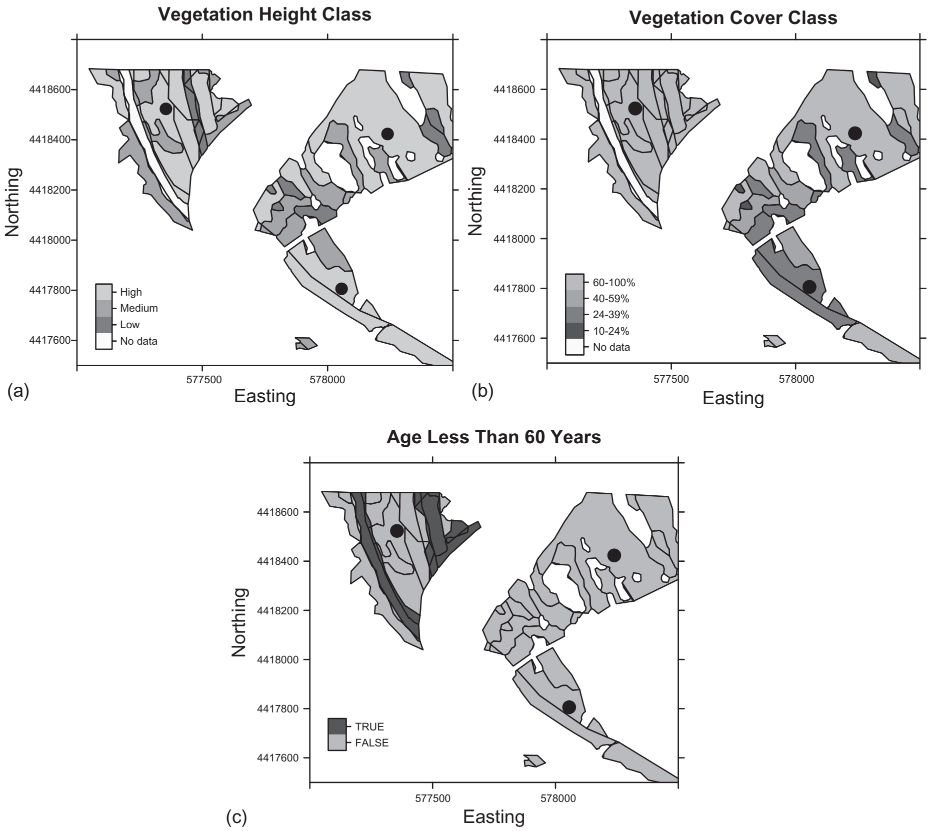

FIGURE 7.3
Geometry of the habitat patches and observation points in the northern area (Figure 7.2a): thematic maps created with the function `spplot()` distinguishing (a) vegetation size classes, (b) cover classes, and (c) age classes. The large black circles indicate patches in which cuckoo presence/absence was recorded.

```
> levels(data.Set1.sp$HtClass)
[1] "0" "h" "l" "m"
> data.Set1.sp@data$HtClass2 <-
+    ordered(as.character(data.Set1.sp@data$HtClass),
+    levels = c("0", "l", "m", "h"),
+    labels = c("No data", "Low", "Medium", "High"))
> greys <- grey(c(250, 100, 150, 200) / 255)
> obs.list = list("sp.points", data.Set1.obs, pch = 19, col = "black",
+    cex = 2)
> spplot(data.Set1.sp, "HtClass2", col.regions = greys, # Fig. 7.3a
+    scales = list(draw = TRUE), xlab = "Easting", ylab = "Northing",
+    main = "Vegetation Height Class", sp.layout = list(obs.list),
+    xlim = c(577000,578500), ylim = c(4417500,4418800))
```

Two new concepts are introduced here. The first is the *ordered factor*, which is generated by the function `ordered()` in the fourth line. Since *low*, *medium*, and *high* together make up an ordinal scale, it makes sense to plot them in this order. An ordered factor as created by the function `ordered()` behaves the same as an ordinary factor except that its ordinal relationships are recorded and preserved in plotting. The second new concept is the use of the argument `sp.layout` to overlay a point map on top of a polygon map created with the `sp` package function `spplot()`. A list called `obs.list` is generated that contains the type of spatial object that is to be plotted, the data set, and any special arguments to control the appearance of the plot.

Examining the figures themselves, the most obvious, and distressing, feature is that the vegetation map is truncated at the north end, so that not all of the area of the northern polygons is included in the map (compare Figures 7.2a and 7.3a). Figure 7.2 displays the data contained in the file *habitatpatches.shp*. This file was created by photointerpretation of a collection of 1997 aerial photographs, and only contains land cover class (riparian, lacustrine, or freshwater wetland, see Appendix B.1). These are the habitat patches whose area and width are a part of the habitat suitability model (Table 7.1). Figure 7.3 displays data contained in the file *set1data.shp*. The polygons in this shapefile are characterized by a combination of vegetation type, height class, cover class (this is not used in the model), and age class. Vegetation type, height class, and cover class were estimated based on the 1997 aerial photographs alone, but age class was estimated based on photointerpretation of the entire set of available aerial photos together with historical vegetation maps. Because of this, the data extent of age class was slightly less in the north end. Unfortunately, this means that the floodplain age ratios in Table 7.1 cannot be measured accurately for the northernmost observation point (the one with 16 observations in Figure 7.2a), and so this data record will have to be dropped from the statistical analysis. The other truncated patch, in the northeast of Figures 7.2a and 7.3a, with zero birds observed, is less affected, and we will leave it in the analysis.

In order to continue the exploratory analysis, we must create a spatial data set that reflects the construction of the CWHR model. This involves two fundamental steps. First, we eliminate those polygons that represent habitat patches where no observation was made and append the presence-absence data to those patches where observations were made.

Second, we compute the explanatory variables in the model as represented in Table 7.1. Elimination of the habitat patches in which there is no observation point is accomplished using the sp function over().

```
> Set1.obs <- over(data.Set1.obs, data.Set1)
```

The object data.Set1.obs is a SpatialPointsDataFrame created from the file *obspts.csv* and object data.Set1.sp is a SpatialPolygonsDataFrame created by reading the shapefile *set1data.shp*. The result of the application of the function over() with first and second arguments in these classes is the data frame Set1.obs containing the attribute data of the subset of the habitat patches in data.Set1 that contain an observation point (i.e., a point in data.Set1.obs).

Now we add the presence-absence data, the ID value, and the coordinates from data.Set1.obs. For future reference, we will also add the abundance data.

```
> Set1.obs$PresAbs <- data.Set1.obs$PresAbs
> Set1.obs$obsID <- data.Set1.obs$ID
> Set1.obs$Abund <- data.Set1.obs$Abund
> Set1.obs$Easting <- data.Set1.obs@coords[,1]
> Set1.obs$Northing <- data.Set1.obs@coords[,2]
```

This completes the first of the two steps listed above, elimination of patches with no observation site. Here is the result.

```
> names(Set1.obs) # Polygons containing an obs point
 [1] "SP_ID" "ID" "AgeID" "AgeArea"
 [5] "Age" "VegType" "PatchArea" "PatchID"
 [9] "HtClID" "HtClArea" "HtClass" "CoverClass"
[13] "PatchWidth" "HtClass2" "CoverClass2" "AgeLT60"
[17] "PresAbs" "obsID" "Abund" "Easting"
[21] "Northing"
```

We can display some of the variables used in the model as follows.

```
> with(Set1.obs, cbind(PatchID, PatchArea, PatchWidth, obsID, PresAbs))
      PatchID PatchArea PatchWidth obsID PresAbs
 [1,]     191 1257345.27     593.88     1       1
 [2,]     175  301897.91     188.28     2       0
   *   *   *   DELETED   *   *   *
[21,]      19 1058784.95     414.52    21       1
```

At this point, we again encounter the dangerous step in the analysis of spatial data discussed in Section 3.6.2. We have detached the attribute data in data.Set1.obs from the spatial data and then attached them to the records in Set1.obs. The observation point data are attached to the habitat patch data according to the order in which the records appear in the respective data sets. Linking each observation point with the correct habitat patch requires that the application of the function over() preserve the order of the records in the attribute table. In Exercise 7.1, you are asked to verify that this linkage is indeed correct.

Two of the variables in the CWHR model of Table 7.1, patch area and patch width, are already present in the object Set1.obs, as are the vegetation type code, patch ID, and value of the response variable *Y* (presence or absence) for each observation point. Now we must add the other variables in Table 7.1, height class ratio and floodplain age ratio. Because it is available in the data set, we will also add the cover class ratio *dense area*:(*medium area* + *sparse area*), even though this is not in the model. Height class and cover class have the same identification value, HtClID.

We will illustrate the computation of the habitat ratios in Table 7.1 by describing the calculation of the height class ratio $HtRatio = H : (H + M + L)$, the ratio of areas of high to total vegetation. The age ratio (the fraction of the total area with vegetation less than 60 years old), and the cover class ratio (the fraction of total area with dense cover) are computed similarly. The computation of the height class ratio could be done using functional programming, but this is one of those instances where an old-fashioned for loop is probably easier and more transparent. The computations involve the object that was initially created by reading the file *set1.data.shp*, which contains all the data, rather than the just-created Set1.obs (Exercise 7.2). Recall that two objects were created, the sf, object data.Set1.sf, and the sp object data.Set1.sp. We will use the former because it has the simpler structure.

First, a data frame MLArea is created to place in the data field Area nonzero values for all those polygons in which the value of HtClass is either "m" or "l".

```
> # Create the Medium/Low data frame with ID data field
> MLArea <- data.frame(PatchID = data.Set1.sf$PatchID)
> # Initially set ML area of each patch to 0
> MLArea$Area <- 0
> # Insert the patch area if the height class is m or l
> for (i in 1:nrow(MLArea)){
+   {if ((data.Set1.sf$HtClass[i] == "m") |
         (data.Set1.sf$HtClass[i] == "l"))
+       MLArea$Area[i] <- data.Set1.sf$HtClArea[i]}}
```

The process is then repeated for the data frame HArea to hold nonzero Area values for all those polygons in which the value of HtClass is "h".

```
> HArea <- data.frame(PatchID = data.Set1.sf$PatchID)
> HArea$Area <- 0
> for (i in 1:nrow(HArea)){
+   {if ((data.Set1.sf$HtClass[i] == "h"))
+       HArea$Area[i] <- data.Set1.sf$HtClArea[i]}}
```

These data frames are then passed as denom and num respectively to the function AreaRatio() to create the height class ratio value in Table 7.1. Here is the function.

```
> AreaRatio <- function(num, denom){
+ # Add the num argument values by PatchID
+    patch.num <- aggregate(num,
+       by = list(AggID = num$PatchID), FUN = sum)
+ # Add the denom argument values by PatchID
+    patch.denom <- aggregate(denom,
+       by = list(AggID = denom$PatchID), FUN = sum)
+ # If the num and denom PatchID values are not equal, return NULL
+    ratio <- NULL
+ # Otherwise, compute the ratio (and avoid 0/0)
+    if (all.equal(patch.num$PatchID, patch.denom$PatchID)){
+       ratio <- patch.num$Area /
+          (patch.num$Area + patch.denom$Area + 0.000001)}
+    return(cbind(patch.num$AggID, ratio))
+}
```

The function AreaRatio() takes two arguments associated with a patch, num and denom, and computes the total area of the patch in the category num divided by the sum of the areas in num and denom. The R function aggregate() in the third and sixth lines does exactly what its name implies; it aggregates the contents of the data frame in its first argument by the data field in the second argument using the function defined in the third argument. The function all.equal() returns FALSE if all of the elements of the two arrays in the argument are not equal, in which case the function AreaRatio() returns NULL. Note that the argument num actually appears in both the numerator and the denominator of the ratio, since the ratio being computed is actually a faction of the total area. The value 0.000001 is added to the denominator so that a value of zero is returned in the case that the aggregated data in both num and denom are equal to zero.

The use of the function AreaRatio() is illustrated in the following code sequence, which computes the height class ratio in Table 7.1.

```
> Ratio <- AreaRatio(HArea, MLArea)
> HtRatio.df <- data.frame(PatchID = Ratio[,1], HtRatio = Ratio[,2])
> Set1.obs1 <- merge(x = Set1.obs, y = HtRatio.df,
+   by.x = "PatchID", by.y = "PatchID")
```

The object Ratio in the first line is a matrix whose first column contains the value of PatchID for each habitat patch and whose second column contains the height class ratio for that patch. This matrix is converted into a data frame. The function merge(x, y, by.x, by.y) in the third line performs what is called in GIS terminology a *join by attribute values* (Lo and Young, 2007, p. 210). That is, the function merge() returns a data frame in which all records in data frame x with a given value of by.x (which must be a data field of x) are merged (or joined) to those records in data frame y having the same value in the data field by.y. This creates the object Set1.obs1, which contains all of the data fields of Set1.obs plus the height class ratio.

The process is repeated for the floodplain age ratio (Exercise 7.3), creating Set1.obs2, and again for the cover class ratio, creating the final object Set1.obs3.

Now that we have created a data frame to hold the explanatory and response variables for each of the observation locations, we can begin to explore the data set. A quick check can be carried out by computing box plots of each of the explanatory variables. We can do this quickly using ggplot().

```
> ggplot(data = Set1.obs3) +
+    geom_boxplot(mapping = aes(as.factor(PresAbs), PatchArea))
```

The function geom_boxplot() expects to see the abscissa variable as a factor, so we coerce it. The code to produce the presentation graphics boxplot for *PatchArea* is shown in Figure 7.4a.

```
> with(Set1.obs3, boxplot(PatchArea ~ PresAbs)) # Fig. 7.4a
> title(main = "Patch Area", cex.main = 2,
+   xlab = "Presence/Absence",
+   ylab = expression(Area~"("*m^2*")"), cex.lab = 1.5)
```

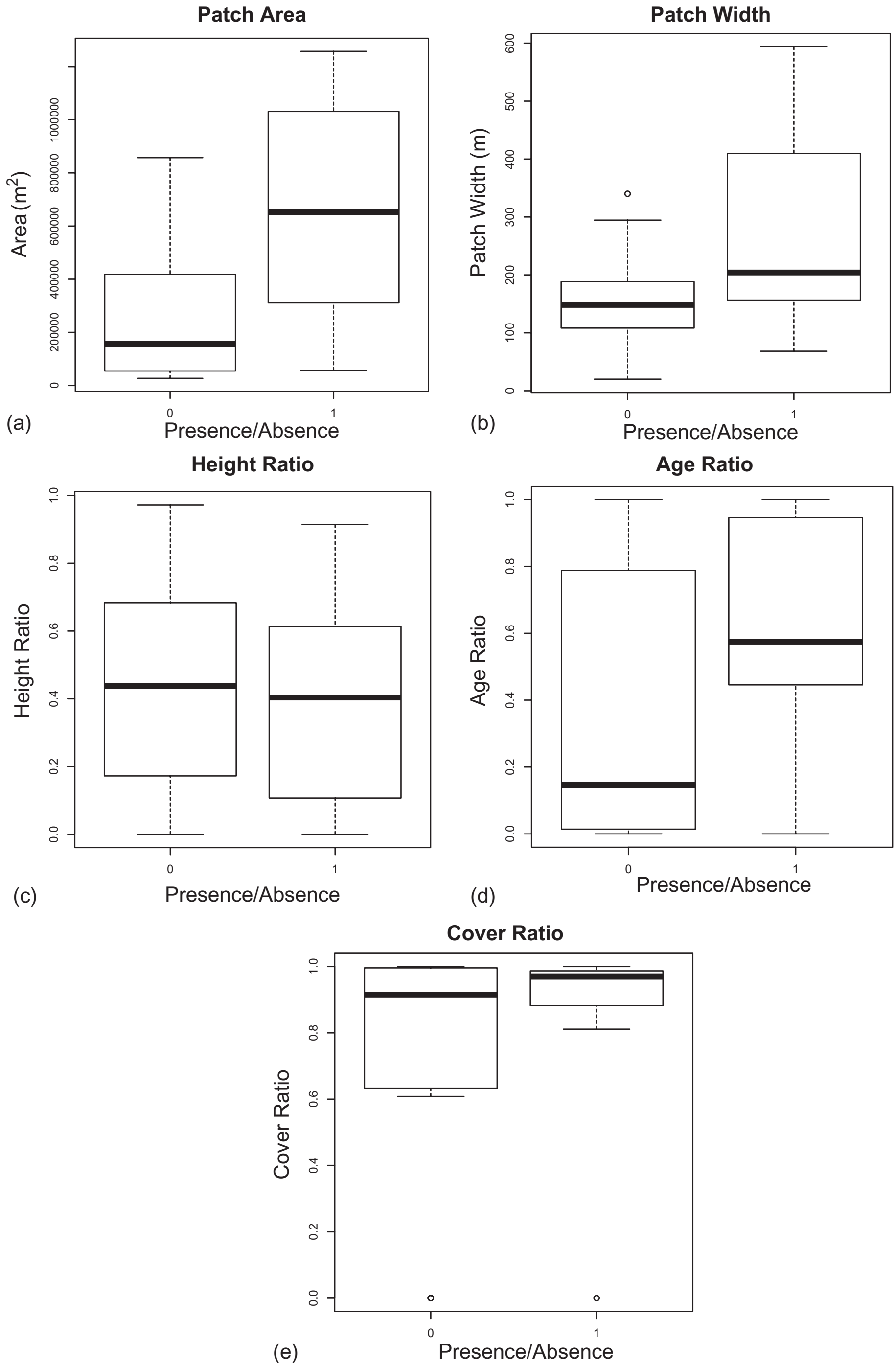

FIGURE 7.4
Boxplots showing distribution of (a) patch area, (b) patch width, (c) height ratio, (d) floodplain age ratio, and (e) cover class ratio for patches with (1) and without (0) cuckoos observed.

The other plots are created in a similar way. The heavy line inside the box indicates the median and the upper and lower extremes of the box indicate the values of the first and third quartile. The whiskers extend to the most extreme data points whose distance from the box is no more than 1.5 (this is the default value, which can be altered) times the interquartile range.

Examining these boxplots (Figure 7.4) reveals that there is considerable overlap in attribute values between occupied and unoccupied patches, and that *HtRatio*, the ratio of area of tall trees to total area, seems to play the least important role in distinguishing cuckoo presence and absence. The boxplots of *CoverRatio* are artificially compressed by one data record in which it was not measured, but this variable also does not appear to play a major role. Patch area and patch width have slightly different boxplots. The median patch area in those patches where birds are present is higher than that where birds are absent. There are also no narrow patches with birds present, but the medians of those with and without birds are fairly close. The minimum and maximum age ratios of patches with and without birds are similar, but the median age ratio of the patches with birds present is higher.

An alternative view of the data can be obtained using a pair of connected dot plots (Figure 7.5), which resemble somewhat the figure on page 49 of Tufte (1983). The graphs in Figure 7.5 are produced with the traditional graphics function `matplot()`. The function `matplot(x, y)`, where `x` and `y` are matrices, plots the columns of `y` against the columns of `x`. If there is only one argument, then this is plotted as the `y` variable. There are other options, which can be seen via `?matplot`. The code to produce Figure 7.5a is as follows.

```
> row.names(Set1.obs3) <- as.character(1:nrow(Set1.obs3))
> obs.trans <- data.frame(t(scale(with(Set1.obs3,
+    cbind(HtRatio, AgeRatio, CoverRatio, PatchArea, PatchWidth)))))
> n.trans <- nrow(obs.trans)
> obs.trans[n.trans + 1,] <- t(Set1.obs3$PresAbs)
> obs.YPres <- obs.trans[1:n.trans,(obs.trans[n.trans+1,] == 1)]
> obs.YAbs <- obs.trans[1:n.trans,(obs.trans[n.trans+1,] == 0)]
```

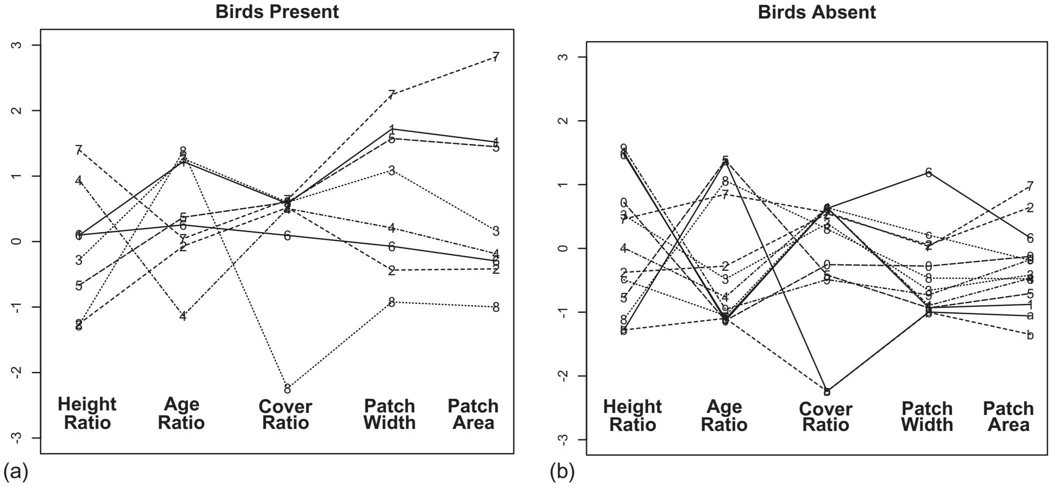

FIGURE 7.5
Connected dot plots of values of the predictor variables for observations with (a) birds present; (b) birds absent.

```
> matplot(obs.YPres, type = "o", col = "black",
+    ylim = c(-3,3), main = "Birds Present", ylab = "",
+    cex.main = 2, xaxt = "n") # Fig. 7.5a
> text(1.1,-2.5, "Height", font = 2, cex = 1.5)
> text(1.1,-2.75, "Ratio", font = 2, cex = 1.5)
   *   *   *   DELETED   *   *   *
> text(4.8,-2.5, "Patch", font = 2, cex = 1.5)
> text(4.8,-2.75, "Area", font = 2, cex = 1.5)
```

The second line applies three functions in succession. The function `scale()` centers and scales each of the explanatory data fields so that they all have zero mean and unit variance. The function `t()` creates a matrix that is the transpose of its argument, so that the data for each site is in a column. The function `data.frame()` then converts this matrix to a data frame. The next line augments the data frame by adding as the last row the presence-absence data, which is not centered and scaled. The next two lines separate the data into those where birds are present and those where birds are absent. The plot is constructed in the usual way with traditional graphics. The function `matplot()` is applied; the argument `xaxt = "n"` suppresses the *x* axis, which would consist of meaningless numbers, and instead the appropriate text is manually added.

The properties of the data in Figures 7.4 and 7.5 indicate that the explanatory data themselves may be highly correlated among themselves, and this is indeed the case.

```
> with(Set1.obs3, (cor(cbind(HtRatio, AgeRatio,
+    CoverRatio, PatchArea, PatchWidth))))
              HtRatio   AgeRatio CoverRatio PatchArea PatchWidth
HtRatio     1.0000000 -0.5313764  0.4829030 0.3103738  0.3732220
AgeRatio   -0.5313764  1.0000000 -0.2243593 0.1080622  0.1294009
CoverRatio  0.4829030 -0.2243593  1.0000000 0.5314379  0.5538397
PatchArea   0.3103738  0.1080622  0.5314379 1.0000000  0.8836189
PatchWidth  0.3732220  0.1294009  0.5538397 0.8836189  1.0000000
```

In particular (and not surprisingly), *PatchArea* and *PatchWidth* are highly correlated. However, the other important variable as indicated by Figure 7.5 is *AgeRatio*, and this is not highly correlated with either of the patch size variables. Again, there appears to be little difference in the distributions of *HtRatio* or *CoverRatio* between occupied and unoccupied patches. In comparing properties of patches with and without cuckoos present, we must bear in mind that the CWHR model is of habitat *suitability*, not of habitat *occupancy*. Thus we might expect, particularly with a very rare species, to find some suitable habitats unoccupied, but we would regard as a potential error in the model any occupied habitat that is unsuitable under the model.

The final step in the exploration process is to apply a preliminary test to the CWHR model of Table 7.1. We will construct simple 2 by 2 contingency tables (Chapter 11, see also Larsen and Marx, 1986, p. 424) of cuckoo presence/absence vs. habitat suitability/unsuitability. A function `suitability.score()` is defined to compute suitability scores for patch area and patch width in Table 7.1

```
> suitability.score <- function(x, cat.val){
+    score <- 0
+    if(x >= cat.val[3]) score <- 1
+    if(x < cat.val[3] & x >= cat.val[2]) score <- 0.66
+    if(x < cat.val[2] & x >= cat.val[1]) score <- 0.33
+    return(score)
+ }
```

This function is applied to compute the suitability scores for patch area as follows. From Table 7.1, the threshold patch areas for *Optimum*, *Suitable*, and *Marginal* are 80, 40, and 17 ha, respectively. This motivates the following code.

```
> AreaScore <- numeric(nrow(Set1.obs3))
> for (i in 1:nrow(Set1.obs3)){
+  AreaScore[i] <-
+     suitability.score(Set1.obs3$PatchArea[i] / 10000, c(17,40,80))
+ }
> Set1.obs3$AreaScore <- AreaScore
```

The same function can be used to compute the suitability scores for patch width. Similar functions are used to compute scores for height ratio and floodplain age ratio. These cannot be computed using the function `suitability.score()` because the suitability values for these variables in Table 7.1 are not monotonic (i.e., strictly increasing or strictly decreasing) (Exercise 7.4). Figure 7.6 shows a plot of the suitability score of the age ratio.

After the suitability scores are computed, the function `apply()` is next used to compute as an overall score the geometric mean of the individual scores (recall that this is used in order to emphasize the suitability of patches that score high in all explanatory variables).

```
> scores <- with(Set1.obs3, cbind(AreaScore, WidthScore,
+   AgeScore, HeightScore))
> print(Set1.obs3$HabitatScore <- apply(scores[,1:4], 1, prod)^(1/4),
```

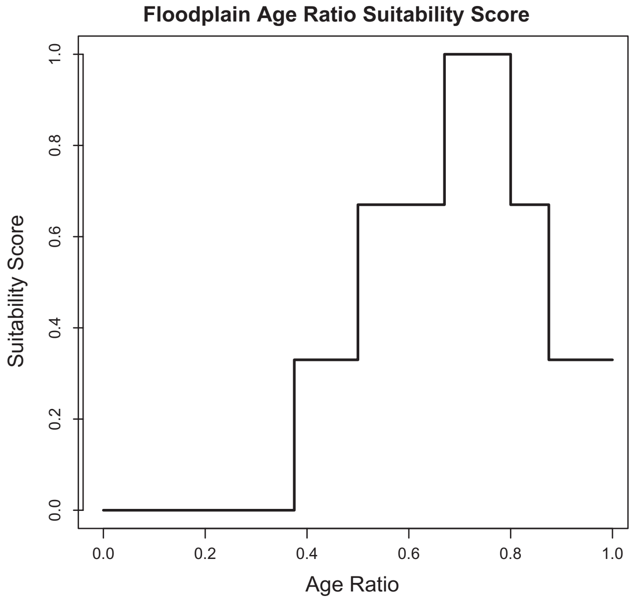

FIGURE 7.6
Plot of the habitat suitability score as a function of floodplain age ratio.

```
+   digits = 2)
 [1] 0.68 0.00 0.33 0.00 0.62 0.00 0.00 0.00 0.00 0.00 0.66 0.00
[13] 0.73 0.52 0.00 0.00 0.00 0.00 0.00 0.00 0.00
```

In this exploration, we set a very low bar: any habitat that is not unsuitable will be considered potentially suitable. Therefore, we define a habitat suitability index that takes on the value 1 if the habitat score is positive.

```
> print(Set1.obs3$HSIPred <- as.numeric(Set1.obs3$HabitatScore > 0))
 [1] 1 0 1 0 1 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0
```

The northernmost observation point in Patch 191, for which we lack floodplain age data, is excluded.

```
> Set1.corrected <- Set1.obs3[-which(Set1.obs3$PatchID == 191),]
```

Finally, we compute and display a preliminary contingency table from the following data.

```
> print(habitat.score <- as.numeric(HabitatScore > 0))
 [1] 1 0 1 0 1 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0
> print(obs <- scores$PresAbs)
 [1] 1 0 1 0 1 0 0 1 0 0 0 0 1 1 0 0 1 0 0 0
```

The contingency table indicates the occupancy of patches that are unsuitable and those that are in some measure suitable. Larger contingency tables are best done with the function `table()` (see Exercise 7.10), but we will construct this one by hand.

```
> UA <- with(Set1.corrected, which(HSIPred == 0 & PresAbs == 0))
> UP <- with(Set1.corrected, which(HSIPred == 0 & PresAbs == 1))
> SA <- with(Set1.corrected, which(HSIPred == 1 & PresAbs == 0))
> SP <- with(Set1.corrected, which(HSIPred == 1 & PresAbs == 1))
> print(cont.table <- matrix(c(length(SP),length(SA),
+    length(UP),length(UA)), nrow = 2, byrow = TRUE,
+    dimnames = list(c("Suit.", "Unsuit."),c("Pres.", "Abs."))))
        Pres. Abs.
Suit.       5    1
Unsuit.     2   12
```

There are two occupied patches (*PatchID* = 100 and 209) with a zero suitability score, and one unoccupied patch (*PatchID* = 122) with a positive suitability score. Some interpretation of this contingency table is carried out in Exercise 7.5. As a last step of preliminary analysis, you are asked in Exercise 7.6 to determine the effect on the contingency table of removing explanatory variables.

7.3 Data Set 2

Data Set 2 (Figure 1.1) 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., 1991). At each location an 81 m^2 (9 m by 9 m) area of land was surveyed. All non-climatic variables in Table 7.2 were recorded. Allen and

TABLE 7.2

Summary of the Variables Included in Data Set 2

Variable	Type	Description
Response variables:		
QUDO	N	Presence/absence of *Quercus douglasii* (Blue oak).
QUDO_BA	R	Basal area of *Quercus douglasii* (Blue oak).
Predictor variables:		
Elevation	R	Elevation (m)
CoastDist	R	Great circle distance from coast (km)
Precip	R	Mean annual precipitation (mm)
MAT	I	Mean annual temperature (C)
JaMin	I	Mean minimum temperature in January (C)
JaMax	I	Mean maximum temperature in January (C)
JaMean	I	Mean temperature in January (C)
JuMin	I	Mean minimum temperature in July (C)
JuMax	I	Mean maximum temperature in July (C)
JuMean	I	Mean temperature in July (C)
TempR	R	Annual temperature range (C) i.e., JUMEAN–JAMEAN
GS28	R	Length of 28°F growing season (days)
GS32	R	Length of 32°F growing season (days)
PE	R	Potential evapotranspiration (mm)
ET	R	Actual evapotranspiration (mm)
SolRad6	R	Potential (cloudless) solar radiation on the average day in June (MJ/m^2)
SolRad12	R	Potential (cloudless) solar radiation on the average day in December (MJ/m^2)
SolRad	R	Yearly potential (cloudless) solar radiation (MJ/m^2)
Texture	O	Soil surface texture (from 0: rock/gravel to 5: clay; 6: no data)
AWCAvg	R	Average (over components) of available water capacity for soil map unit (mm)
PM100	N	Parent materials in 100 series: igneous intrusive rocks
PM200	N	Parent materials in 200 series: igneous extrusive, flow rocks
PM300	N	Parent materials in 300 series: igneous extrusive, pyroclastic rocks
PM400	N	Parent materials in 400 series: metamorphic rocks
PM500	N	Parent material in 500 series: soft sedimentary rocks, soft
PM600	N	Parent material in 600 series: hard sedimentary rocks

co-workers entered these data into an electronic database (Allen et al., 1991). Data Set 2 is a subset of the full VTM database and includes all locations at which at least one oak species was recorded. Evett (1994) added geographic coordinates and climatic data to this database. Records in the original database include presence-absence data for six oak species, but we will concern ourselves exclusively with one of these: blue oak (*Quercus douglasii* Hook. & Arn.) This is one of the oak species considered to be in severe population decline (Zavaleta et al., 2007). Allen et al. (1991) used TWINSPAN (Two-Way Indicator Species Analysis) (Hill, 1979a) and DECORANA (Detrended Correspondence Analysis) (Hill, 1979b) to carry out a classification based on these data. Evett (1994) used direct gradient analysis (Austin et al., 1990) to construct models of the environmental niche of all six oak species.

The VTM survey achieved fairly complete coverage of the regions of California containing oak species, although portions of the southern Sierra Nevada are not included. Because the data set developed by Allen et al. (1991) and Evett (1994) includes only those sites on which at least one oak was recorded, the question we address in our analysis of this data set is not what characterizes suitability of sites in California for blue oak, but rather what characterizes suitability for blue oak of those sites in California in which some species of oak is observed.

Pavlik et al. (1991, p. 16) describe the blue oak as an extremely drought-tolerant species that populates the foothills of the Central Valley (visible in Figure 1.1 as the large valley in the center of California) at elevations less than 1,000 m. Annual precipitation in these regions ranges from 500 to 1,000 mm and soils are generally not well developed. McClaran (1986), citing Jepson (1910), describes blue oak woodland as being characterized by its exclusivity, with a complete lack of other tree species, with the occasional exception of foothill pine (*Pinus sabiniana*). In the analysis of Data Set 1, we had to take into account the distinction between habitat suitability for the yellow-billed cuckoo and presence/absence of the cuckoo, recognizing that cuckoos may be absent from suitable habitat. In the case of Data Set 2, every site contains at least one species of oak, and therefore the geographic exclusivity of blue oak woodland renders the issue of habitat suitability versus species presence less important. An important feature of Data Set 2 is that although the data are all recorded as points, they are not all measured at the same spatial scale. The climatic variables were derived by Evett (1994) using the procedure described in Appendix B.2 and are based on climatic measurements that are made at a scale that, although it is not well defined, is considerably larger than that of the non-climatic variables.

We can get a good overview of the data relationships via the spatial features `plot()` function. Appendix B.2 indicates that Data Set 2 contains several quantities that may influence blue oak growth. We select for initial exploration elevation (`Elevation`), mean annual temperature (`MAT`), evapotranspiration (`ET`), total precipitation (`Precip`), and soil texture class (`Texture`). Here is the code to set the data up and invoke the plot function.

```
> data.Set2 <- read.csv("set2\\set2data.csv", header = TRUE)
> x <- with(data.Set2, cbind(Longitude, Latitude, QUDO, Elevation, MAT,
+   ET, Precip, Texture))
> data.plot <- data.frame(x)
> data.plot.sf <- st_as_sf(data.plot, coords = c("Longitude",
+   "Latitude"))
> st_crs(data.plot.sf) <- "+proj=longlat +datum=WGS84"
> plot(data.plot.sf)
```

As always with these initial plots, the intent is to be quick and not to generate a publication quality figure, so the plot is not displayed here. Instead, Figure 7.8 below shows nice-looking figures of the same data (the code is included in the R file but not shown here). Examination of the output of the plotted spatial features object (or of Figure 7.8) indicates that, especially in the Sierra Nevada, where the maximum elevation is higher, there is a close relationship between elevation and blue oak presence. Therefore, we begin by examining in more detail the relationship between elevation and blue oak presence and absence.

Again in preliminary exploration we can use the function `ggplot()` to get a quick look at the relationship between elevation and blue oak presence, and how this interacts with other environmental variables. When we only plot *QUDO* against *Elevation* we don't see much (try it!), but when we add a call to `geom_smooth()`, which fits a smooth curve to the data, we see that blue oak presence increases with elevation until about 400 m and then declines. We will return to this in Section 9.2. We can also visualize the interaction between elevation and other variables in their influence on blue oak presence. This is most effectively done if we introduce a factor with the same values as *QUDO*. For example, the code sequence

```
> data.Set2$QF <- as.character(data.Set2$QUDO)
> ggplot(data = data.Set2) +
+   geom_point(aes(x = Elevation, y = MAT, color = QF)) +
+   geom_smooth(aes(x = Elevation, y = MAT))
```

produces a graph that summarizes how mean annual temperature and elevation interact to influence blue oak presence. The use of the geom geom_histogram(aes(data.Set2$Elevation)) or the base graphics function hist(data.Set2$Elevation) produces a histogram that indicates that the majority of data locations lie between about 250 and 1000 m, and the statement max(data.Set2$Elevation) reveals that the highest elevation measurement was taken at 2240 m. With this general knowledge, we can examine the effect of elevation on blue oak presence. The first step is to aggregate the presence and absence data at 100 m intervals. We could use the R function aggregate() to do this, but it is probably a case where a for loop is more transparent than functional programming. We first write a function that allows us to compute the number of occupied or unoccupied sites of any oak species at any range of another variable in the data set.

```
> n.oaks <- function(var.name, oak, low, high, PresAbs){
+ length(which(var.name >= low &
+ var.name < high & oak == PresAbs))}
```

Next, we aggregate the presence and absence data.

```
> pres <- numeric(22)
> absent <- numeric(22)
> for(i in 0:22) pres[i] <- with(data.Set2,
+     n.oaks(Elevation, QUDO, i*100,(i+1)*100, 1))
> for(i in 0:22) absent[i] <- with(data.Set2,
+     n.oaks(Elevation, QUDO, i*100,(i+1)*100, 0))
```

Now we are ready to plot the fraction of occupied sites versus elevation.

```
> pa <- pres / (pres + absent)
> x <- seq(100,2200,100)
> par(mai = c(1,1,1,1))
> plot(x, pa, type = "o")
> title(xlab = "Elevation (m)", ylab = "Portion with Blue Oak",
+    cex.lab = 1.5, main = "Blue Oak Presence vs. Elevation",
+    cex.main = 2)
```

If you ran the ggplot() code, you have already seen that the portion of sites having a blue oak, when aggregated over the entire data set, declines almost linearly with elevation (Figure 7.7).

As mentioned in Section 7.1, the trees are probably not directly sensitive to elevation. It is likely that elevation is serving as an exogenous variable in the sense of Figure 7.1, and that the direct influence on blue oak presence or absence is one or more endogenous variables associated with elevation. One possibility is that climate has an influence and a second, not mutually exclusive, is that soil texture is important. Figure 7.8 shows thematic maps of blue oak presence/absence, elevation, mean annual precipitation, mean annual temperature, and soil texture class (the pretty versions of the earlier plotted spatial features object). The code is similar to that used to construct Figure 1.4 and is described in Section 2.6.2. The maps indicate an apparent association between climate variables and blue oak presence.

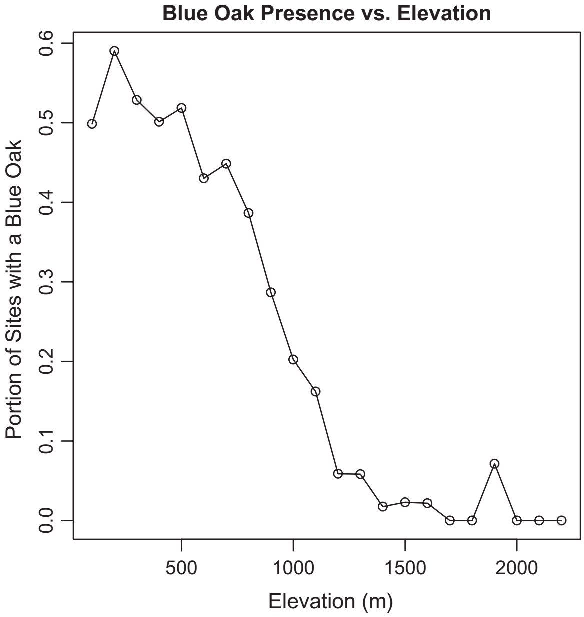

FIGURE 7.7
Plot of the fraction of sites in Data Set 2 at which a blue oak is present vs. elevation.

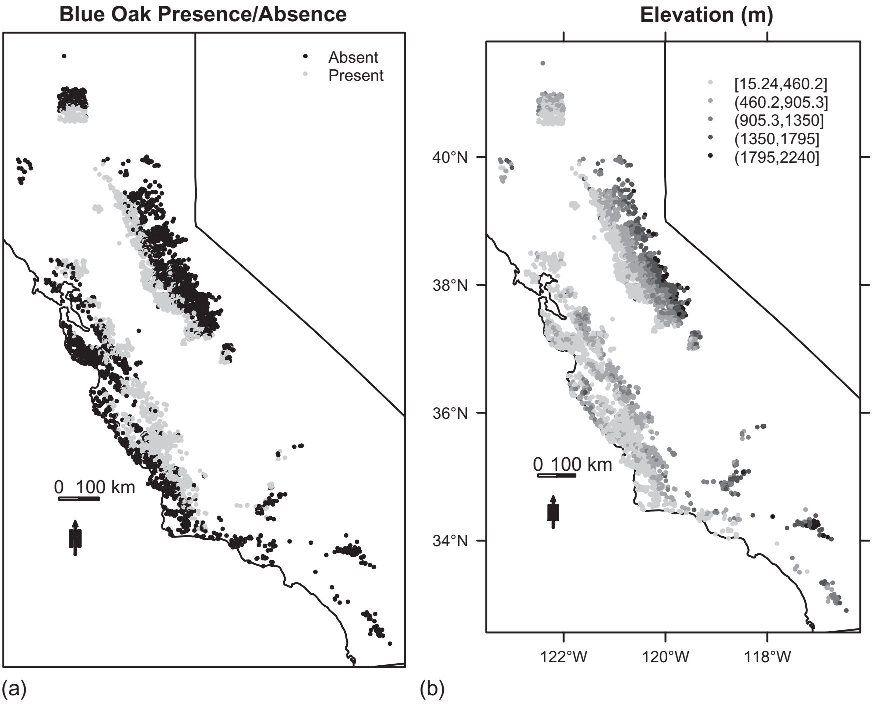

FIGURE 7.8
Thematic maps of the data set showing (a) blue oak presence/absence, (b) elevation. (*Continued*)

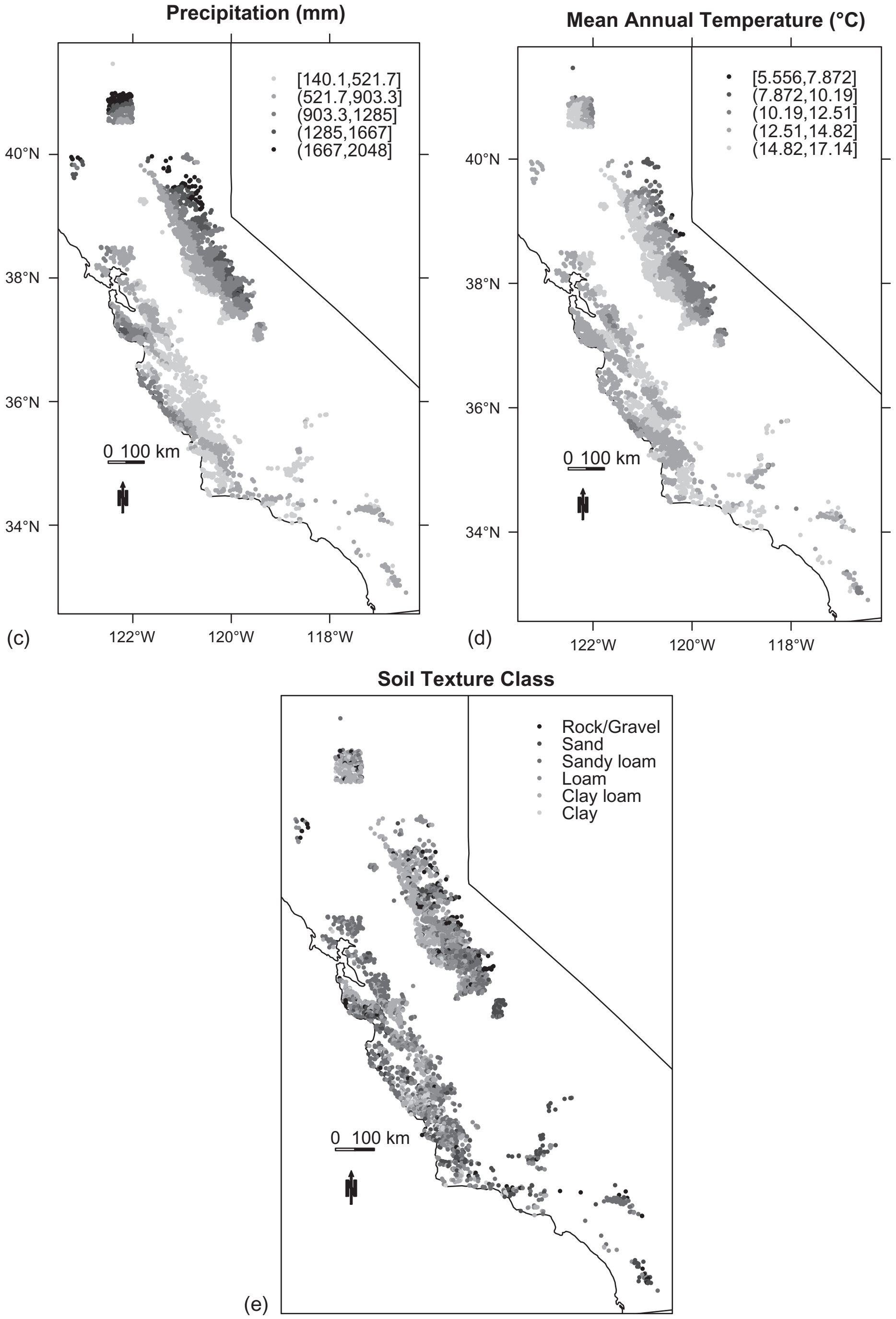

FIGURE 7.8 (Continued)
Thematic maps of the data set showing (c) mean annual precipitation, (d) mean annual temperature, (e) soil texture class at each of the sample sites.

Another issue that we can now examine is that of "secondary" spatial variables. One of the explanatory variables in Table 7.2 is *CoastDist*, the distance in kilometers from the nearest point on the coast. The maps in Figure 7.8 indicate that this might be a more appropriate east–west spatial variable than longitude, since clearly the ocean might have a strong influence. This is a situation similar to that encountered by Venables and Dichmont (2004) in their study of tiger prawn fisheries off the Australian coast. Even though it is not a climate variable per se, we will include *CoastDist* in the preliminary analysis since it may play an important role as a predictive quantity. In this context it is important to recognize that exogenous variables such as *CoastDist* and *Elevation*, while they may have great predictive value, have no explanatory value except insofar as they might lead to the identification of important endogenous variables through a process of scientific, rather than statistical, analysis.

Let's look at how these climate variables relate to each other by using a scatterplot matrix (Cleveland, 1985). The function `splom()` in the `lattice` package (Sarkar, 2008) plots one directly. We first create a data frame of the attribute data using the `data.frame()` function. Then we separate out the climate data. Not all of the climate-related data are included; we mostly focus on temperature and precipitation.

```
> climate.data <- with(data.Set2@data, data.frame(Precip, MAT, JaMin,
+    JaMax, JaMean, JuMin, JuMax, JuMean, GS32, CoastDist))
```

While we are at it, we would like to make the font of the main title of the scatterplot matrix twice the nominal size, as we have been doing with our other plots. With the `plot()` function this is done using the argument `cex.main = 2`, either in `plot()` itself or in the `title()` function. In trellis plots with `lattice` functions, we must use another approach. First, we need to find out what the term is that is used to control the plot settings. An easy way that usually works (and is often faster than wading through the documentation) is to apply the function `trellis.par.get()` to list all of the parameter settings.

```
> library(lattice)
> trellis.par.get()
$grid.pars
list()
$fontsize
$fontsize$text
[1] 12
   *   *   *   DELETED   *   *   *
$par.main.text$cex
[1] 1.2
   *   *   *   DELETED   *   *   *
```

Scrolling through a very long list of settings, we come to `par.main.text$cex` set at 1.2. This looks like it should govern the settings for the main text of the plot (and indeed that is exactly what it does), so we will set its value to 2.

```
> trellis.par.set(par.main.text = list(cex = 2))
```

This parameter setting is permanent (unless we change it) for the remainder of the R session. This is all right because we always want our main text to have this `cex` value, but trellis parameter settings can also be achieved for one time only by including them as arguments to the trellis plotting function. We can use this technique for some settings in our application of `splom()`.

```
> splom(climate.data, par.settings = list(fontsize=list(text=9),
+   plot.symbol = list(col = "black")), pscales = 0,
+   main = "Climate Data") # Fig. 7.9a
```

Arguments in `lattice` functions are often in the form of lists, for example, `plot.symbol = list(col = "black"))`. The argument `pscales = 0` suppresses printing of scales in the diagonal boxes.

The scatterplot matrix (Figure 7.9a) reveals some very tight linear relations, but also some scatterplots, such as the one relating `JaMean` to `JuMean`, that seem to be a combination

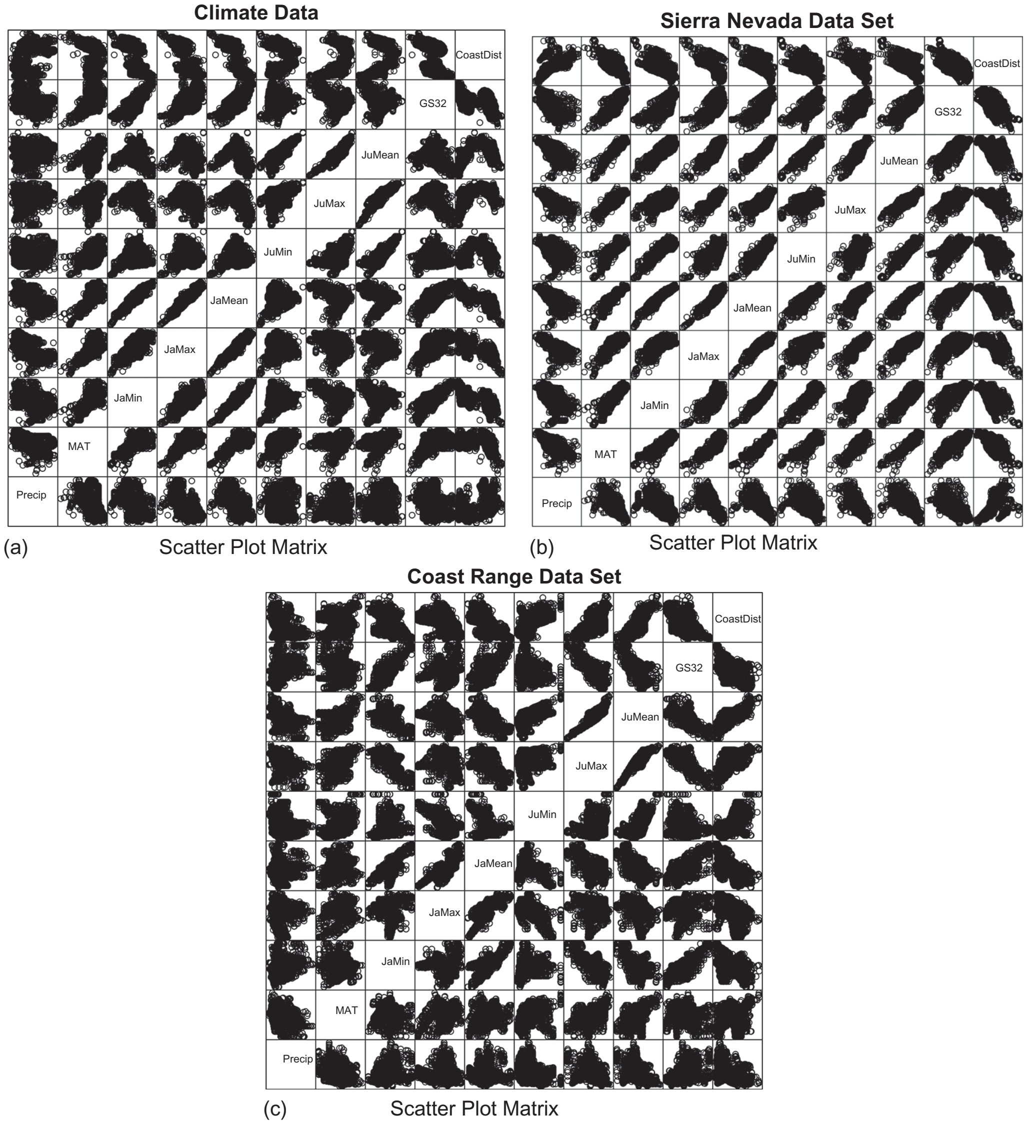

FIGURE 7.9
Scatterplot matrices of Data Set 2 climate data for (a) the full data set; (b) the Sierra Nevada subset; (c) the Coast Ranges subset.

of two or more relationships. Examination of Figure 7.8 indicates that the data sites are located in four separate groups, each of which is in a separate mountain range. The small, square northernmost group is in the Klamath Range. The two long groups running north to south are in the ranges on either side of the Central Valley. The eastern group is in the Sierra Nevada, and the western group is in the Coast Range. The small, isolated groups in the southern inlands are in the Traverse Range (http://en.wikipedia.org/wiki/Geography_of_California). The climatic conditions of these ranges are different, and it is reasonable to expect that at least some of the data relationships will be different as well. To investigate this, we will start by separating out the Sierra Nevada locations. First draw a map of the data set.

```
> plot(cal.poly, axes = TRUE) # Fig. 7.10a
> title(main = "Set 2 Sample Locations",
+    xlab = "Longitude", ylab = "Latitude",
+    cex.lab = 1.5, cex.main = 2)
> points(data.Set2, pch = 1, cex = 0.4)
```

We use the function locator() to create a boundary surrounding this group of data points. First read ?locator and note that in Windows to stop you right click (in RStudio you hit the escape key—I actually find the use of locator() easier in TINN-R). Now type the following

```
> bdry.pts <- locator(type = "o")
```

Then, by using the mouse as a digitizing tool and clicking at selected points on the screen, one can produce a list of coordinates generated by the digitizing process. Figure 7.10a shows my actual digitized points. Yours will be slightly different. The

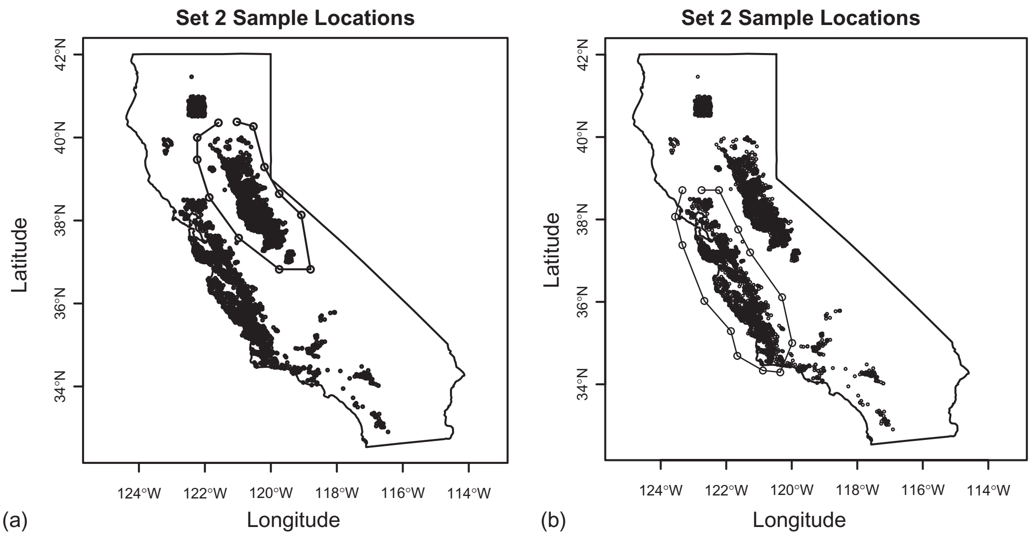

FIGURE 7.10
Boundary points drawn around the data values in (a) the Sierra Nevada and (b) the Coast Range using the function locator().

argument type = "o" causes the digitizing to be displayed as in the figure. Note that the polygon is not closed. Next, the function unlist() is used to convert the list into a vector, and then in the same line the function matrix() converts the vector into a matrix with two columns.

```
> coords.mat <- matrix(unlist(bdry.pts), ncol = 2)
```

The polygon is closed by appending the coordinates of the first point.

```
> coords.mat <- rbind(coords.mat, coords.mat[1,])
```

The sf polygon object is then created using the sequence of steps described in Section 2.4.3.

```
> coords.lst <- list(coords.mat)
> coords.pol = st_sfc(st_polygon(coords.lst))
> bdry.sf = st_sf(z = 1, coords.pol)
> proj4string(bdry.spdf) = CRS("+proj=longlat +datum=WGS84")
```

A SpatialPolygonsDataFrame called bdry.spdf representing the boundary is created by coercion from the sf object. The data points located in the Sierra Nevada are extracted from the full data using the sp function over() to do a point in polygon operation.

```
> data.ol <- over(data.Set2,bdry.spdf)
> region.spdf <- data.Set2[which(is.na(data.ol$ID) == FALSE),]
```

This process was used to create separate data frames containing data from the Sierra Nevada and the Coast Range (Figure 7.10b) and to generate the scatterplot matrices in Figure 7.9b and c. The decision of where to establish the southern boundary of the Coast Range was somewhat subjective.

It is evident that the relationship between climatic variables is very different in the two mountain ranges. It is also evident that a very close linear relationship exists between climatic variables in the Sierra Nevada, and that the relationship is more complex in the Coast Range. If the relationship among climatic variables is different between the Sierra Nevada and the Coast Range, then it may be that the relationship between elevation and presence/absence of blue oaks may also be different. Figure 7.11 confirms that this is indeed the case. The fraction of blue oak sites declines steadily with elevation in the Sierra Nevada, while it actually increases slightly in the Coast Range.

At this point, we come to a fork in the road. The scatterplot matrix of Figure 7.9a indicates that a regression model that incorporates all of the data records will be quite complex, involving higher order terms and interactions. The scatterplot matrices of Figure 7.9b and c imply that regression models restricted to these geographic regions may be much simpler. We must therefore decide whether to analyze the entire data set or split it up and analyze regions separately. There are at least two primary arguments in favor of analyzing the entire data set. First, although the regression model may be complex, it may also represent real biophysical phenomena. For example, blue oaks apparently prefer xeric environments, but it is likely that they would not be found in completely dry areas. Therefore, one might expect that in an area that ranged from very dry to very moist locations, a regression model for blue oak presence might include a quadratic term that first increased and then decreased. The region could be split into xeric and mesic sub-regions in which the portion of sites with a blue oak was respectively increasing and decreasing, but this would not capture the biophysical process as well as a model of the entire region. In any case,

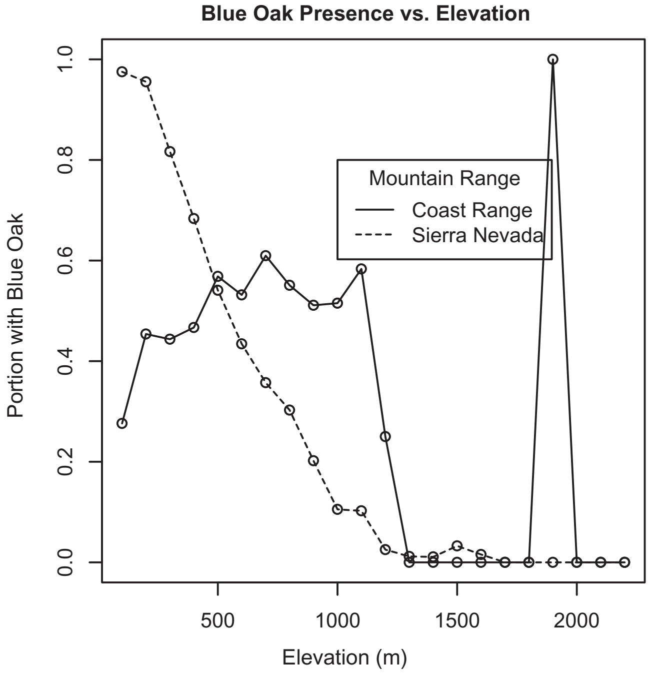

FIGURE 7.11
Plot of blue oak presence vs. elevation for the Sierra Nevada and Coast Ranges.

a spatial data set should never be split unless the two sub-regions are spatially contiguous and meaningful. A second, more practical argument in favor of analyzing the entire data set involves degrees of freedom. Splitting the data set in two means that each of the sub-regions will have only about half the number of observations, and in some cases this could have a substantial effect on the analysis by reducing the amount of available data.

Arguments in favor of a split include the following. In general, simpler models are easier to interpret and more reliable than complex ones. If the sub-regions are governed by different biophysical processes (e.g., if temperature is important in one sub-region and soil texture in the other) then two separate models may be more easily interpreted than a single model with an interaction term. Moreover, comparing and contrasting the models in the context of their geography may provide further insights into the processes they describe.

In the case of Data Set 2, with over four thousand data records, the degrees of freedom issue is not important. Examination of the scatterplots in Figure 7.9 indicates that some of the explanatory variables have a nonlinear relationship over the whole region, but the relationship is better approximated as linear over the Sierra Nevada and to a lesser extent the Coast Range by themselves. If we look at the elevation map in Figure 7.8b, we can get a sense of the geographical relationships. The elevation in the Sierra Nevada gradually increases in a more or less uniform manner from west to east. The terrain in the Coast Range is more complex. As one moves inland from the coast, elevation increases rapidly and then declines as one moves into the Salinas Valley (made famous in the works of John Steinbeck), as well as other, smaller valleys that do not show up in Figure 7.8b. Elevation rises rapidly again on the east side of the Salinas Valley, and then declines rapidly as one enters the Central Valley. The prevailing wind is from the west off the Pacific Ocean, and the terrain influences precipitation patterns. Precipitation in the Sierra Nevada generally increases with elevation. Precipitation in the Coast Ranges also tends to increase with elevation, but there is an apparent rain shadow effect on the eastern side and increased precipitation on the western side (Figure 7.8c).

It is clear that precipitation is strong associated with blue oak habitat suitability. What is not so clear right now is whether the relationship between precipitation and other climatic factors is the same in both the Sierra Nevada and the Coast Range. It does not appear to be. There may be some insight to be gained in comparing separate models for these two systems, and we can hold back the Klamath Range and the Transverse Ranges for validation purposes. Therefore, we make the decision to split the data and model the Sierra Nevada and the Coast Ranges separately. To analyze the presence-absence data, we create vectors `pa.coast` and `pa.sierra`, each housing values in groups by 100 m in elevation. Here is the code to create `pa.coast`.

```
> pres <- numeric(22)
> absent <- numeric(22)
> for(i in 0:22) pres[i] <- with(coast.sf,
+     n.oaks(Elevation, QUDO, i*100,(i+1)*100, 1))
> for(i in 0:22) absent[i] <- with(coast.sf,
+     n.oaks(Elevation, QUDO, i*100,(i+1)*100, 0))
> pa.coast <- pres / (pres + absent + 0.00001)
```

The relationship between blue oak presence/absence and climatic conditions is clearly complex, and we will delay further exploration until later chapters. The relationship among factors not included in Figure 7.9 is not very well organized. For our final exploration step of Data Set 2 in this chapter, we plot blue oak presence/absence against precipitation (Figure 7.12a) and soil texture category (Figure 7.12b). Soil texture category is an ordinal scale variable indicating the fineness of the soil particles, from 0 (rocky) to 5 (clay). Figure 7.12a is consistent with the statement of Pavlik et al. (1991, p. 16) that blue oaks are extremely drought tolerant but are out-competed by other oak species on more mesic soils. Figure 7.12b is somewhat of a surprise in the context of a statement by McDonald (1990) that blue oaks tend to be found in rocky, poorly developed soils. This apparent contradiction was also noticed by Evett (1994, p. 91).

Because soil texture is an ordinal scale variable, we plot it using a bar chart to avoid (at least for now) the idea that we might compute something like a regression line against this variable. Actually, we shall see in a later chapter that this is not always such a bad idea. In any case, the code to produce Figure 7.12b is

```
> pa <- rbind(pa.coast, pa.sierra)
> barplot(pa, beside = TRUE, # Fig. 7.12b
+   names = c("0", "1", "2", "3", "4", "5"),
+   ylim = c(0,1), xlab = "Texture Class",
+   ylab = "Percent Occupied Sites", cex.lab = 1.5,
+   legend.text = c("Coast", "Sierra"),
+   main = "Percent Blue Oak Occupation vs. Texture",
+   cex.main = 1.5)
```

Occupancy in soils in the intermediate texture class 2 is clearly different between the Coast Ranges and the Sierra Nevada, but the difference is not so obvious for the finer textured soils (classes 3, 4, and 5). Blue oak presence also declines with increasing permeability class (Figure 7.12c). However, texture class and permeability class are not closely related (Exercise 7.10).

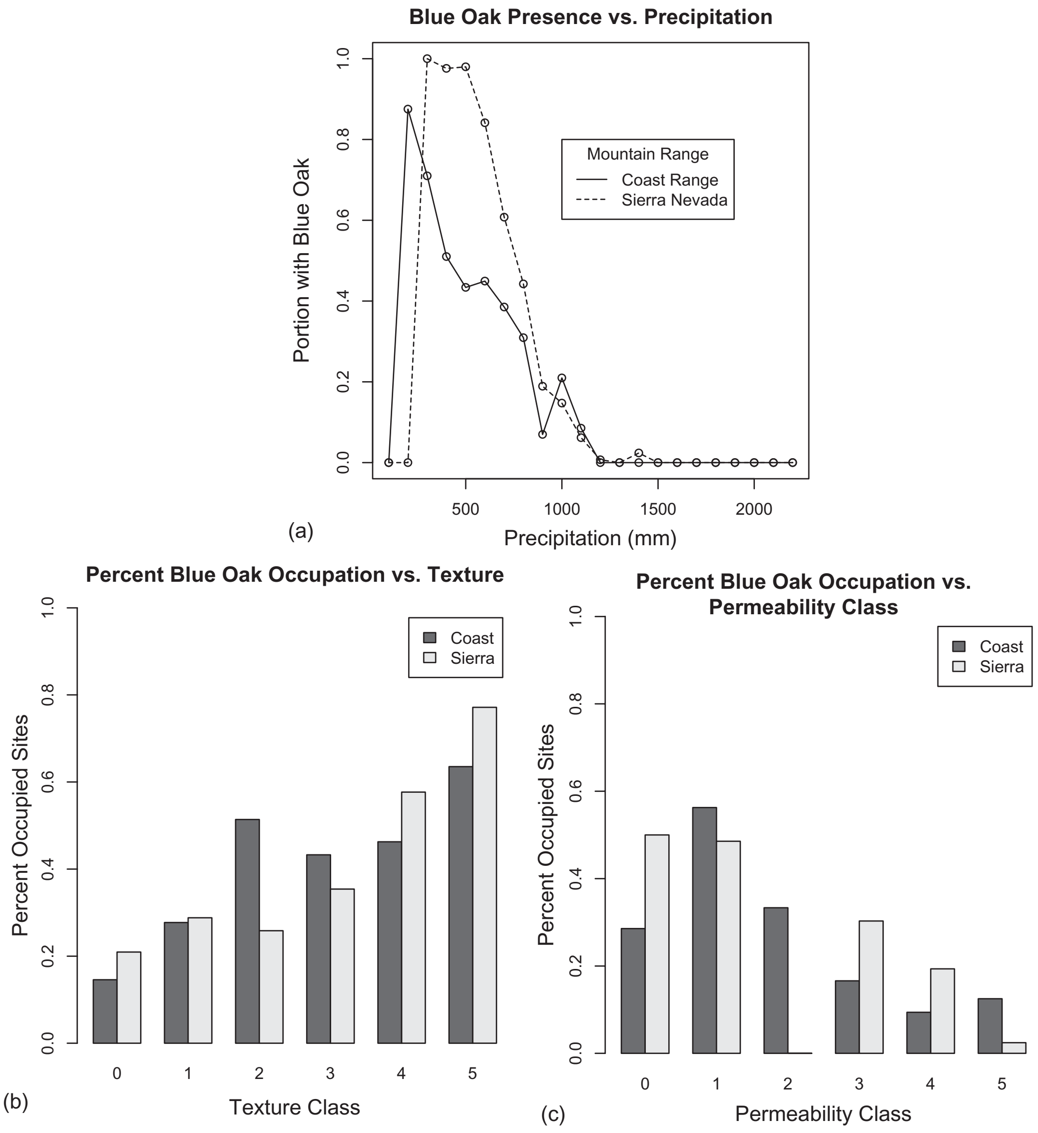

FIGURE 7.12
Plot of blue oak presence vs. (a) precipitation, (b) soil texture class, and (c) soil permeability class.

Figure 7.12a suggests that something affects the response of blue oaks to precipitation, which differs between the Sierra Nevada and the Coast Range. One possibility is temperature. We will carry out a last exploration step that allows us to introduce the function `hexbin()` from the package of the same name (Carr et al., 2016). This provides a convenient means to visualize bivariate frequency plots. Here is the code to generate a hexagon bin frequency plot of precipitation and mean annual temperature in the Coast Range.

```
> library(hexbin)
> plot(hexbin(coast.sf$Precip, coast.sf$MAT),
+    xlab = "Precipitation", # Fig. 7.13
```

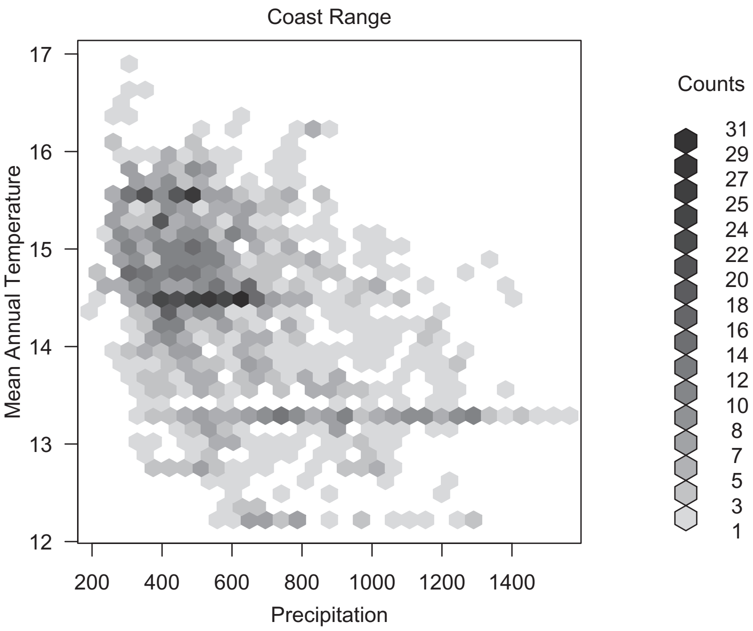

FIGURE 7.13
Hexagonal bin frequency plot of *Precip* and *MAT* in Data Set 2.

```
+    ylab = "Mean Annual Temperature",
+    main = "Coast Range")
```

Figure 7.13 shows the result. There is a bimodal distribution of *MAT* over a range of *Precip* between 200 and 700 mm. The effect is even more pronounced with the variable *JuMax* (Exercise 7.11).

In summary, we have found that blue oaks differ in their response to elevation between the Coast Range and the Sierra Nevada, with a more uniform distribution with respect to elevation in the Coast Range. Precipitation is clearly a limiting factor, with virtually no blue oak presence at sites with an annual precipitation greater than 1250 mm. The response to precipitation, however, appears to be different in the Coast Range from that in the Sierra Nevada. This may have something to do with temperature differences between the mountain ranges. Blue oaks appear to have a higher percent occupation in sites with fine textured soil, in contradiction to published anecdotal observations. It is disconcerting, however, that measured soil texture and measured permeability are not closely related.

Finally, we can use the `spatstat` package to analyze the spatial pattern of sample points to try to determine whether it is reasonably dispersed and does not favor any particular regions (see the discussion in Chapter 5 about spatial sampling patterns). Figure 7.14a shows a point pattern selected from the Sierra Nevada. The computation is made using the Ripley's *K* statistic (Ripley, 1977; Diggle, 1983, p. 47). This is defined as follows. Let λ be the mean number of points per unit area. Then

$$K(r) = \frac{E_p(r)}{\lambda}, \tag{7.1}$$

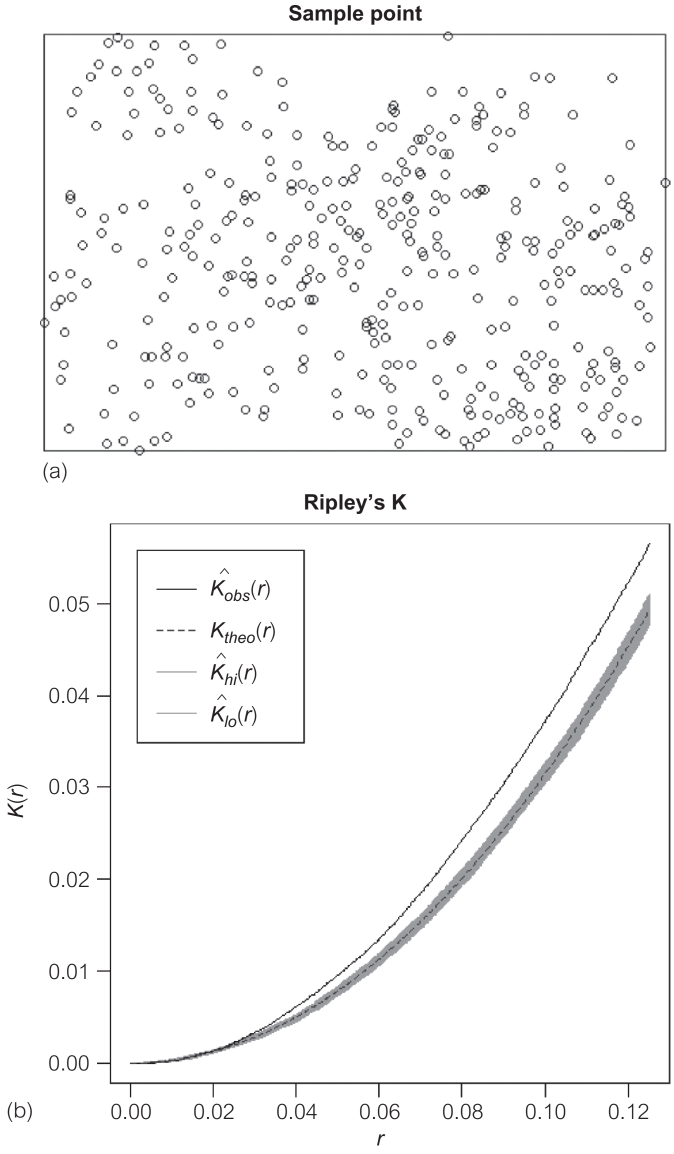

FIGURE 7.14
(a) Point pattern of a small section of sample points in the Sierra Nevada and (b) Plot of the theoretical and observed values of Ripley's K for the point pattern of Figure 7.14a.

where $E_p(r)$ is the expected number of additional points within a distance r of an arbitrarily chosen point. We can use the spatstat functions Kest() and envelope() to carry out this computation. Here is the code.

```
> W <- -120.8994
> S <- 38.0425
> E <- -120.1537
> samp.pts <- which(coordinates(data.Set2)[,1] <= E & coordinates(data.
Set2)[,1] >= W &
+    coordinates(data.Set2)[,2] >= S & coordinates(data.Set2)[,2] <= N)
```

```
> longitude <- coordinates(data.Set2)[samp.pts,1]
> latitude <- coordinates(data.Set2)[samp.pts,2]
> samp.ppp <- ppp(longitude, latitude, window = owin(c(W, E),c(S, N)))
> plot.ppp(samp.ppp, main = "Sample points") # Fig. 17.14a
> plot(envelope(samp.ppp, Kest), main = "Ripley's K") # Fig. 17.14b
```

Figure 7.14a shows the location of the points, which are in the central Sierra Nevada. Figure 7.14b shows a plot of the observed values of $K(r)$ as a function of r together with an envelope containing the highest and lowest simulated values obtained using a Monte Carlo simulation consisting of 99 runs of simulated data in which the same number of points are arranged at random. These are not confidence intervals in the normal sense, but they do give an indication of how different the observed value of $K(r)$ is from the theoretical value. Values of r for which the observed value of $K(r)$ is higher than the theoretical value are taken as an indication that the data are more clustered in space than randomly arranged data would be, and values of the observed $K(r)$ below the theoretical value are taken as indicating more dispersed point locations. Figure 7.14b indicates that the data are clustered at all values of r. This is interesting as it would be hard (at least for me) to draw this conclusion by eye. In Exercise 7.12, you are asked to gain further insight into Ripley's K by plotting it for the sampling patterns of Chapter 5.

7.4 Data Set 3

Data Set 3 (Appendix B.3) was collected over a period of three cropping seasons from a total of 16 rice fields in the region around Treinta y Tres, Uruguay (which is named for the 33 individuals who initiated the revolution that ultimately resulted in Uruguayan independence). The fields are all relatively large (generally about 30–50 ha). At the time of data collection, rice was typically grown in Uruguay in a five-year rotation, with two years of rice followed by three years of pasture. Many farmers do not own the land but rather rent it on a year to year basis, so that some fields were farmed by different people in different years. The fields are located in three geographic regions, northern, central, and southern (Figure 1.8). The northern region is located about 50 km to the north of the central region, and the southern region is about 75 km to its south. The data fields are given in Table 7.3. The objective of the study was to determine the practices that, by altering them, would have greatest impact on improving farmers' yields. The approach to realizing this objective was to determine the most effective practices that distinguished the farmers getting the highest yields from those getting lesser yields, taking into account field conditions.

A quick plot of UTM coordinates using `ggplot()` indicates that all fields in the northern region have a Northing greater than 6,340,000 m and all fields in the southern region have a Northing less than 6,280,000 m.

```
> data.Set3$SeasonFac <- factor(data.Set3$Season,
+    labels = c("2002-03", "2003-04", "2004-05"))
> ggplot(data = data.Set3) +
+   geom_point(mapping = aes(x = ID, y = Northing, color = SeasonFac))
```

The study spanned four calendar years (three Southern Hemisphere summers), and the variable *Season* represents a growing season (i.e., 1, 2, or 3). In the code we introduce the factor `SeasonFac`, which will serve as one of the grouping factors moving forward. Based

TABLE 7.3

Field Variables Included in Data Set 3

Variable Name	Units	Quantity Represented	Type	Scale
pH		pH	Exogenous	Ratio
Corg	%	Soil organic carbon	Exogenous	Ratio
SoilP	%	Soil P conc.	Exogenous	Ratio
SoilK	%	Soil K conc.	Exogenous	Ratio
Sand	%	Sand content	Exogenous	Ratio
Silt	%	Silt content	Exogenous	Ratio
Clay	%	Clay content	Exogenous	Ratio
Weeds		Weed level prior to herbicide	Exogenous	Ordinal
Irrig		Irrigation effectiveness	Exogenous	Ordinal
DPL	15 days	Planting date after 1 October	Management	Ratio
Cont		Level of weed control	Management	Ordinal
Farmer		Farmer ID letter	Management	Nominal
Fert	kg ha^{-1}	Total fertilizer	Management	Ratio
N	kg ha^{-1}	Fertilizer N	Management	Ratio
P	kg ha^{-1}	Fertilizer P	Management	Ratio
K	kg ha^{-1}	Fertilizer K	Management	Ratio
Var		Variety	Management	Nominal
Yield	kg ha^{-1}	Yield	Response	Ratio
Emer	Days	Emergence	Response	Ratio
D50	Days	Days after 1 Jan. to 50% flower	Response	Ordinal

on this, a `Location` data field was added to the data set that had the three values `North`, `Center`, and `South`. Let's look at examples of boxplots created using traditional and trellis graphics (the code also contains the plots created with `ggplot()`). The data frame `data.Set3` holds the contents of the file *set3.data.csv* (Appendix B.3). Figure 7.15 shows a boxplot of yield by farmer created using traditional graphics.

```
> boxplot(Yield ~ Farmer, data = data.Set3,
+   main = "Rice Yields by Farmer", xlab = "Farmer", cex.main = 2,
+   ylab = "Yield (kg/ha)", cex.lab = 1.5) # Fig. 7.15
>
```

Trellis graphics were used to create a set of box-and-whiskers plots of the yields for each farmer for the three regions (Figure 7.16), with the single grouping variable *Location*. Here is the code.

```
> trellis.par.set(par.main.text = list(cex = 2))
> trellis.device(color = FALSE)
+ labels = c("2002-03", "2003-04", "2004-05"))
> bwplot(Yield ~ Location | SeasonFac, data = data.Set3,
+   main = "Rice Yields by Year", # Fig. 7.16
+   xlab = "Location", ylab = "Yield (kg/ha)", layout = c(3,1),
+   aspect = 1)
```

Median yield in the central region is very consistent from one season to the next, although variability is somewhat higher in the first season (Figure 7.16). Yield in the south is reduced

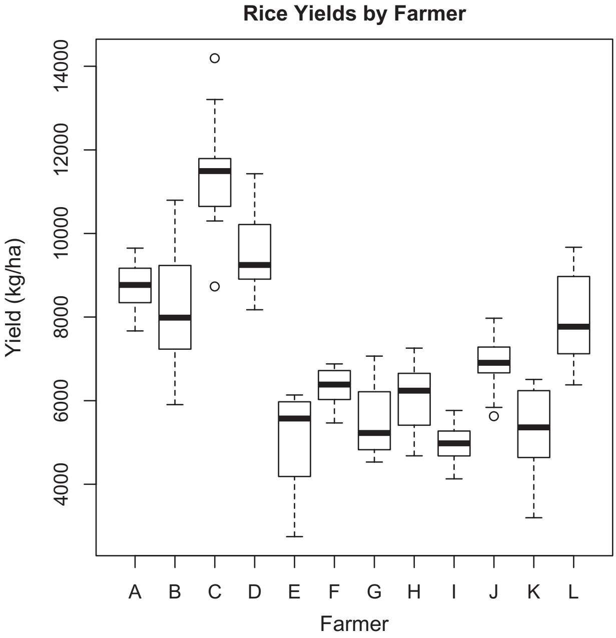

FIGURE 7.15
Box-and-whiskers plots of individual farmer yields in Data set 3, constructed using traditional graphics.

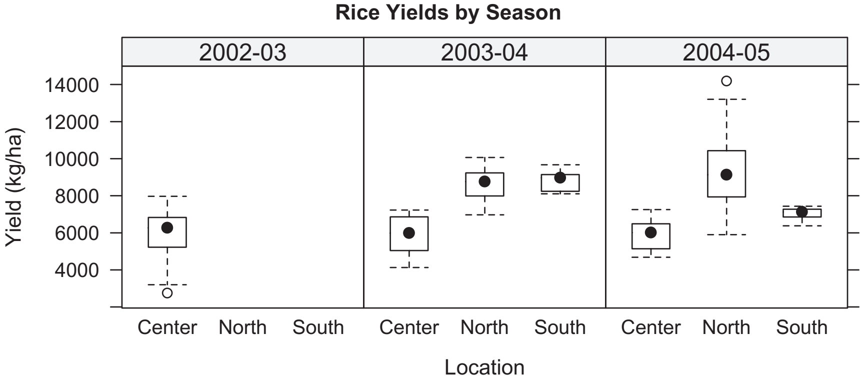

FIGURE 7.16
Trellis box-and-whiskers plots of yield by region with season number as a grouping variable.

a bit, and yield variability in the north is increased considerably in the third year. Farmers *A* through *D* are in the northern region, Farmers *E* through *K* are in the central region, and Farmer *L* is in the south. The farmers in the north obtained the highest yields in general. The farmer in the south had median yields somewhat below those of the northern farmers. The farmers in the central region tended to get much lower yields. Although he has one bad harvest, farmer *C* in the north clearly leads the pack. Farmer *L* in the south is about on par with the other three northern farmers, *A*, *B*, and *D*. The farmers in the center fall in behind with differing amounts of variability; farmer *I* gets consistently mediocre yields and farmers *E* and *K* have some good and some truly dreadful seasons. Farmer *J* has the highest average yield in the center.

Let us first examine in more detail the mean yield of each of the three regions in each of the three years. We can use the function `tapply()` to compute the mean yields over region and season by numbering the regions as 1 for the northern, 2 for the central, and 3 for the southern, and assigning a code to each data record identifying the region–season combination. The code for location–season combination consists of the location code plus 10 times the season code.

```
> data.Set3$LocN <- 2
> data.Set3$LocN[(data.Set3$Northing > 6340000)] <- 1
> data.Set3$LocN[(data.Set3$Northing < 6280000)] <- 3
> data.Set3$LocSeason <- with(data.Set3, Season + 10 * LocN)
> print(round(mean.yields <- tapply(data.Set3$Yield,
+    data.Set3$LocSeason, mean)))
  12   13   21   22   23   32   33
8605 9173 5999 5900 5909 8832 7064
```

Mean yield stayed fairly consistent in the north and center and dropped substantially in the south. There was, however, only one field in each season in the south, and the fields were different. Since the differences in yield seem to be due to field differences more than season differences, we will not normalize yield over seasons.

We can also determine whether there is evidence of a second year effect, that is, whether yield tends to decline in the second year in a field in which rice is grown in two years in succession. The data field `RiceYear` indicates the year of the field in the rotation.

```
> unique(data.Set3$RiceYear)
[1] 1 2
```

Let's see which fields were in the study in both first and second years.

```
> sort(unique(data.Set3$Field[which(data.Set3$RiceYear == 1)]))
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
> sort(unique(data.Set3$Field[which(data.Set3$RiceYear == 2)]))
[1]  3 5 12 13 14 16
```

Fields 3, 5, 12, 13, and 14 had both first and second rice years in the study. We can do a *t* test on the null hypothesis that the yields were the same in each rice year.

```
> t.test(data.Set3$Yield[which((data.Set3$Field == 3)
+      & (data.Set3$RiceYear == 1))],
+      data.Set3$Yield[which((data.Set3$Field == 3)
+      & (data.Set3$RiceYear == 2))])
        Welch Two Sample t-test
t = 1.4235, df = 43.61, p-value = 0.1617
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -211.4933 1227.8933
sample estimates:
mean of x mean of y
  8472.84   7964.64
```

In Field 3 the second rice year yield declined, but not significantly, in Fields 5 and 12 the yield declined significantly in the second year. In Fields 13 and 14, the yields increased significantly. Here is a list of every farmer and field in each year.

```
> data.Set3$YearFarmerField <- with(data.Set3,
+    paste(as.character(Season),Farmer, as.character(Field)))
> print(tapply(data.Set3$Yield, data.Set3$YearFarmerField, mean),
+    digits = 4)
1 E 11  1 F 7  1 G 8 1 H 10 1 J 12 1 J 13 1 J 14  1 J 5  1 K 6  1 K 9
  4771   6331   5559   5995   6898   6408   7153   7404   5904   4091
 2 A 1  2 B 3  2 I 5 2 J 14 2 L 15  3 B 3  3 C 2  3 D 4 3 E 13 3 H 12
  8738   8473   4982   6742   8832   7965  11456   9542   5698   6023
3 L 16
  7064
```

Yields appeared to be more determined by the farmer than by location or season in those cases where different farmers farmed the same field in different seasons. This is most evident in the central region, in which *J* farmed Fields 5, 12, 13, and 14 in season 2, while *H* farmed Field 12 in season 4, *J* farmed Field 14 in season 3, and *I* farmed Field 5 in season 3. Of these fields, only that farmed by *J* maintained its yield in both seasons. The only field in the north farmed in two successive seasons was Field 3, which was farmed by *B* and maintained a roughly similar yield in both seasons. In summary, certain farmers appear to be much more successful in obtaining high yields than others. Can we find out the secrets to *J*'s and *C*'s success? Is it better management or better field conditions? If it is better management, what are the most influential management factors?

Our initial question is whether something in the physical environment could be affecting the regions. One possibility is terrain. Rice is grown in Uruguay as a flood-irrigated crop (Figure 1.6), and a high level of relief can affect water distribution. The SRTM (Shuttle Radar Topography Mission) digital elevation model (DEM), which has a 90 m cell size, was downloaded from the website provided by the Consultative Group on International Agricultural Research (CGIAR), http://srtm.csi.cgiar.org/. It is the tile between 50°W and 55°W and 30° and 35°S. The files in the *auxiliary* folder are included with the data. We can analyze the grid using functions from the `raster` package. We use the function `raster()` to read the file, then we assign a projection.

```
> library(raster)
> dem.ras <- raster("auxilliary\\dem.asc")
> projection(dem.ras) <- CRS("+proj=longlat +datum=WGS84")
```

The DEM is quite large, and we only need a small portion of it, so we determine the ranged of longitude and latitude values of the data and then use the function `crop()` to create a smaller grid.

```
> range(data.Set3$Latitude)
[1] -33.75644 -32.77624
> range(data.Set3$Longitude)
[1] -54.52640 -53.74849
> crop.extent <- matrix(c(-54.6,-53.7,-33.8,-32.7),
+     nrow = 2, byrow = TRUE)
> dem.Set3 <- crop(dem.ras, extent(crop.extent))
```

Next, we use the function terrain() to compute the slope.

```
> slope.Set3 <- terrain(dem.Set3, opt = "slope")
```

To find the slopes in the fields, we copy data.Set3 to a file that we convert to a SpatialPointsDataFrame and use the function extract() to obtain the slope values at the field locations.

```
> Set3.WGS <- data.Set3
> coordinates(Set3.WGS) <- c("Longitude", "Latitude")
> slopes <- extract(slope.Set3, Set3.WGS)
> print(range(slopes), digits = 3)
[1] 4.91e-18 8.93e-02
```

There is a large range in slopes, with a maximum of almost 9%. An application of the function hist() (not shown) shows that almost all of the slope values are less than 3%. Let's see which fields are in highly sloped regions.

```
> sort(unique(data.Set3$Field[which(slopes > 0.03)]))
[1]  3 13 14
```

There is, however, only a very small correlation between slope and yield.

```
> print(cor(data.Set3$Yield, slopes), digits = 2)
[1] -0.034
```

Moreover, the fields with higher slopes do not tend to be those displaying low yields. Therefore, we do not consider terrain an important factor.

A second question is the extent to which there was a season effect. Addressing the question of a season effect allows us to use a trellis plot with two grouping variables, farmer and season. Although data were collected during three seasons, the first season's data were only collected from the central region. The box-and-whiskers plot of Figure 7.17 does not make very effective use of space, but it is informative.

```
> bwplot(Yield ~ Farmer | Location + SeasonFac, data = data.Set3,
+   main = "Rice Yields by Farmer and Season",
+   xlab = "Farmer", ylab = "Yield (kg/ha)") # Fig. 7.17
```

Once again, we see the farmer effect. Farmers *C* and *D*, who only entered the study in the last year, accounted for the greatly increased yields in that year. The only northern farmer observed in both year 2 and year 3 was *B*, whose yields were pretty consistent between years.

We can begin to examine the possibility that management actions are associated with differences in yield between the northern, central, and southern regions. We can use the function tapply() to get a quick look at the mean amount of fertilizer applied by region.

```
> tapply(data.Set3$Fert, data.Set3$Location, mean)
  Center North South
109.4405 158.0000 125.3200
```

The farmers in the north apply the most fertilizer, the farmers in the center apply the least, and the farmer in the south is in between. Of course, it may be that the farmers in the north apply too much fertilizer, or that the soil in the center is not suitable for higher levels of fertilizer application. This is what we are trying to determine. It is also possible that some of the farmers cannot afford more fertilizer, but we have established at the outset that we are going to ignore economic considerations.

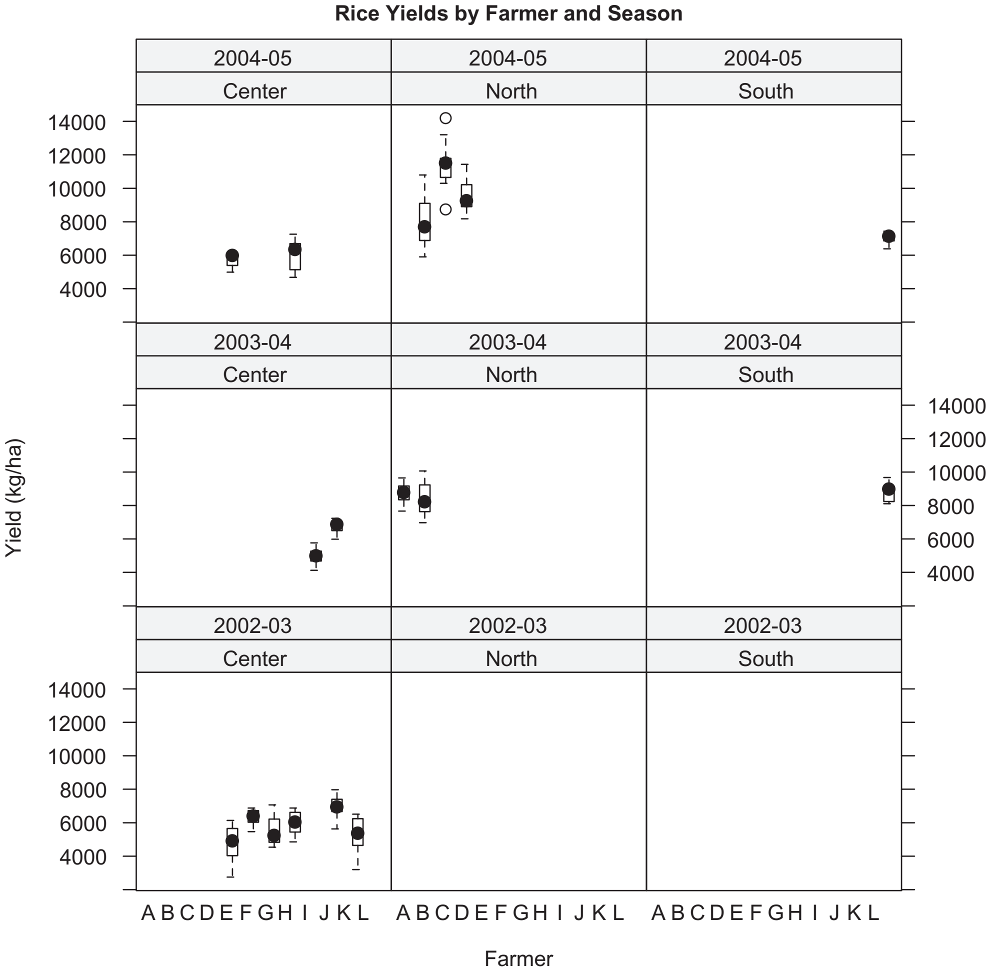

FIGURE 7.17
Trellis box-and-whiskers plots of yield grouped by farmer and season.

Using `tapply()` with other variables generates Table 7.4. In each of the management actions the farmers in the northern region managed at a higher level, the farmers in the central region managed at a lower level, and the farmer in the southern region was intermediate between the two. For example, although farmers in the north generally had lower weed problems prior to applying control measures, as indicated by a smaller average value of *Weeds*, they controlled weeds more effectively, indicated by a larger value of *Cont*. Similarly, they irrigated more effectively. Finally, although they had higher mean soil test

TABLE 7.4

Mean Values of Measured Quantities in the Three Geographic Regions of Data Set 3.

Region	*Weeds*	*Cont*	*Irrig*	*SoilP*	*SoilK*	*Fert*	*N*	*P*	*K*
North	2.3	4.8	4.2	6.2	0.20	158	62	72	23
Center	3.5	3.5	3.4	5.6	0.22	109	54	57	0
South	3.0	4.1	3.8	3.6	0.30	125	61	64	0

P and K values on average, the northern farmers still applied more fertilizer P and K, in addition to more fertilizer N.

The final stage of preliminary exploration in this section is to begin to search for insight as to which of the measured quantities has the greatest impact on yield. First, we consider climatic factors. Temperature, hours of sun per day, and rainfall were recorded three times a month at the Paso de la Laguna agricultural experiment station located near Treinta y Tres. The data frame wthr.data holds the contents of the file *set3weather.csv* (Appendix B.3). The code to plot the temperature in each year as well as the normal for that day is as follows.

```
> plot(wthr.data$Temp0203, type = "l", lty = 2, xaxt = "n",
+    ylim = c(10,27), main = "Regional Temperature Trends",
+    ylab = expression(Temperature~"("*degree*C*")"),
+    xlab = "Month", cex.main = 2, cex.lab = 1.5) # Fig. 7.18a
> lines(wthr.data$Temp0304, lty = 3)
> lines(wthr.data$Temp0405, lty = 4)
> lines(wthr.data$TempAvg, lty = 1)
> axis(side = 1, at = c(1,4,7,10,13,16,19),
+    labels = c("Oct","Nov","Dec","Jan","Feb","Mar","Apr"))
> legend(8, 15, c("2002-03", "2003-04", "2004-05",
+   "Normal"), lty = c(2,3,4,1))
```

There are a few points to be made about this plot. The first is that the default abscissa would indicate the index of each record, which is much less useful than indicating the month. For this reason, the default abscissa is suppressed through the argument xaxt = "n" of the function plot(), and the labels are added manually using the function axis(). In line 3, the function expression() is used to insert a degree symbol in the ordinate title.

The resulting plot is shown in Figure 7.18a. It is fairly clear that in each season the temperatures late in the season tended to be higher than normal, while these temperatures tended to be lower than normal earlier in the season. It is not easy, however, to visualize how much higher. Cleveland (1985, p. 276) points out that the vertical distance between two curves of varying slopes is very difficult to evaluate, and that a plot of the difference between these curves is more effective. Figure 7.18b plots the difference between each season's temperature measurements and the normal temperature. The temperature did tend to move from cooler than normal early seasons to warmer than normal late seasons, with a particularly warm late southern hemisphere fall in 2004. Differences from the normal in hours of solar irradiation did not follow any particular pattern (Figure 7.18c). None of the years were particularly dry relative to the normal (Figure 7.18d), and indeed there was heavy rainfall in the late fall in 2003.

We now consider the effect of soil-related variables. Here are the texture components of each region.

```
> Sand <- with(data.Set3, tapply(Sand, Location, mean))
> Silt <- with(data.Set3, tapply(Silt, Location, mean))
> Clay <- with(data.Set3, tapply(Clay, Location, mean))
> print(cbind(Sand, Silt, Clay), digits = 3)
  Sand Silt Clay
1 46.6 32.0 21.5
2 15.5 61.5 23.0
3 30.3 37.9 31.8
```

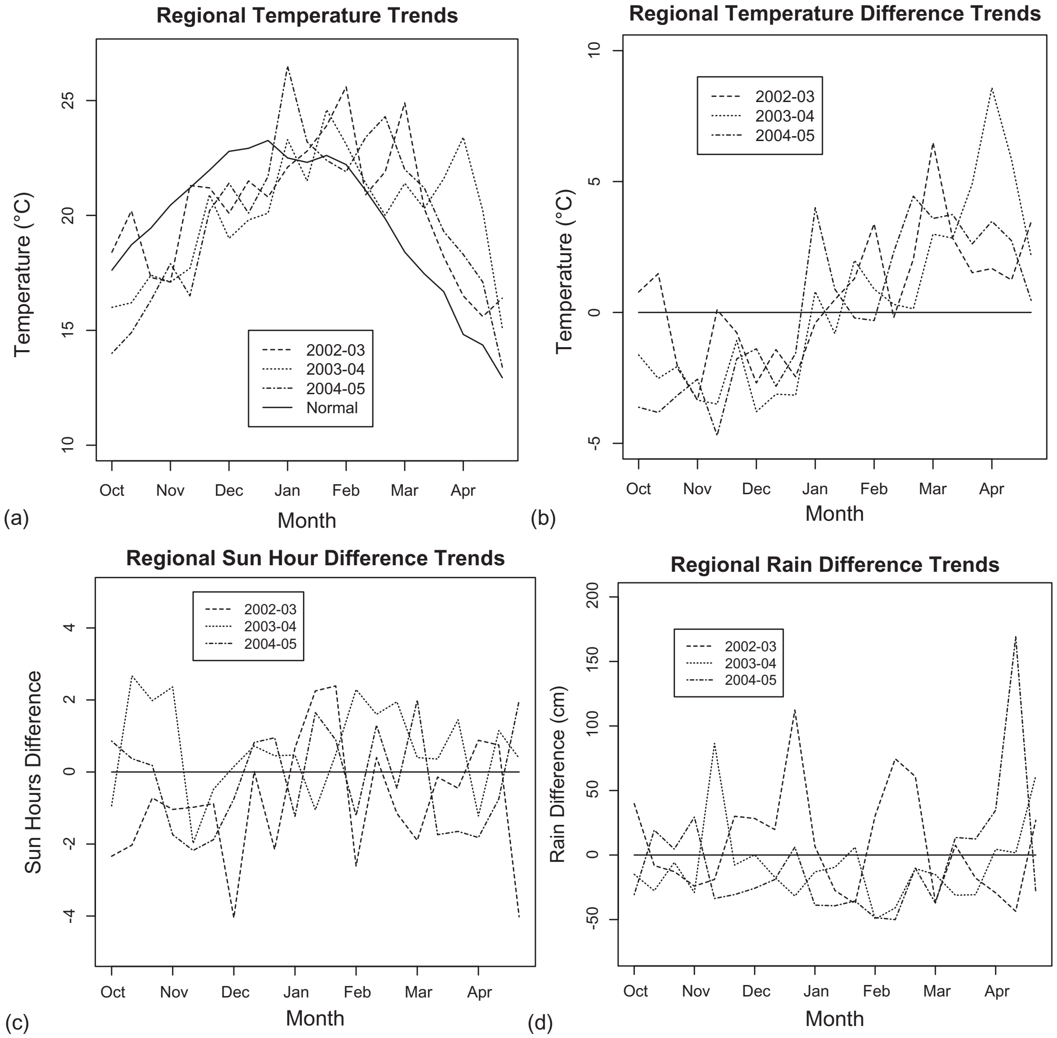

FIGURE 7.18
(a) Plot of temperatures recorded three times a month at the Paso de la Laguna experiment station, Treinta y Tres, Uruguay; (b) plot of the difference between the normal temperature and that recorded in each growing season; (c) plot of the difference between the normal number of hours of sun per day for each season; (d) plot of the difference between the normal rainfall per day for each season.

The northern region has a higher sand content on average, and the central region has a much higher silt content. A scatterplot matrix of Data Set 3 yield together with soil-related factors indicates that yield may have a positive association with sand and a negative association with silt, and possibly with pH (Figure 7.19). The scatterplot matrix provides information on relationships at the landscape scale; it is of interest to determine whether the relationship between yield and silt content “scales down” to the field level. We can make this determination using the function `xyplot()` from the `lattice` package.

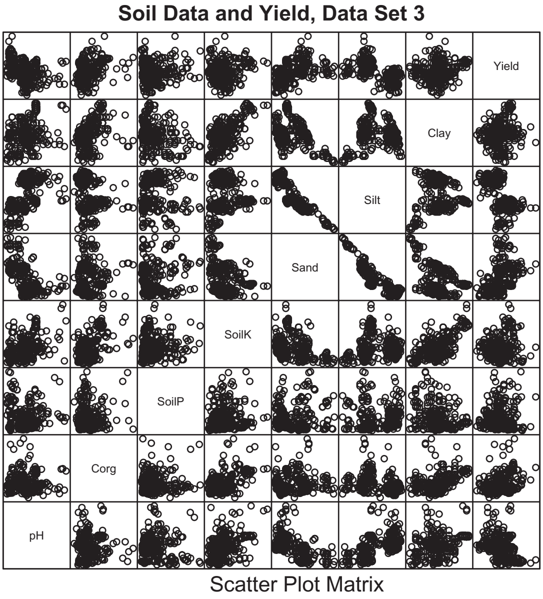

FIGURE 7.19
Scatterplot matrix of soil-related data of Data Set 3.

```
> trellis.par.set(par.main.text = list(cex = 2))
> trellis.device(color = FALSE)
> data.Set3$FieldFac <- factor(data.Set3$Field,
+    labels = c("Field 1","Field 2","Field 3",
+    "Field 4","Field 5","Field 6","Field 7",
+    "Field 8","Field 9","Field 10","Field 11",
+    "Field 12","Field 13","Field 14","Field 15",
+    "Field 16"))
> xyplot(Yield ~ Sand | FieldFac, data = data.Set3,
+    main = "Yield vs. Sand by Field") # Fig. 7.20
```

The trellis plot (Figure 7.20) indicates that there is no consistent relationship at the field level between yield and silt content. A couple of fields show a negative relationship, a couple of fields show a positive relationship, and most of the fields show no relationship. A similar lack of consistency is present in the field by field relationships with sand and pH (not shown). This is an example of a phenomenon called the *ecological fallacy*, which will be discussed in Section 11.5.2.

Finally, we examine the effect of variety. Four varieties were planted. They are coded numerically in the data set; the corresponding variety names are given by Roel et al. (2007). We can use these to compute mean yields by variety.

```
> data.Set3$Variety <- "INIA Tacuarí"
> data.Set3$Variety[which(data.Set3$Var == 2)] <- "El Pasol"
> data.Set3$Variety[which(data.Set3$Var == 3)] <- "Perla"
> data.Set3$Variety[which(data.Set3$Var == 4)] <- "INIA Olimar"
> with(data.Set3, tapply(Yield, Variety, mean))
    El Pasol  INIA Olimar INIA Tacuarí        Perla
    7064.099     7737.281     6931.632     9307.667
```

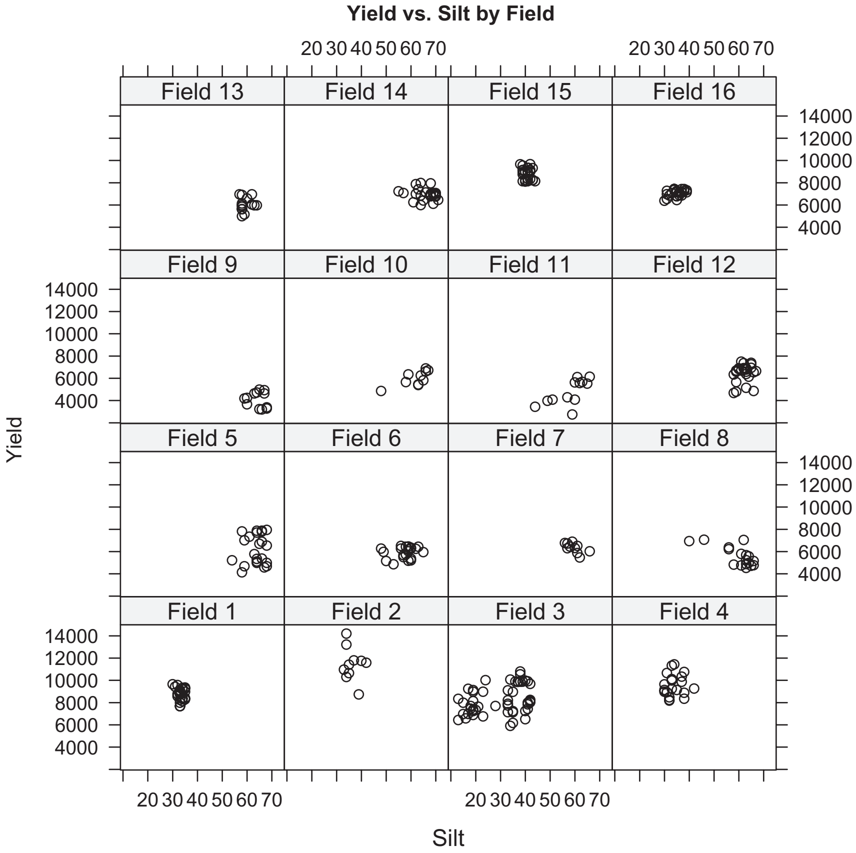

FIGURE 7.20
Trellis plot of yield vs. silt content for each of the individual fields of Data Set 3.

Obviously, there is a substantial difference in yields between the varieties. Let's look at who used which variety.

```
> with(data.Set3, unique(Farmer[which(Variety == "INIA Tacuarí")]))
[1] K I L D
Levels: A B C D E F G H I J K L
> with(data.Set3, unique(Farmer[which(Variety == "El Pasol")]))
[1] K F J G H E A L C
Levels: A B C D E F G H I J K L
> with(data.Set3, unique(Farmer[which(Variety == "INIA Olimar")]))
[1] B E H
Levels: A B C D E F G H I J K L
> with(data.Set3, unique(Farmer[which(Variety == "Perla")]))
[1] A
Levels: A B C D E F G H I J K L
```

Only one (northern) farmer used Perla, and he used El Pasol as well. The farmers in all of the regions tended to use all of the varieties. To further disentangle the data, we can plot a trellis box and whiskers plot (Figure 7.21).

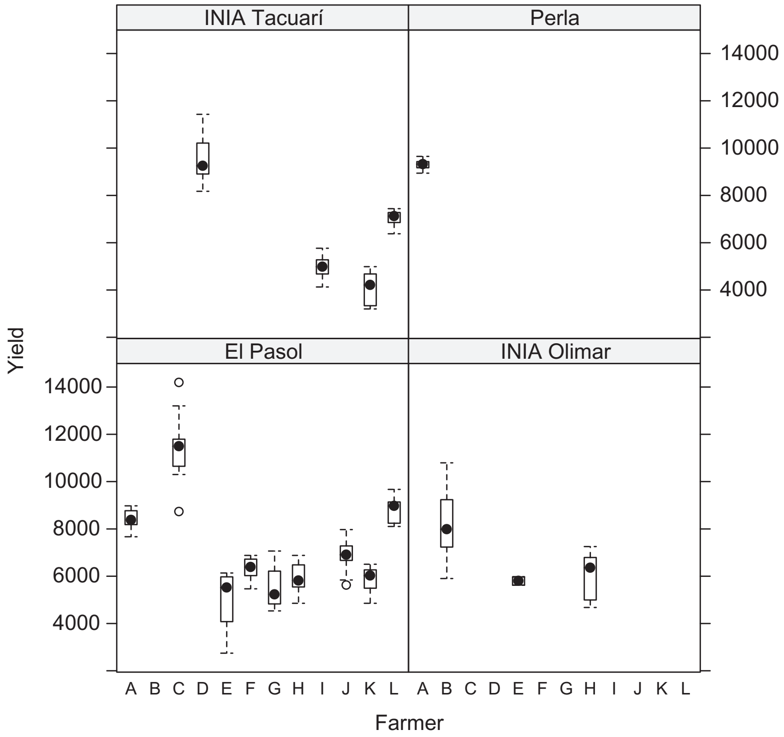

FIGURE 7.21
Trellis plot of yield by variety and farmer.

```
> bwplot(Yield ~ Farmer | Variety, data = data.Set3,
+    xlab = "Farmer") # Fig. 7.21
```

This does not appear to indicate any pattern of yield by variety.

To summarize the results we have obtained so far in the exploratory analysis of Data Set 3, we have found a pair of farmers, one in the central region and one in the north, who seem to consistently obtain better yields than the rest. Climate does not appear to have had an important influence on year-to-year yield variability in the three seasons in which data were collected, and terrain does not appear to vary much between fields. Variety does appear to have a strong association with yield, but the pattern of variety and yield is complex. In the exercises, you are asked to further explore the management variables associated with Data Set 3.

7.5 Data Set 4

Data Set 4 consists of data from the 1995–96 through the 1999 growing seasons from two fields in central California. During the 1995–1996 season the fields were planted to wheat in December 1995 and harvested in May 1996. During the remaining years, the fields were both planted to summer crops. Both fields were planted to tomato in 1997. One field, denoted Field 1, was planted to beans in 1998 and to sunflower in 1999. The other field, denoted Field 2, was planted to sunflower in 1998 and to corn (i.e., maize) in 1999.

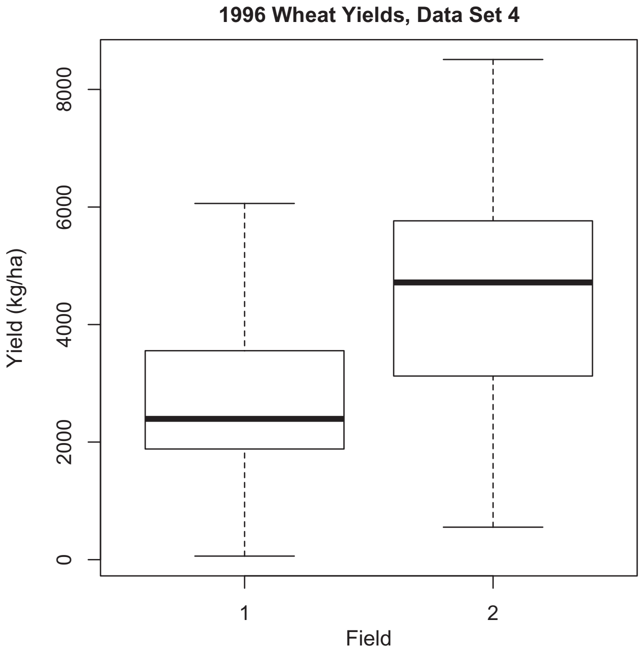

FIGURE 7.22
Boxplots of 1996 wheat yields from the two fields in Data Set 4.

The data from 1997 through 1999 will be analyzed in Chapter 15. Prior to that chapter, we will focus exclusively on the wheat data from the 1995–96 season. The primary factor influencing economic return from wheat is grain yield, with protein content as a secondary influence. The principal objective of the analysis is to determine the principal factors underlying spatial variability in grain yield.

To create the object `data.Set4.1` describing Field 1 we first read the file *set4.196sample.csv*, which contains the values at the 86 sample locations. We then add the data field `Yield`, which contains data from the file created in Exercise 6.10 of cleaned yield monitor data interpolated to the 86 locations (see R code in Appendix B.4). Then we do the same for Field 2. As an initial exploratory step, we compare the yield distributions of the fields using a set of box-and-whisker plots. Figure 7.22 shows boxplots of yield for the two fields. Plotting of extreme values is suppressed. Typical wheat yields in this part of California range between 6,000 and 8,000 kg ha^{-1} (Anonymous, 2008), so both fields clearly had some problem areas. The yield distribution of Field 1 is right-skewed, with a mass of relatively low values, while the distribution of Field 2 is somewhat left-skewed. As with Data Set 2, creating a similar boxplot with `ggplot()`requires that the ordinate be an ordinal scale variable.

The discussion of data exploration will focus on Field 1. The exploration of Field 2 is left to the exercises. As a first step, we examine the high spatial resolution data, looking for similarly shaped spatial patterns. For Field 1, these data consist of the yield map and gauge data derived from the three aerial images taken in December 1995, March 1996, and May 1996. Each image consists of three radiometric bands (Lo and Yeung, 2007, p. 294), infrared, red, and green. The numerical value of each band is an integer between 0 and 255 (which is $2^8 - 1$), with higher values indicating higher intensity in that band. The quantity most often used in remote sensing to estimate vegetation density is the *normalized difference vegetation index*, or NDVI, which is defined as (Tucker, 1979)

$$NDVI = \frac{IR - R}{IR + R} \tag{7.2}$$

Since healthy vegetation reflects infrared much more strongly than red, a high NDVI indicates dense vegetation, whereas an NDVI near zero indicates little or no green vegetation. Figure 7.23a shows the infrared band (IR) from the December image, the NDVI calculated from the March and May images, the IR band from the May image, and the interpolated yield. These were found by a process of trial and error to provide the most information

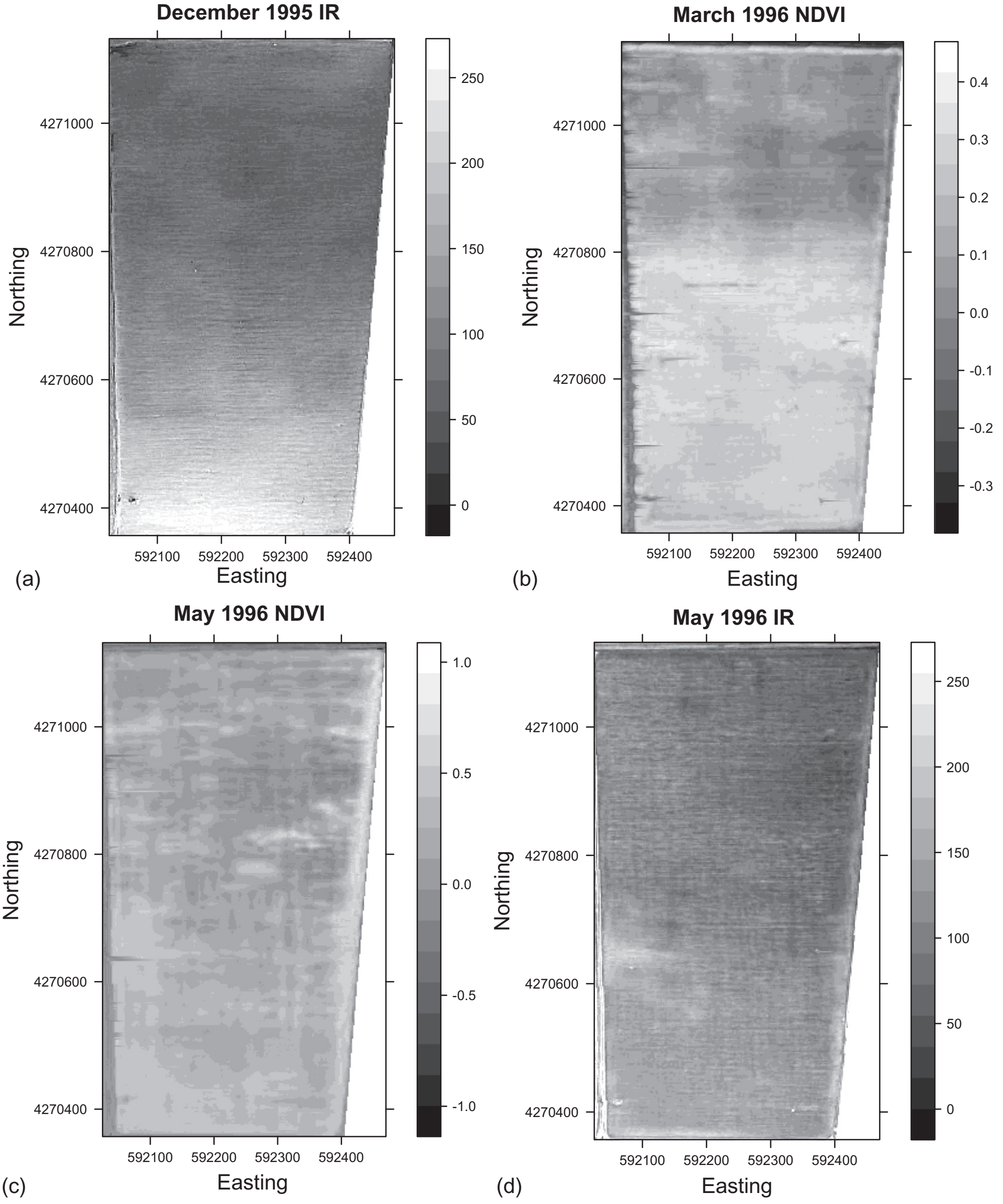

FIGURE 7.23
Plots of data from three aerial images taken of Field 4.1 during the 1995–96 season together with interpolated yield monitor data: (a) December infrared; (b) March NDVI; (c) May NDVI; (d) May infrared. (*Continued*)

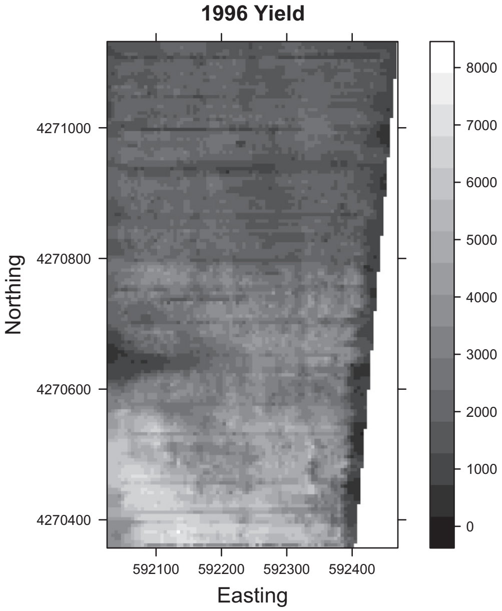

FIGURE 7.23 (Continued)
Plots of data from three aerial images taken of Field 4.1 during the 1995–96 season together with interpolated yield monitor data: (e) interpolated yield.

about the field conditions. To create this figure it was necessary to crop the December infrared image so that it was restricted to the field boundary. We saw how to do this in Section 5.4.3 using the function `over()`. In this section's code, we use a convenient shortcut made possible by polymorphism and the fact that the left bracket `[` is actually a function (Section 2.3.2). Because of this, we can write

```
    f.img <- img.df[bdry.spdf,]
>
```

and the expression on the right is equivalent to `img.df[!is.na(over(img.df, bdry.spdf),]` (Pebesma, 2018). This makes the operation look just like the data frame equivalent.

The December infrared reflectance in Figure 7.23a is lower in the north than in the south. The aerial image was taken a few days after a heavy rainstorm (Plant et al., 1999). Water absorbs infrared radiation, so that wet soil tends to absorb this radiation while dry soil tends to reflect it. The image is an indication that the soil in the north end of the field remained wet after the south end had begun to dry. This indicates that the clay content of the soil in the northern part of the field may be higher than that in the south. This is confirmed by the soil sample data presented in Section 3.2.1. Figure 7.24 shows a map of the soil types in the field. This map was constructed by downloading a shapefile of SSURGO soil classification data from the NRCS Soil Data Mart http://soildatamart.nrcs.usda.gov and clipping this shapefile in ArcGIS with that of the field boundary. The northernmost soil type (Ca) is Capay silty clay (Andrews et al., 1972). This soil is characterized by a low permeability. The soil type in the center (BrA) is Brentwood silty clay loam, which is characterized by a moderate permeability. The soil in the south (Ya) is Yolo silt loam, which is also characterized as moderately permeable and well drained. Taken together, Figures 7.23 and 7.24 indicate

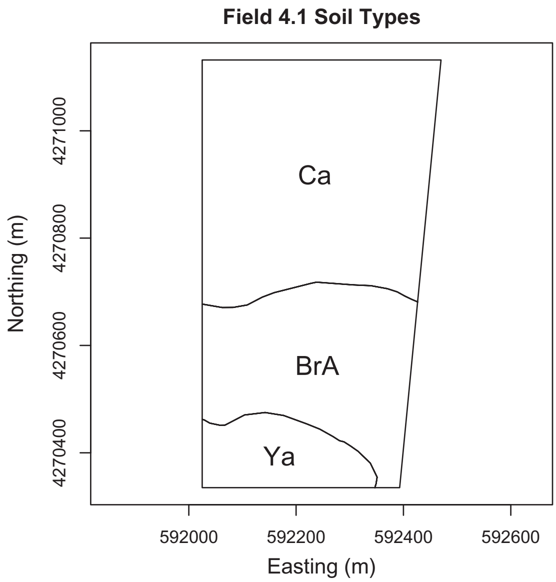

FIGURE 7.24
Soil types of Field 4.1. Ca = Capay silty clay, BrA = Brentwood silty clay loam, Ya = Yolo silt loam.

that Field 4.1 displays a pattern of variation in soil texture. Clay content and sand content have a strong negative association.

```
> with(data.Set4.1, cor(cbind(Sand, Silt, Clay)))
           Sand        Silt        Clay
Sand  1.0000000 -0.30338274 -0.93432993
Silt -0.3033827  1.00000000 -0.05615113
Clay -0.9343299 -0.05615113  1.00000000
```

Therefore, we will eliminate *Sand* from the analysis and instead focus on *Clay* and *Silt* as the texture components.

The southern part of the field has higher yield overall (Figure 7.23e), with the southwest corner having the highest yield. The yield trend generally appears fairly smooth except for two anomalously low areas: a triangular-shaped region on the western edge and the entire eastern edge. The March NDVI data (Figure 7.23b) generally indicates that vegetation density decreases as one moves from south to north. The May infrared data indicates this same trend, but also indicates anomalously high values in the same areas as the low areas in the yield map, namely, the triangular-shaped region and the eastern edge. In the actual study, heavy infestations of wild oat (*Avena fatua* L.) and canary grass (*Phalaris canariensis* L.) were observed in these areas. By May, the wheat was beginning to senesce, but the weeds were still green, and show up as brighter areas in the May infrared image. In summary, visual inspection of the high spatial resolution data suggests that the low yields in the northern half of the field may have been caused by aeration stress and that low yields along the eastern border and in the triangular region on the southwest border were probably caused by weed competition.

We now move to an exploration of the point sample data. In organizing the exploration, we will make use of the variable type concept illustrated schematically in Figure 7.1. Table 7.5 shows a list of variables in Data Set 4.1 together with their type

TABLE 7.5

Variables in Data Set 4

Variable Name	Quantity Represented	Type	Spatial Resolution	Comment
Sand	Soil sand content (%)	Exogenous	Low	
Silt	Soil silt content (%)	Exogenous	Low	
Clay	Soil clay content (%)	Exogenous	Low	
SoilpH	Soil pH	Exogenous	Low	
SoilTOC	Soil total organic C (%)	Exogenous	Low	
SoilTN	Soil total nitrogen (%)	Exogenous	Low	
SoilP	Soil phosphorous content	Exogenous	Low	
SoilK	Soil potassium content	Exogenous	Low	Field 1 only
Weeds	Weed level (1 to 5)	Exogenous	Low	
Disease	Disease level (1 to 5)	Exogenous	Low	
CropDens	Crop density (1 to 5)	Endogenous	Low	
LeafN	Leaf nitrogen content	Endogenous	Low	
FLN	Flag leaf N content	Endogenous	Low	Field 1 only
SPAD	Minolta *SPAD* reading	Gauge	Low	
EM38F	Furrow EM38 reading	Gauge	Low	date sampled
EM38B	Bed EM38 reading	Gauge	Low	date sampled
IR	Infrared digital number	Gauge	High	date sampled
R	Red digital number	Gauge	High	date sampled
EM38	High density EM38	Gauge	High	Field 2 only
GrainProt	Grain protein (%)	Response	Low	
GrnMoist	Grain moist. content (%)	Response	High	
Yield	Yield (kg ha^{-1})	Response	High	

and spatial resolution. No management variables were recorded in the data set. As a first step, we will construct scatterplot matrices of some of the variables. If we consider endogenous variables to be those variables that contribute directly to grain yield or protein content, either through the harvest index (Donald and Hamblin, 1976; Hay, 1995), or through plant physiological processes, then the only endogenous variables are *CropDens*, *FLN*, and *LeafN*. The exogenous variables are considered to influence grain yield indirectly through their impact on the factors contributing to the harvest index. We use scatterplot matrices to eliminate redundant variables and to suggest relationships that we may wish to explore further. We will start with some of the endogenous and exogenous variables and their relationships with the response variables (we refer to this combination as the "agronomic variables"). The code to construct a scatterplot matrix is as follows.

```
> library(lattice)
> agron.data <- subset(data.Set4.1,
+     select = c(Clay, Silt, SoilP, SoilK, SoilpH, SoilTOC, SoilTN,
+ LeafN, FLN, GrainProt, Yield))
> splom(agron.data, par.settings = list(fontsize=list(text=9),
+   plot.symbol = list(col = "black")), pscales = 0,
+   main = "Agronomic Relationships, Field 4.1") #Fig. 7.25
```

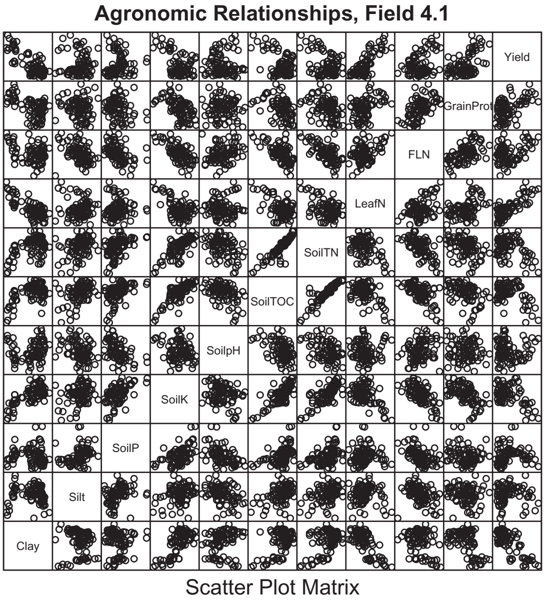

FIGURE 7.25
Scatterplot matrix for "agronomic" data of Field 4.1.

Based on the scatterplot matrix (Figure 7.25), *SoilTOC* and *SoilTN* have a strong positive association. They are sufficiently collinear that one can be removed. *SoilTN*, *SoilK*, and *Clay* are positively associated. *SoilP* appears to have a parabolic relationship with *Clay*. The scatterplots of relationships between *Clay* and some of the other variables appear to indicate that they actually describe two separate relationships. Since *Clay* has a strong north-south trend, it might be suspected that these two separate relationships occur in two separate regions of the field.

Soil phosphorous and potassium are generally regarded as reliable indicators of crop fertilizer demand in California wheat (University of California, 2006). Soil nitrogen level is not, because plant-available forms of nitrogen are highly water soluble and thus very labile, and moreover, a good deal of the nitrogen content of soil is not readily plant-available (Singer and Munns, 1996, p. 215). Fertilization guidelines for nitrogen and potassium in California are based on the level of nitrate-nitrogen and potassium measured in the stem (University of California, 2006). Phosphorous guidelines are based on soil sampling. Data collected in the field were flag leaf nitrogen content, soil phosphorous level, and soil potassium level. In precision agriculture applications, the principal initial questions are whether these data are uniform over the field, and whether any of them are at a level that indicates that an input application is warranted. If an application is warranted, then a further question is whether application at a variable rate is justified, both agronomically and economically, or whether application at a uniform rate are sufficient. The *null hypothesis of precision agriculture* (Whelan and McBratney, 2000) is that precision applications are not economically justified.

The recommended stem nitrate-nitrogen level for wheat in California, measured at the boot stage of plant growth (early in the reproductive phase), is 4,000 ppm, or 0.4% (University of California, 2006). Measurements in Field 4.1 were made between the boot and anthesis stages, so one might expect the level to be a bit lower. The ratio between flag leaf and stem and nitrate concentration at anthesis is about 2.3–2.5 to 1 (Jackson, 1999).

Therefore, one would expect to see a response to fertilizer nitrogen if the flag leaf nitrate-nitrogen level is below about 1 ppm. The soil phosphorous level below which a phosphate application is recommended is 6 ppm (University of California, 2006). Yield response to potassium is rare in California wheat production and is limited to soils with potassium levels less than 60 ppm (University of California, 2006). As an initial check of the measured values relative to these guidelines, we can create stem-and-leaf plots (Tukey, 1977 p. 8; Chambers et al., 1983, p. 26; Tufte, 1990, p. 46) or histograms (Cleveland, 1985, p. 125; Chambers et al., 1983, p. 24); for this application, we choose the former. Applying the function `apply()` to the function `stem()` rapidly performs the calculation of all of the plots as shown here.

```
> apply(with(data.Set4.1, cbind(FLN, SoilP, SoilK)), 2, stem)
```

The results are shown in table 7.6. Based on the stem and leaf plots, one can conclude that nitrogen and potassium levels in the field are adequate, but that phosphorous levels may be low in some areas.

Returning to Figure 7.25, let's explore the relationship between the response variables and some of the quantities that may affect them. *Yield* has a boundary line relation (Webb, 1972) with *Weeds* and, to a lesser extent, with *Disease* and flag leaf nitrogen (*FLN*). Again, this may indicate that in some parts of the field weed level, disease, or nitrogen is the most limiting factor. *Yield*, like *CropDens*, actually has a negative association with *SoilK*, and it has a sort of "sideways parabolic" relationship with *SoilP* and *SoilTN*; that is, if the explanatory variables are plotted on the abscissa and *Yield* on the ordinate, the relationship is parabolic. This is a further indication that difference relationships may occur in different parts of the field, and it indicates the presence of an *interaction*, in which the relationships of yield with these explanatory variables depend on the value of some other explanatory variable, either measured or unmeasured. All of the parabolas appear to have their turning points at a *Yield* value of about 3800 kg/ha.

Table 7.7 shows stem and leaf plots that allow us to examine the distributions of three of the variables. *Clay*, which is representative of soil conditions, is positively skewed and has a bimodal distribution. *Yield*, on the other hand, is negatively skewed. *GrainProt* is fairly symmetric, but is interesting in another context. Wheat at a higher protein content than 13% receives a price premium, which is why this represents a second response variable. About half the field is above this 13% value.

TABLE 7.6

Stem and Leaf Plots for Variables Related to Soil Fertility.

<table>
<tr><th><i>FLN</i> (%)</th><th><i>SoilP</i> (ppm)</th><th><i>SoilK</i> (ppm)</th></tr>
<tr><td>The decimal point is 1 digit(s) to the left of the |</td><td>The decimal point is at the |</td><td>The decimal point is 1 digit(s) to the right of the |</td></tr>
<tr><td></td><td>2 | 5815678</td><td></td></tr>
<tr><td>26 | 0</td><td>4 | 2233348000023335556689</td><td>8 | 869</td></tr>
<tr><td>28 | 21</td><td>6 | 00011112445712223578888</td><td>10 | 7</td></tr>
<tr><td>30 | 22455779011246788</td><td>8 | 1123355667799991446788</td><td>12 | 360</td></tr>
<tr><td>32 | 0134557789001112334455567</td><td>10 | 12236</td><td>14 | 045567789</td></tr>
<tr><td>34 | 000133566788900233555688</td><td>12 | 014345</td><td>16 | 034669023689</td></tr>
<tr><td>36 | 02313478</td><td>14 |</td><td>18 | 4578899011222445577788</td></tr>
<tr><td>38 | 1137898</td><td>16 | 9</td><td>20 | 01111367899025669</td></tr>
<tr><td>40 | 1</td><td>18 | 1</td><td>22 | 0333581144579</td></tr>
<tr><td></td><td></td><td>24 | 447069</td></tr>
</table>

TABLE 7.7

Stem-and-Leaf Plots for Clay, Grain Protein Percent, and Yield.

Clay	GrainProt	Yield
The decimal point is at the |	The decimal point is at the |	The decimal point is 3 digit(s) to the right of the |
22 | 279		
24 | 68	11 | 77899	1 | 2344556666777777788888999999
26 | 559934578	12 | 0000122344	2 | 0000111111223344555577899999
28 | 0338	12 | 5566777788889	3 | 12334557889
30 | 5	13 | 000011111122233344444	4 | 0112444888
32 | 2	13 | 5555566667777777888899999	5 | 124569
34 | 483456	14 | 0001244	6 | 267
36 | 245670011	14 | 5579	
38 | 1145678935	15 | 2	
40 | 1225567801244555999		
42 | 44456677790355678		
44 | 0135		
46 | 9		

Scatterplots of *Clay* versus several other variables indicate that the boundary *Clay* value between the two groups is about 0.3 (i.e., 30% clay). We can use Thiessen polygons (Section 3.4.4) to visualize the location of sites with a *Clay* value less than 30%. As usual, we need to be careful to ensure that the Thiessen polygon ID variables match with the attribute data ID variables.

```
> all.equal(thsn.spdf$ThPolyID, data.Set4.1$ID)
[1] TRUE
> slot(thsn.spdf, "data")$Clay30 <- factor(data.Set4.1$Clay <= 30,
+   labels = c("Greater than 30%", "Less than 30%"))
> greys <- grey(c(100, 200) / 255)
> spplot(thsn.spdf, zcol = "Clay30", col.regions = greys,
+   main = "Field 4.1 Clay Content Categories") # Fig. 7.26
```

All of these locations are in a contiguous area in the south of the field (Figure 7.26). In Exercise 7.15, you are asked to plot a high spatial resolution version of Figure 7.26 by using linear regression to predict clay content based on December IR value.

A different perspective on the spatial relationships among the attribute values may be obtained using star plots (Chambers et al., 1983, p. 161), sometimes called sun ray plots (Jambu, 1991), in the form of a map. Star plots, like other multivariate graphical devices, come very close to being what Tufte (1983, p. 153) calls a “graphical puzzle,” that is, they can be difficult to interpret (Cleveland, 1985, p. 20), but they do provide an overview of general patterns. My own opinion is that star plots with four variables are easiest to interpret; more variables produce a graphical puzzle and fewer variables produce stars with too few “points.” The code for a plot of soil nutrients and clay content is as follows.

```
> star.data <- with(data.Set4.1, data.frame(SoilP, Clay,
+   SoilK, SoilTN))
> star.loc <- 2* cbind(data.Set4.1$Column, 13 - data.Set4.1$Row)
> stars(star.data, locations = star.loc,
+   labels = NULL, key.loc = c(18,2),
+   main = "Nutrient Star Plot", cex.main = 2) # Fig. 7.27
```

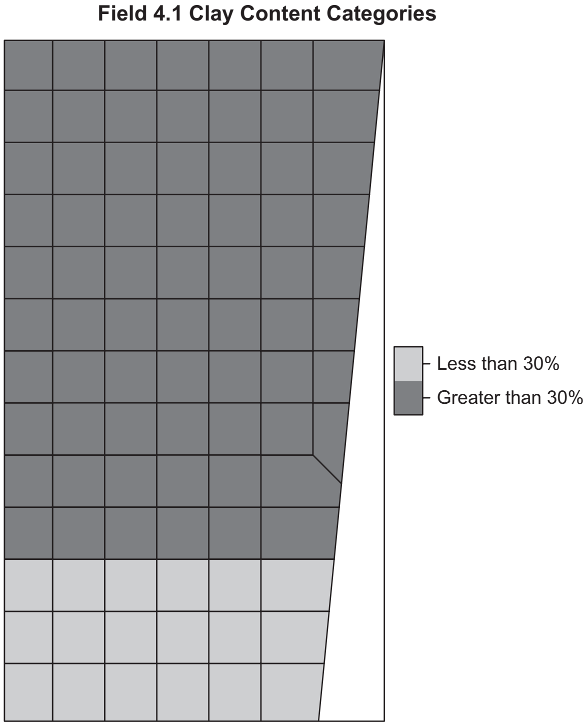

FIGURE 7.26
Thiessen polygon plot of regions of Field 4.1 as defined by clay content. The Thiessen polygons were created in ArcGIS.

The row and column numbers of the attribute data set are manipulated into x and y coordinates. The calculation of these x and y coordinates in line 3 involved a bit of trial and error. The function `stars()` is called in line 4 with arguments that suppress the plotting of labels (these would be the number of each data record) and that locate the legend (again this involved a bit of trial and error). The star map in Figure 7.27 indicates that all soil nutrient levels are lowest in the southern (high yielding) region and that phosphorous is also low in the northern part of the field.

In Exercise 7.16, you are asked to create a star map for yield in relation to weed and disease level. The highest weed and disease levels occur in the central portion of the field and along the northern and eastern edges. Star plots are useful for identifying interactions. One potential interaction of interest involves soil potassium content and leaf nitrogen content. Potassium deficiency is known to inhibit the utilization of nitrogen (Gething, 1993; Marschner, 1995, p. 299). If such an interaction is present in these data, it should manifest itself in lower values of flag leaf N and yield in low K soils than in high K soils, conditional on soil N levels. This is not indicated in the star plot of Exercise 7.16b. The last phase of the initial data exploration involves recognizing that two quantities, soil phosphorous and grain protein content, are below a meaningful threshold in some parts of the field. To determine these locations precisely, we can apply the function `spplot()` to the Thiessen polygons file. The results are shown in Figure 7.28.

In summary, the preliminary analyses indicate that the gross trend of declining yield in the north end of the field is probably caused by aeration stress associated with moisture retention by the heavy clay soil. In other parts of the field, weed competition may

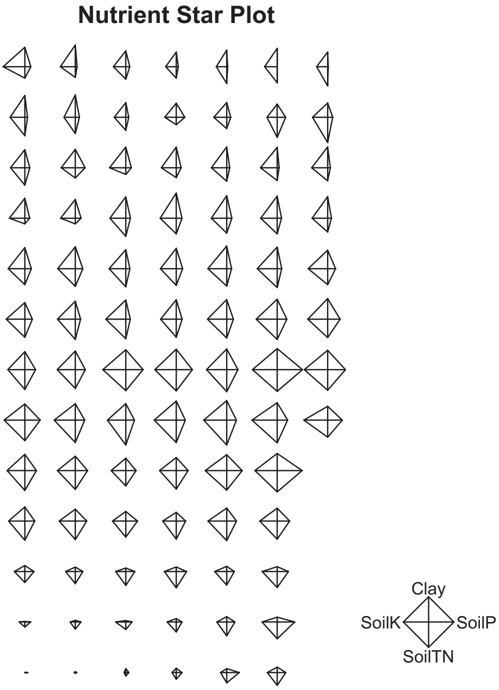

FIGURE 7.27
Star maps of the relationship between soil mineral nutrient levels and clay content.

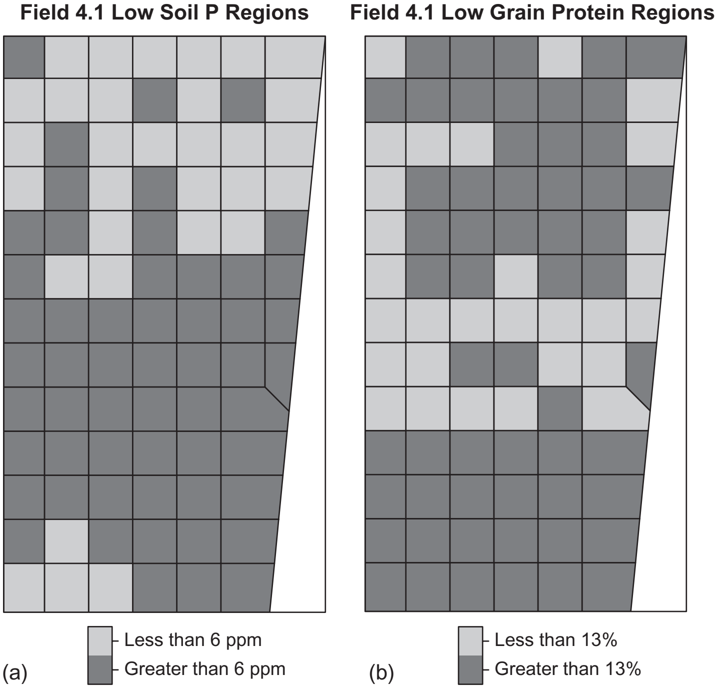

FIGURE 7.28
Relationship to a meaningful threshold for two quantities: (a) soil phosphorous level, and (b) grain protein percent. In each case, the gray lattice cells are those in which the quantity falls below the threshold.

limit yield. The crop appears to have adequate nutrient levels for nitrogen and potassium, but there may be some areas in which phosphorous is deficient. Of the mineral nutrients, neither potassium nor nitrogen is at any sample point below the level at which a yield response to increased fertilization would be expected. However, there is an indication that soil phosphorous levels might be low in the northern part of the field. The "sideways parabolic" relationship of yield with soil phosphorous level is a further indication of this possibility. Yield has a boundary line relationship with weeds, disease, and flag leaf nitrogen level. This might be consistent with a "law of the minimum" (Barbour et al., 1987) interpretation that, for example in the case of nitrogen, in some areas of adequate nitrogen level, some other factor inhibits yield. The area of high flag leaf N level and low yield is in the north of the field. Yield and grain protein content are highly correlated.

7.6 Further Reading

The books by Tufte (1983, 1990) provide an excellent discussion of the use of graphics in data analysis, as do those of Cleveland (1985, 1993). The utility of trellis graphics was first pointed out by Cleveland (1993), who referred to them as "multiway plots." Jambu (1991) provides a discussion of exploratory data analysis using the methods covered in this chapter and others as well. Andrienko and Andrienko (1998) provide a very extensive discussion of map analysis and graphical methods associated with spatial data. Sibley (1988) is also a good early source in this area.

An excellent introduction to mapping with R is provided by the NCEAS Scientific Computing: Solutions Center Use Case: Creating Maps for Publication using R Graphics, located at the link "Create Maps With R Geospatial Classes and Graphics Tools" on the website https://www.nceas.ucsb.edu/scicomp/usecases/CreateMapsWithRGraphics. The free software package GeoDa (https://spatial.uchicago.edu/geoda/) contains many spatial data exploration tools.

Cressie and Wikle (2011) advocate a hierarchical modeling approach that is very similar in concept to the division of variables into categories as represented in Figure 7.1.

Exercises

7.1 By comparing the values `ID` and `obsID`, verify that the creation of the object `Set1.obs` in Section 7.2 correctly links cuckoo observation points and habitat patches. If you have access to a GIS, you should do this by comparing habitat and observation point data files.

7.2 The object `Set1.obs`, which contains the information about each polygon in Data Set 1 that contains a cuckoo observation point, includes variables *HtClass*, *AgeClass*, and *CoverClass*. When applying the function `AreaRatio()` to compute the habitat ratios, why is it necessary to use the data in `data.Set1` rather than simply using `Set1.obs`?

7.3 In Section 7.2, a set of size class ratios is computed for use in the Wildlife Habitat Relationships model for the yellow-billed cuckoo expressed in Table 7.1. Carry out a similar construction for the ratio of age classes in this model.

7.4 (a) Compute the suitability score for age class in the CWHR model of Section 7.2. (b) Repeat the computation for the height class suitability score.

7.5 The CWHR model for the yellow-billed cuckoo produces two misclassified patches having a suitability score of zero but where cuckoos are observed. Examine the properties of these locations. Can you conclude anything from them?

7.6 Repeat the contingency table analysis of Section 7.2 with subsets of the explanatory variables and determine which have the most explanatory power.

7.7 Use the function `locator()` to separate out of Data Set 2 that part located in the Klamath range (this is the square shaped northernmost portion of the data). Plot a map of the region and compare it with your maps generated in Exercise 6.1.

7.8 Plot the presence and absence of blue oaks in the Klamath range against elevation.

7.9 Construct scatterplots of precipitation against elevation for the Sierra Nevada and the Coast Range and compare them.

7.10 Use the function `table()` to construct a contingency table of texture class vs. permeability class for Data Set 2. Are the two closely related?

7.11 The application of the function `hexbin()` at the end of Section 7.3 showed that for values of *Precip* less than 800, the distribution of *MAT* is bimodal. Generate the `hexbin()` frequency plot of *MAT* and *Precip* when *Precip* is restricted to this range, Then repeat for *JuMax* and *JaMin*.

7.12 Use the functions of the `spatstat` package to plot the theoretical and observed plots of Ripley's *K* for (a) the random sample pattern of Section 5.3.1; (b) the 32 point grid sample pattern of Section 5.3.2; and (c) the clustered pattern of Section 5.3.5. How do your results (especially those of the clustered pattern) correspond with the interpretation of Ripley's *K*?

7.13 Use the function `xyplot()` to determine whether any of the measured soil quantities in Data Set 3 appears to have a linear association with yield across fields and, if so, which appears to have the highest level of association.

7.14 Repeat Exercise 7.12 plotting yield against management quantities by location and year.

7.15 The C:N ratio, that is, the ratio of total carbon to total nitrogen, is an important ecological quantity in soil science. The soil total organic carbon content is equal to 0.58 times the soil total nitrogen content. Compute the C:N ratio for the 86 sample locations in Field 4.1. If the C:N ratio becomes too high (say, above 30:1), then the demand from soil microorganisms for nitrogen becomes sufficiently high that nitrate will not remain in soil solution and thus will be unavailable for root uptake (Brady, 1974, p. 291).

7.16 Use the December IR image to assist in assessing the distribution of clay content in Field 4.1. (a) Use `over()` to select the pixels values in the December IR image that align with sample points, and compute a regression of IR digital number on clay content. (b) Use this regression relationship to predict clay content at each pixel based on IR values and plot the regions above and below 30% clay.

7.17 (a) Construct a star plot of disease, weeds, and yield for Field 4.1. (b) Construct a star plot of soil total nitrogen, flag leaf nitrogen, potassium, and yield to determine whether there is an interaction between N and K in this field.

7.18 Compare using stem and leaf plots the clay content of Field 4.2 and Field 4.1. Then create a scatterplot showing the relationships between yield and clay content in both fields. Are the relationships similar? Do you expect that clay content will play as significant a role in Field 4.2 as it does in Field 4.1?

7.19 Construct a scatterplot matrix of yield in Field 4.2 with some of the variables that might affect it. Can you identify with this the most likely cause of the large low yield areas on the western side of the field? Check your answer by plotting thematic point maps of yield and this variable.

7.20 Determine whether there are any areas in Field 4.2 that are nitrogen or phosphorous deficient according to University of California standards for wheat in California.

7.21 (a) Construct a star plot of yield, grain moisture, protein content, and leaf nitrogen. (b) Construct a star plot of yield vs. weeds and disease.