17

Assembling Conclusions

17.1 Introduction

Formulating conclusions is the most important part of the analysis of a data set, and the most subjective. To interpret the results of the analysis, the investigator must apply his or her knowledge of the biophysics of the system, and as a result different people may legitimately look at the same statistical results and reach different conclusions. To the extent that the statistics support prior assumptions they may be regarded as an affirmation of these assumptions, but to the extent that they conflict with them, one must decide how to respond to this conflict. The analysis of each of the four data sets that has occupied this book has been a simulation of an actual research project. The objectives that were established in Chapter 1 are not actual current research objectives. Rather, the objectives were established, based on the research objectives of the original projects, for expository purposes, as questions that could be could legitimately be posed prior to the collection of the data and then addressed by the analysis of that data. Nevertheless, the fact that different people may use the same data set to reach entirely different conclusions still applies. For this reason, my goal in this chapter is not to convince you that my conclusions are correct, but rather to present an example of how one might approach the highly personal process of developing an interpretation of the implications of their analytical results. Others may come to different conclusions based on the same data, and the intent of this chapter is not to advocate for one particular approach but rather to provide an example of an approach, some parts of which you may find appropriate and some parts of which you may choose to ignore.

17.2 Data Set 1

The objective of the analysis of Data Set 1 is to test the predictive capacity of the California Wildlife Habitat Relationships (CWHR) model of habitat suitability for the yellow-billed cuckoo. The model involves four variables: patch area, patch width, floodplain age distribution as a surrogate for vegetation species distribution, and the ratio of area of tall vegetation (>20 m) to total area (Table 7.1). The latter two variables enter into the model in a non-monotonic fashion in which an intermediate value is optimal. A fifth variable, patch distance to water, is accounted for by excluding patches farther than 100 m from the river.

The data set does not allow us to distinguish effectively between patch area and patch width as limiting factors in determining habitat suitability, but a comparison of contingency

tables in which one or the other was excluded indicates that for this data set, patch area plays a more critical role. In the exercises of Chapter 7, further contingency table tests with various combinations of the variables excluded were carried out. The results of the contingency table analysis indicate that a model including only *PatchArea* and *AgeRatio* suitability scores generates the same contingency table as the model with the full set of four variables. The table contains five true positives, twelve true negatives, two misclassified positives, and one misclassified negative. Among the single variable models, the contingency table of the model with only *AgeRatio* is closest to this. The Fisher test was used in Section 11.2 to test the null hypothesis that the model is not related to the presence-absence data. The test generated a p value of about 0.01 uncorrected and 0.07 when corrected for spatial autocorrelation.

Logistic and zero-inflated Poisson regression models were constructed for the system in Sections 8.3.3 and 8.3.4. These models also indicated that the most important predictors in the model are *PatchArea* and *AgeRatio*, although the manner in which these enter the model differed depending on the model (logistic vs. Poisson). Because of the relatively small number of data records, no attempt was made to incorporate spatial autocorrelation directly into these models.

Let's take a closer look at the habitat score data. After re-running the code from Section 7.2 and Exercises 7.2 and 7.3 we generate the `data frame Set1.corrected`, which contains the data use to carry out the analyses in the subsequent chapters. Here are the data, ordered by observation point ID number.

```
> d.f <- data.frame(with(Set1.corrected, cbind(PatchID,
+    AreaScore, AgeScore, HeightScore, obsID, HSIPred, PresAbs)))
> d.f[order(d.f$obsID),]
   PatchID AreaScore AgeScore HeightScore obsID HSIPred PresAbs
16     175      0.33     0.00        0.00     2       0       0
15     170      0.00     0.00        0.00     3       0       0
14     164      0.33     0.66        1.00     4       1       1
13     155      1.00     0.66        0.66     5       1       1
12     130      0.33     0.00        0.33     6       0       0
11     122      0.66     0.66        0.66     7       1       0
10     116      1.00     0.00        0.00     8       0       0
19     224      0.00     0.00        0.33     9       0       0
9      106      0.00     0.33        0.33    10       0       0
8      100      0.66     0.00        0.00    11       0       1
7       97      0.00     0.00        0.66    12       0       0
6       86      0.00     0.00        0.66    13       0       0
20     400      0.66     0.00        0.66    14       0       0
18     210      0.00     0.33        0.33    15       0       0
4       60      0.66     0.00        0.66    16       0       0
5       63      1.00     0.33        0.66    17       1       1
17     209      0.00     0.33        0.33    18       0       1
3       31      0.33     0.33        0.33    19       1       1
2       22      0.00     0.00        0.00    20       0       0
1       19      1.00     0.33        1.00    21       1       1
```

Of the 20 locations, there are none in which *HeightScore* is 0 and no other habitat score is 0. Indeed, every patch with *HeightScore* = 0 also has *AgeScore* = 0. The data set, therefore, cannot tell us anything about the capacity of *HeightScore* to limit patch suitability. One patch, patch 209, has *AreaScore* = 0 and positive values for both *AgeScore* and *HeightScore*, and is a false negative. This is misleading, however, because it happens that patch 209 is

a very narrow patch classified as lacustrine that is in the center of a much larger patch classified as riverine (Exercise 7.4). In conclusion, as with patch width, we cannot provide any solid evidence for or against the predictive power of vegetation height in the model.

Our objective is to use a data set to test a classification of data based on an already specified model, as opposed to generating a new model. There is a statistic available, Cohen's kappa (Cohen, 1960; Lo and Yeung, 2007, p. 121), that is specifically designed for this objective. This statistic, which was briefly discussed in Section 15.4.1, accounts for the fact that, when comparing a predicted classification with the actual observed classification, some of observations will be correctly classified by chance. Cohen's kappa corrects for this chance agreement. It is defined as follows. Consider again the site classification matrix.

```
> 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
```

This matrix describes the level of agreement between the predicted classification of sites of the CWHR model and the actual observed classification of sites. If it were a pure diagonal matrix, then all of the sites would be correctly classified. The off-diagonal elements are the number of misclassified sites. In this context, the matrix is called a *confusion matrix*. Let P_0 be the total portion of the sites that are correctly classified. Let P_c be the portion of the sites that would be correctly classified by chance if the classification were random. In terms of the marginal sums defined in Section 11.3.1, $P_c = n_{m1}n_{im} + n_{2m}n_{m2}$. The kappa statistic is given by

$$\kappa = \frac{P_0 - P_c}{1 - P_c}. \tag{17.1}$$

If all of the data are correctly classified then $\kappa = 1$, and if the classification is no better than random, then $\kappa = 0$. There are numerous packages available that contain a function to compute Cohen's kappa. We will use the function `Kappa.test()` of the `fmsb` package (Nakazawa, 2017).

```
> library(fmsb)
> print(kappa.stat <- Kappa.test(cont.table))
$Result
        Estimate Cohen's kappa statistics and test the null hypothesis
that the extent of agreement is same as random (kappa=0)
data:  cont.table
Z = 2.6127, p-value = 0.004491
95 percent confidence interval:
 0.3034305 1.0147513
sample estimates:
[1] 0.6590909
$Judgement
[1] "Substantial agreement"
```

Cohen (1960) showed that the kappa statistic satisfies an asymptotic normality property that permits it to be tested against the null hypothesis of random classification. As with the analysis of contingency table data, however, the small size of Data Set 1 limits the applicability of the results. We can carry out a permutation test similar to that used in Section 11.3.2 with the function `fisher.test()`. First, we carry out the test without restricted randomization.

```
> set.seed(123)
> sample.blocks <- function() sample(Set1.corrected$PresAbs)
> U <- replicate(1999,calc.kappa())
> print(obs.kappa <- Kappa.test(cont.table)$Result$estimate)
[1] 0.6590909
> print(p <- sum(u >= obs.kappa - 0.001) / 2000)
[1] 0.0075
```

Next, we carry out a restricted randomization using the blocks in Figure 11.4. Most of the code is identical to that used in Section 11.3.2 and is not shown here. The function `calc.kappa()` replaces `calc.fisher()`, and the function `sample.blocks()` from Section 11.3.2 replaces the function in the code sequence above.

```
> calc.kappa <- function(){
+    PA <- sample.blocks()
+    UA <- sum(Set1.corrected$HSIPred == 0 & PA == 0)
+    UP <- sum(Set1.corrected$HSIPred == 0 & PA == 1)
+    SA <- sum(Set1.corrected$HSIPred == 1 & PA == 0)
+    SP <- sum(Set1.corrected$HSIPred == 1 & PA == 1)
+    n <- matrix(c(SP, SA, UP, UA), nrow = 2, byrow = TRUE)
+    kappa.stat <- Kappa.test(n)$Result$estimate
+ }
```

Now we run the test.

```
> set.seed(123)
> u <- replicate(1999,calc.kappa())
> print(p <- sum(u >= obs.kappa - 0.001) / 2000)
[1] 0.0725
```

Taking spatial autocorrelation into account, the p value for the kappa statistic is about the same as that for the chi-squared statistic computed in Section 11.3.2. This is not terribly surprising since most of the sites are correctly classified.

The second issue raised above is that the model we are testing is intended to predict habitat suitability, not habitat occupancy. Therefore, an error of bird presence in a habitat patch classified as unsuitable would be considered more serious than an error of bird absence in a habitat patch classified as suitable. The remote sensing and GIS literature deals with this issue through the concept of *producer's accuracy* and *user's accuracy* (Lo and Yeung, 2007, p. 120). In the case of our data set, there is only one suitable site with an absence, and one of the sites with a misclassified presence (patch 209) has been accounted for. Therefore, this is not such an important issue for us.

Let's now consider the results of the generalized linear modeling. Because of the unreliable nature of the abundance data, we will focus on the logistic model. As with the contingency table model, the variables *PatchArea* and *AgeRatio* were picked out as most important. Here is a brief summary of the results:

```
> Set1.logArea <- glm(PresAbs ~ PatchArea,
+       data = Set1.norm1, family = binomial)
> AIC(Set1.logArea)
[1] 25.14631
> Set1.logAge <- glm(PresAbs ~ AgeScore,
+       data = Set1.norm1, family = binomial)
> AIC(Set1.logAge)
[1] 23.79338
> Set1.logAreaAge <- glm(PresAbs ~ PatchArea + AgeScore,
+       data = Set1.norm1, family = binomial)
> anova(Set1.logAge, Set1.logAreaAge, test = "Chisq")
Analysis of Deviance Table
Model 1: PresAbs ~ AgeScore
Model 2: PresAbs ~ PatchArea + AgeScore
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        18     19.793
2        17     17.086  1   2.7069   0.09991.
> summary(Set1.logAreaAge)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7308     0.8444  -2.050   0.0404*
PatchArea     0.9876     0.6570   1.503   0.1328
AgeScore      4.8061     2.7064   1.776   0.0758.
AIC: 23.086
```

The single most important variable in the logistic regression model is *AgeScore*, and the model that includes both this and *PatchArea* is perhaps very slightly better. This is basically in concurrence with the contingency table analysis.

Figure 17.1 is a plot of the data values of *PresAbs* vs. *PatchArea*. The open circles are sites where *AgeScore* = 0, and the black circles are sites where *AgeScore* > 0. Also shown are plots of the fit of the model `Set1.logAreaAge` for *AgeScore* = 0 and for *AgeScore* = 1. The model appears to have been heavily influenced by the data record with *PresAbs* = 1, *AgeScore* = 0, and *PatchArea* ≅ 0.5. Overall, however, the logistic regression model does a reasonable job of fitting the data, with the exception that it appears to overestimate the probability of a patch being occupied when the patch area is high but the age score is low.

In summary, the data provide some support for the CWHR model. The relatively small sample size limits the power of the analysis, and this effect is exacerbated by spatial autocorrelation. To the extent that the data support the model, they also support the notion that the patch suitability is limited by its least suitable aspect. The data do not provide sufficient variability to distinguish the impact of patch width from that of patch area, nor do they provide evidence regarding the importance of the ratio of area of tall vegetation to total area. The one anomalous data record that cannot be explained is patch 100. This is a large patch with uniformly high floodplain age and a relative abundance of tall vegetation, indicating the strong possibility that it is dominated by oaks. The fact that there are four recorded sightings indicates that the chance of a false positive is low. This remains an unsolved mystery.

There is an epilogue to this story. Girvetz and Greco (2009) analyzed a similar data set consisting of 102 sites surveyed along a longer stretch of the Sacramento River. A major improvement in the analysis was to use the *Patch Morph* algorithm (Girvetz and Greco, 2007) to incorporate small isolated areas of different vegetation type into the habitat patches. This provided a better model of the patches as they are perceived by the cuckoos.

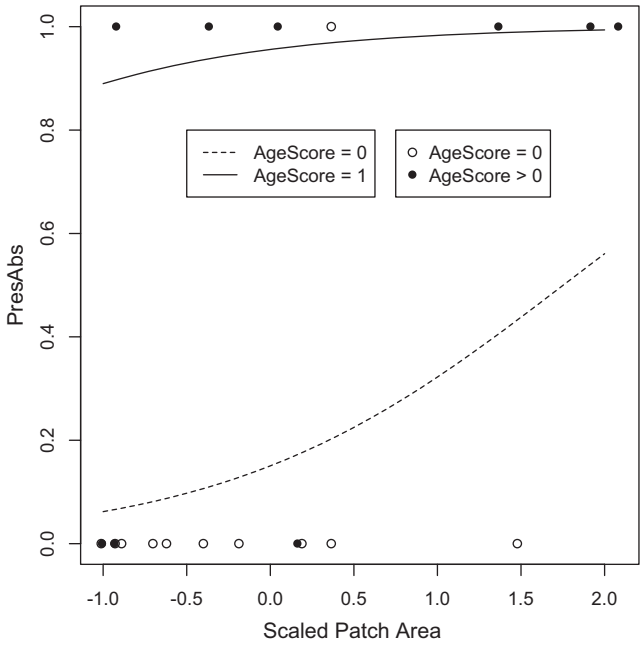

FIGURE 17.1
Plots of yellow-billed cuckoo presence/absence vs. scaled patch area and age score. The curves are predicted values based on the generalized linear model, and the circles are data values.

Girvetz and Greco (2009) used the CWHR model, logistic regression, and classification tree analysis to determine the factors having the greatest influence on patch suitability. Similarly to our analysis, they found that the total area of cottonwood forest within the patch was the most important factor in determining patch suitability. Their larger study area is bisected by a state highway and, surprisingly, they found a large reduction in patch occupancy on the north side of this highway. A legitimate inference from the data would be that the highway serves as some sort of barrier. This is surprising because, as Girvetz and Greco (2009) point out, the birds, having flown all the way from South America, should not find the crossing of a two-lane highway to be a major challenge. Taken together with our own anomalous patch 100 this shows that, despite our best efforts, the yellow-billed cuckoos have not yet revealed all of their secrets.

17.3 Data Set 2

Data Set 2 consists of observations of the presence or absence of blue oak trees together with measurements and calculated estimates of explanatory variables at 4,101 locations in California. Each of these locations was identified as containing at least one species of oak at the time it was sampled. The initial objective was to determine suitability characteristics of the individual sites for blue oak. In Section 12.6, we added a second related but different objective: to determine the characteristics of a geographic region, as defined by elevation and spatial proximity, that affect the average fraction of sites in the region containing a blue oak. We will refer to these as the pointwise and the regional analyses, respectively.

The data set naturally subdivides into four mountain ranges: the Sierra Nevada to the east, the Coast Range to the west, the Klamath Mountains to the north, and the Transverse Range to the south. We analyzed data from the Sierra Nevada and the Coast Range, leaving the Klamath Mountains and the Transverse Range for verification purposes.

The Sierra Nevada has a fairly simple topography, with a gradually increasing elevation as one moves from west to east. Blue oak prevalence declines as elevation increases (Figure 7.7), and there is a strong indication that mean annual precipitation, represented by the variable *Precip*, plays a prominent role in determining blue oak presence. This conclusion is supported by exploratory analysis (Figure 7.12), by the pointwise general linear model analysis under the assumption of independent data records (Section 8.4.2), and by the regional analysis in which spatial autocorrelation is accounted for (Section 12.6). The second explanatory variable in the pointwise model is *MAT*. In the Sierra Nevada, however, *MAT* is highly correlated with *Precip*, and it is unclear whether the estimated *MAT* effect is due to *MAT* or represents a quadratic effect of *Precip.* In the regional Sierra Nevada subset, the primary and secondary roles of *Precip* and *MAT* are exchanged. In both analyses, the soil-related variables *Permeab* and *AWCAvg* are also in the model. The variable *SolRad* is also in the pointwise model, and *PM400* is in the regional model.

Possibly because of its more complex topography, in the Coast Range subset the correlations among the explanatory variables are generally lower. This may help to disentangle the relationships a bit. We first focus on the issue of whether the blue oaks are responding to temperature. Regression tree analysis (Figure 9.10, Section 9.3.3) and contingency table analysis (Exercises 11.2 and 11.3) suggested that high mean July temperatures, low mean January temperatures, or both may favor blue oaks. Figure 17.2a shows a scatterplot of *JuMean* vs. *Precip* in the Coast Range subset, in which blue oak presence and absence is also shown. There appears to be a fairly strong boundary line effect (Abbott et al., 1970; Webb, 1972) of *JuMean*. A similar graph is obtained using *JuMax.* The use of the boundary line model represents an implicit acceptance of Leibig's law of the minimum (Loomis and Connor, 1992, p. 55). This law, as used by agronomists, asserts that yield is limited by the scarcest resource. As interpreted by ecologists, it states that species distribution will be controlled by the environmental factor for which the organism has the lowest range of adaptability (Krebs, 1994, p. 40), or what Bartholomew (1958) calls the *ecological tolerance*.

Figure 17.2a could also be interpreted as indicating that mean annual precipitation influences the boundary line effect of mean July temperature, with lower values of *Precip* corresponding to lover boundary values of *JuMean*. Figure 17.2b shows a plot of *JaMean* vs. *Precip* in the same format. There appears to be a similar boundary line effect, but in the opposite direction: blue oak presence is favored by values of *JaMean* below about 9.2. Figure 17.3a and b show maps indicating the locations that lie on the "wrong" side of the boundary lines in Figure 17.2a and b, respectively. Somewhat surprisingly, there is a strong negative correlation ($r = -0.57$) between *JaMean* and *JuMean* in the Coast Range. Figure 17.4 shows a scatterplot of the two variables identifying the values for data records lying on the "wrong" side of the boundary lines and with $QUDO = 1$. There may be some small indication that mean January temperature is more decisive than mean July temperature, since there are more anomalous data records of the latter type, but the evidence is not very strong. What is the effect of *JaMean* and *JuMean* in the Sierra Nevada? There are very few sites at which the either the value of *JaMean* is above 9.2 or the value of *JuMean* is below 19.8. Therefore, as with precipitation in the Coast Range, the data of the Sierra Nevada subset do not provide a sufficient range to determine the effect of temperature. Neither the Sierra Nevada subset nor the Coast Range subset of the pointwise data provided any convincing evidence for or against the effect of the soil related variables.

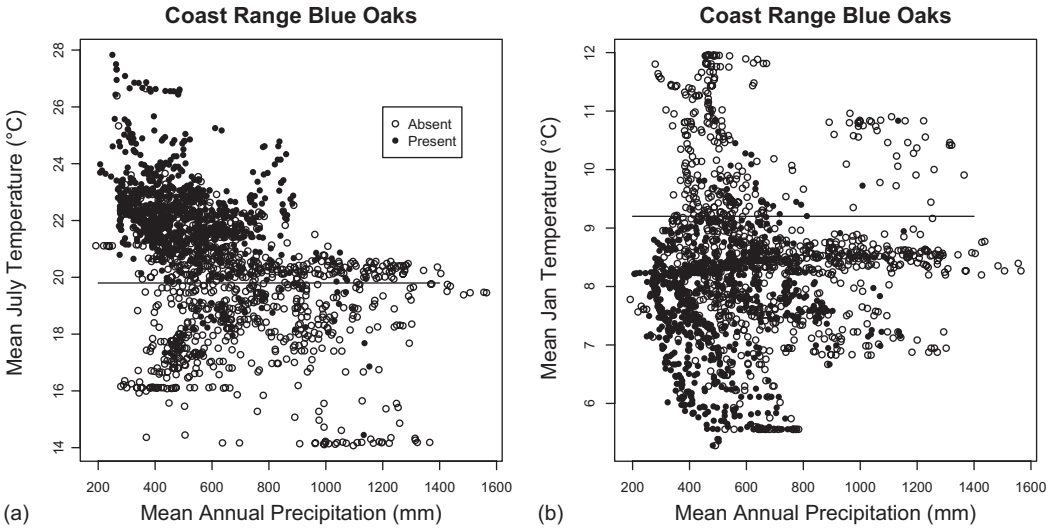

FIGURE 17.2
Scatterplots of (a) mean July temperature and (b) mean January temperature vs. mean annual precipitation. The filled circles indicate locations where a blue oak is present.

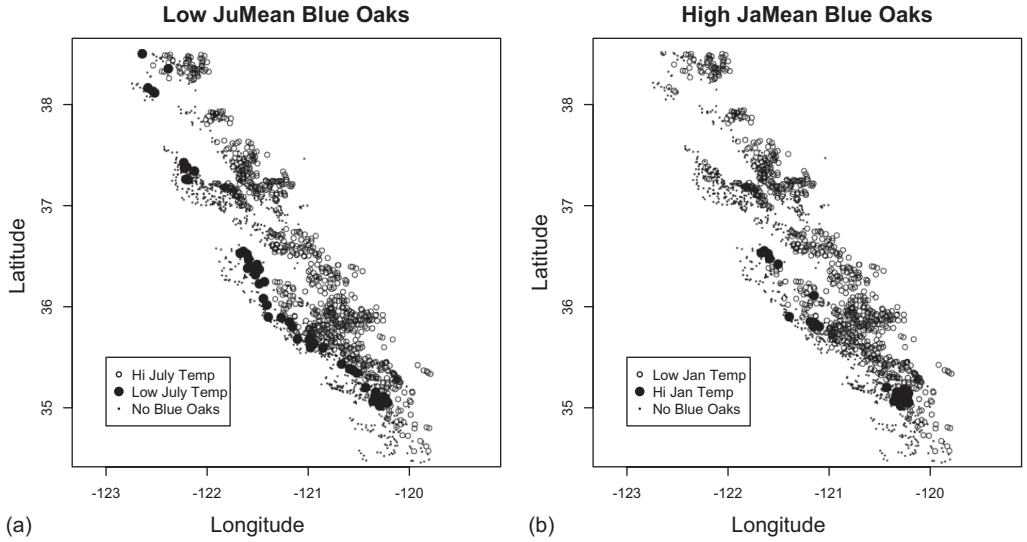

FIGURE 17.3
"Anomalous" locations in the Coast Range. The large filled circles indicate locations where a blue oak is present and: (a) the daily mean July temperature is less than 19.8°C, (b) the daily mean January temperature is greater than 9.2°C.

As a final check, we will test the models on the Klamath Range and Transverse Range data sets, which were not included in the model development. Figure 17.5 shows the results. Unfortunately, neither of these two mountain ranges has a sufficient range of values to truly test the models. The data are not, however, inconsistent with the notion that there is a boundary line temperature effect and that the location of the boundary is influenced by precipitation.

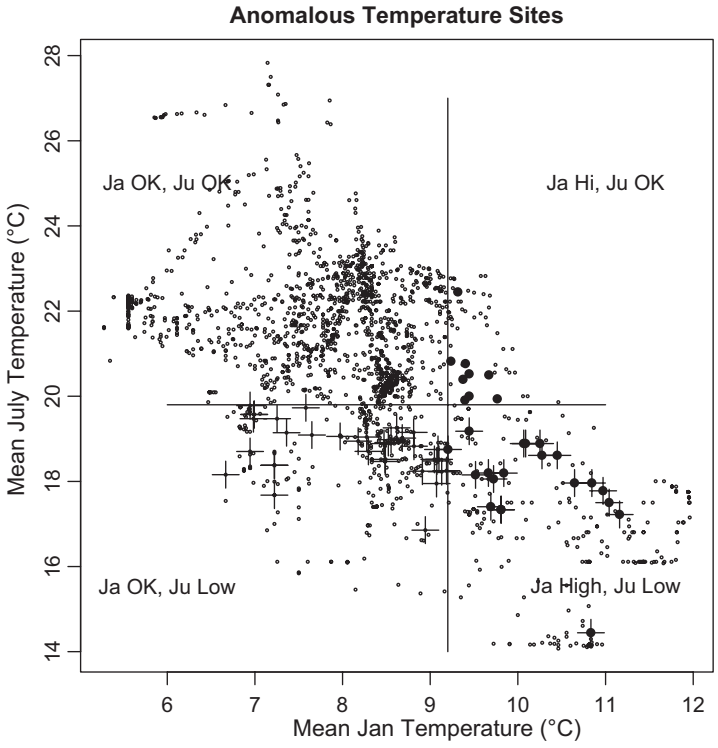

FIGURE 17.4
Scatterplot of daily mean January and July temperatures. The crosses indicate "anomalous" (i.e., low mean) July temperature sites containing a blue oak, and the filled circles indicate anomalous" (i.e., high mean) January temperature sites containing a blue oak.

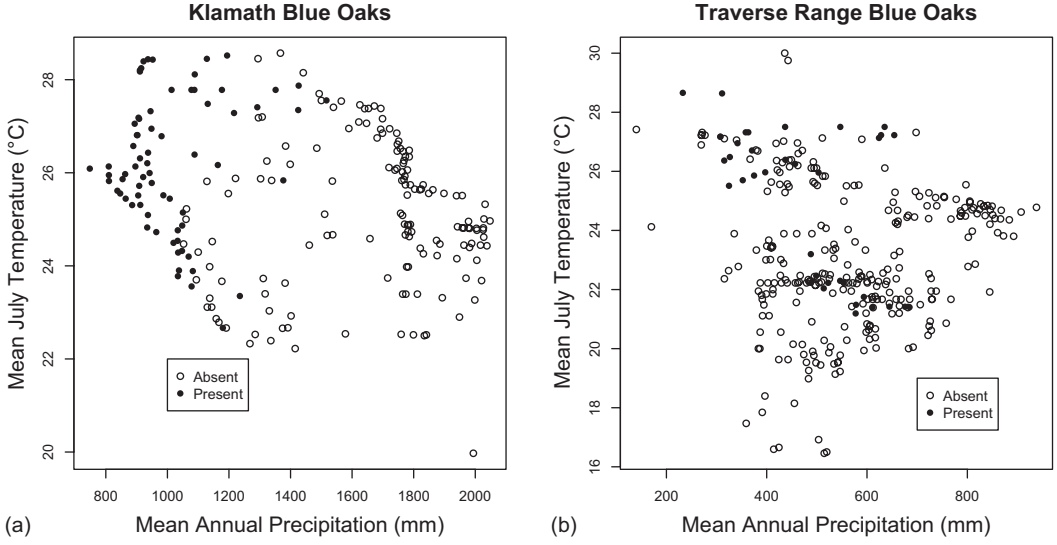

FIGURE 17.5
Scatterplots of mean July temperature vs. mean annual precipitation for (a) the Klamath Range and (b) the Transverse Ranges. The filled circles indicate locations where a blue oak is present.

We turn now to the regional data. The mixed model results in Section 12.6 identified soil permeability *Permeab*, length of growing season over 32°F *GS32*, and potential evapotranspiration *PE* as the principle explanatory variables in the Coast Range, and mean annual temperature *MAT*, mean annual precipitation *Precip*, and average available water capacity *AWCAvg* as the principle explanatory variables in the Sierra Nevada. Let's construct scatterplots of some of the regional data to see if we can get a handle on what the data are trying to say. Figure 17.6a shows a scatterplot of *QUDO* vs. *Permeab* for the Sierra Nevada and the Coast Range. The latter data set shows a boundary line effect with the maximum fraction of oak-occupied sites declining as the soil permeability increases. The Sierra Nevada only displays this effect to a much lesser extent. It is possible that this is a real effect, and that permeability is interacting with some other variable that differs between the Coast Range and the Sierra Nevada. It is also possible that permeability is serving as a surrogate for some other variable. The logical suspect is, of course, precipitation. There is indeed a positive correlation between *Permeab* and *Precip* ($r = 0.59$). Figure 17.6b shows a plot of *QUDO* vs. *Precip* for the regional data. There is again a negative relationship, but something seems to be interacting with *Precip*. Once again, we think of a logical suspect: something related to temperature. Playing around with scatterplots indicates that *JuMean* is a good possibility.

In summary, the evidence for the influence of precipitation is incontrovertible. The high level of autocorrelation among the variables makes further conclusions more or less speculative, but it appears that temperature, possibly either summer high temperatures or winter low temperatures, also limit the range of blue oaks. The xerophilic nature of blue oaks lends ecophysiological credence to the idea that the trees prefer hot summers, but regions that have hot summers tend to have cool winters, so it is hard to tease apart these effects with data. There are indications, both in the pointwise data and in the regional data, of an interaction between precipitation and temperature effects. The correlations between the data make it difficult to positively identify the effect of soil permeability.

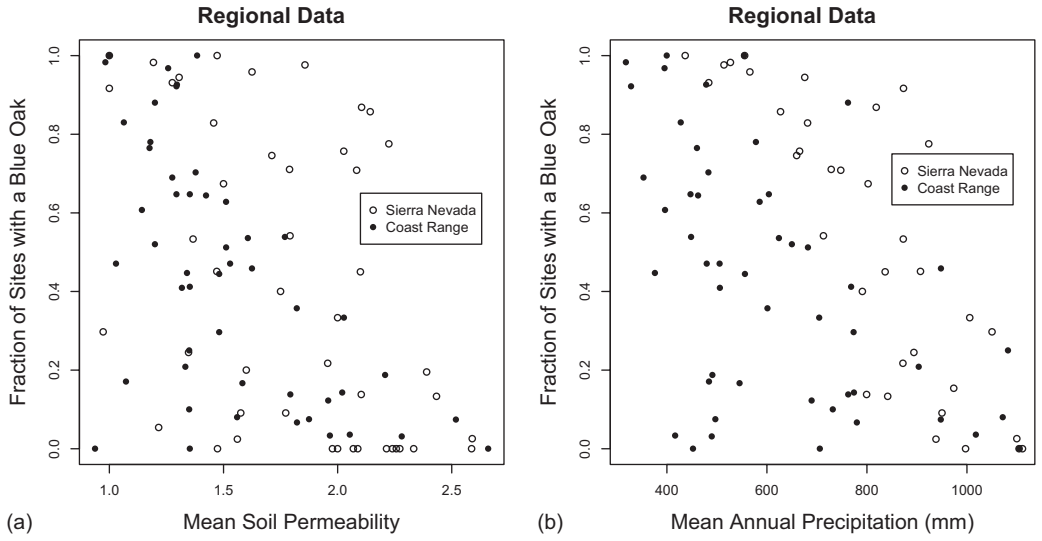

FIGURE 17.6
Plots of the fraction of sites in the regional data having a blue oak vs. (a) permeability and (b) mean annual precipitation for the Sierra Nevada and the Coast Range.

The initial exploratory analysis in Chapter 7 indicated that blue oak prevalence declined with increasing permeability, in opposition to the generally accepted properties of blue oak. Permeability and precipitation are highly correlated, however, and this may confound the results of the analysis.

17.4 Data Set 3

There is a wide range of values of yields among the rice farmers who participated in the study that led to Data Set 3. The ultimate goal in the analysis of the data set was to determine the management actions that those farmers take that would provide the greatest increase in their yield. No assumptions were made about either the economic value of the yield or the cost to the farmer of the management actions. As with Data Set 2, this data set and objective provides us with the opportunity to study a data analysis problem that involves multiple spatial scales. Two important early findings were that there was no significant effect of the season (i.e., the year of the study), and no significant rice year effect (i.e., of the position in the crop rotation). Based on these findings, data were pooled for the analysis. Much of the data analysis, especially the recursive partitioning in Chapter 9, examined the data at the landscape level, identifying the management variables that most influenced yield over this extent. The improvement in yield, however, must take place at the field level. In between these, there are differences in patterns at the regional level.

The landscape level recursive partitioning analysis in Section 9.3.4 identified the management variables *N*, *Irrig*, and *DPL* as playing an important role in the tree relating yield to management variables. When both management and exogenous variables are included in the model, nitrogen rate *N* and soil clay content *Clay* are identified as most important. An alternative approach was to identify the exogenous and management factors that most distinguished each farmer. The primary distinguishing characteristics of the farmers obtaining the highest yields in their respective regions were *DPL*, *Irrig*, and *N*.

The potential confounding factor is the quality of the land on which each farmer grows rice. Farmers with worse soil conditions might not improve their yield by mimicking the practices of farmers who get higher yield on better soil. Therefore, we attempted to determine the extent to which climatic or environmental factors played a role in determining the differences in rice yield observed between farmers. In other areas, clay-related quantities such as cation exchange capacity provide the best predictors of rice yield (Casanova et al., 1999; Williams, 2010, p. 5). Most of our analyses supported the conclusion that for the fields in our study, management variables played a more important role than field conditions (i.e., exogenous variables). The primary distinguishing characteristic at the landscape level between the lower yielding central region fields and the higher yielding northern and southern region fields is the high silt content in the central region. Figure 17.7 shows a soil texture triangle (Singer and Munns, 1996, p. 24) made with the function `soil.texture()` from the `plotrix` package (Lemon, 2006). Within region, however, there is no indication that silt content influences yield (Figure 7.20). This was cited in Section 11.5.2 as an example of the ecological fallacy. Normally, coarser textured soils require a higher level of applied nitrogen than high clay soils (Williams, 2010, p. 34), but farmers in the central region tended to apply fertilizer *N* at lower rates than farmers in the north and south.

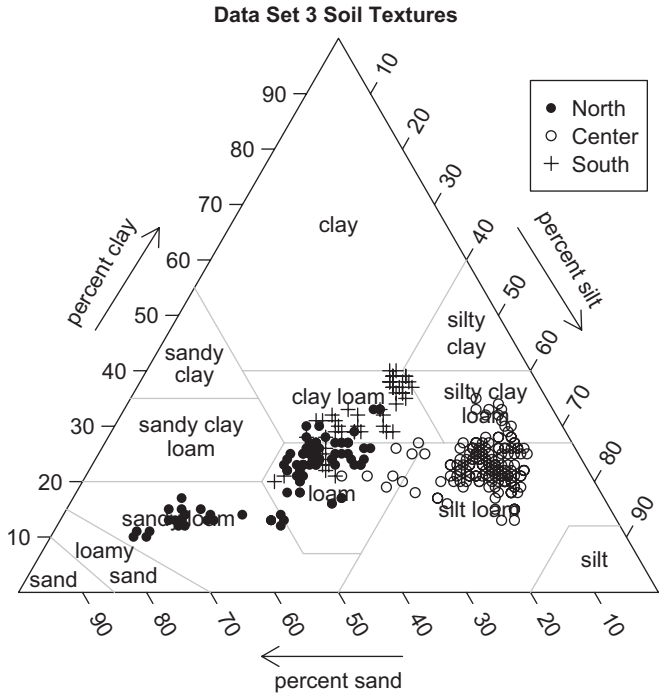

FIGURE 17.7
Soil texture triangle showing the soil textures of the fields in Data Set 3.

Our final analysis will try to predict the effect on yield that would occur if the farmers in the central region adopted the management actions of the northern and southern farmers. Here we must again be careful of running afoul of the ecological fallacy. The recursive partitioning analyses indicate that *DPL*, *N*, and *Irrig* are the most important management variables distinguishing high yield farmers from low-yield farmers. This conclusion is based on a landscape level analysis, but if we are going to compare farmers and regions, then we must accept this limitation. Gelman and Hill (2007, p. 252) point out that, just as the individual level model tends to understate regional variation, the regional means model tends to overstate it. We will try as much as possible, however, to predict the effect of modifications in farmer management practices using field scale analyses.

Our strategy will be the following. For each of the inputs *N*, *Irrig*, and *DPL*, we will carry out a regression analysis similar to that of Section 12.3 and determine the regression coefficients on a field by field basis for those fields at which the input was measured at more than one rate. We will then use these regression coefficients to predict the yield that would be achieved in each field in the central region if the farmer applied the input (*N*, *Irrig*, or *DPL*) at the rate deemed optimal by the analysis. Because of the scarcity of the data, we will not take spatial autocorrelation into account in the model. Because we are interested only in the specific fields, we will treat both the input and the field as a fixed effect, and therefore we will use ordinary least squares.

For nitrogen fertilization rate *N*, only fields 5, 11, 12, 13, and 16 contain measurements at different rates. We will use the function `lmList()` from the `nlme` package (Pinhiero et al., 2011) to obtain the regression coefficients.

```
> data.Set3N <- data.Set3[which(data.Set3$Field %in%
+     c(5,11,12,13,16)),]
> data.lis <- lmList(Yield ~ N | Field, data = data.Set3N)
> print(coef(data.lis), digits = 3)
   (Intercept)     N
5         37.5 126.1
11      -323.7 104.4
12       662.9 106.8
13      1351.1  86.6
16      3759.4  56.2
> b0.N <- coef(data.lis)[,1]
> b1.N <- coef(data.lis)[,2]
```

Now we compute the current yield and the predicted change in yield if *N* is applied at the maximal rate.

```
> print(Nmax <- max(data.Set3$N))
[1] 62.8
> print(Yield.predN <- b0.N + b1.N * Nmax, digits = 4)
[1] 7959 6233 7368 6789 7288
> print(Yield. N <- tapply(data.Set3N$Yield, data.Set3N$Field, mean),
+    digits = 4)
   5   11   12   13   16
6193 4771 6461 6053 7064
> print(delta. Y <- Yield.predN - Yield. N, digits = 3)
   5   11   12   13   16
1766 1462  908  736  224
```

We can carry out a similar analysis for *Irrig* and *DPL*. The code is analogous to that for *N*, and is not displayed. Here are the corresponding changes in yield.

```
> print(delta. YI <- Yield.predI - Yield. I, digits = 3)
      1       3       4       5       6       7       8       9      10
 154.20 1308.42  511.13 2873.80   -4.67  844.17 2811.60 2275.33 -448.35
     11      12      13      14      15      16
1363.78 1083.96 1065.00  395.56 -216.50    9.90
> print(delta. YD <- Yield.predD - Yield. D, digits = 3)
    3     4     5    10    11    12    16
 -254  3419  3953  2827  5011 -2500   494
```

It is difficult to draw general conclusions, especially given the high variability in the predicted response to planting date. If we assume that there are no scheduling constraints on the growers that would prevent them from moving up their planting dates, then it would appear that this would be the lowest cost change that would improve their yield. Unless the cost of fertilizer becomes very high, increased nitrogen rate should also be economically justified. Given the high cost associated with land moving, changes in irrigation effectiveness within the field appear to be the least justified.

17.5 Data Set 4

The two fields that make up Data Set 4 were each sampled on a grid in 1995 at the start of the study, and yield data were collected in each of four growing seasons after that. Both fields were planted to a crop rotation that began with winter wheat in the first season and coursed to tomatoes in the second season. Field 4.1 was planted to beans and sunflower in years three and four, while Field 4.2 was planted to sunflower and corn (maize) during those years. The southern part of Field 4.1 (which is, alas, probably the most interesting part) was removed from the study after the second year. Both fields are laser leveled, so, although the topography prior to leveling almost surely influenced the current soil patterns, the current topography has no influence. The fields are fully irrigated, and there is virtually no rainfall (or even clouds) in the summer (cf. Figure 15.14b), so the last three years were not influenced by rainfall patterns. Temperature also does not appear to play much of a role, as the years were very consistent (Figure 15.14a).

The primary objective is to develop a methodology for determining the factors that influence spatial patterns of yield. There is no implication that the same patterns of influence will be present in other fields, but one might hope that the methodology can be applied more broadly. Field 4.2 is only about two kilometers away from Field 4.1, and is farmed by the same person, so that we might also hope that the scope of inference can be extended to Field 4.2. Although some of the exercises have involved graphical exploration and parameter estimation of Field 4.2, none of the model development has used information from this field.

Our model construction was founded on the assumption, based on several lines of evidence, that soil texture plays an important role in influencing yield in Field 4.1. Figure 17.8 shows the clay content over the field. There is one data record that is probably an outlier. Because of their high level of association with other variables, *Sand* and *SoilTOC* were not used in the analysis. *SoilK* and *SoilTN* are also highly associated (Figure 7.25), and both are associated to some degree with *SoilP*, with the association being stronger in the southern portion of the field (Exercise 8.3). There was no indication that any part of the field suffered from a nitrogen or potassium deficiency, but the northern and southern parts of the field had soil phosphorous levels below those recommended by University of California guidelines for wheat production (Figure 7.28a).

In Section 8.3, we developed several candidate models based on ordinary least squares regression. The linear regression models developed in Section 8.3 and later analyzed in Chapter 13 are the following:

```
> model.1 = lm(Yield ~ Clay + Silt + SoilpH + SoilTN + SoilK +
+      SoilP + Disease + Weeds + I(Clay*SoilP) + I(Clay*SoilK) +
+      I(Clay*SoilpH) + I(Clay*SoilTN), data = data.Set4.1)
> model.2 <- lm(Yield ~ Clay + Silt + SoilTN +
+      Weeds + I(Clay*SoilTN), data = data.Set4.1)
> model.3 <- lm(Yield ~ Clay + SoilpH + SoilP + Weeds +
+      I(Clay*SoilP) + I(Clay*SoilpH), data = data.Set4.1)
> model.4 <- lm(Yield ~ Clay + SoilpH + SoilP +
+      Weeds + I(Clay*SoilP) + Weeds, data = data.Set4.1)
> model.5 <- lm(Yield ~ Clay + SoilP + I(Clay*SoilP) +
+     Weeds, data = data.Set4.1)
```

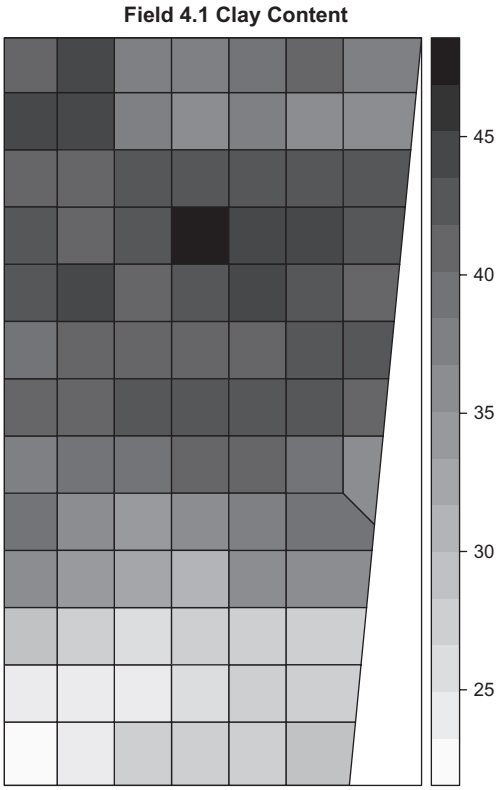

FIGURE 17.8
Thematic map of clay content in Field 4.1.

The first model, `model.1`, includes everything, the second model, `model.2`, was selected by backward selection, and the last model, `model.5`, is the model picked out by Mallow's C_p analysis. A real study would also have included exploration via forward and bidirectional selection. None of the models were strongly affected (i.e., there was no change in the significance level of any of the coefficients) by the inclusion of spatial autocorrelation in the model (Exercise 13.3a). These models identify clay content, soil pH, soil P, weeds, and soil TN as potentially important contributors to yield. Of these, the negative influence of weeds in certain parts of the field is easiest to distinguish. There is an interaction indicated between clay content and soil P, as well as possibly soil pH. This is also evident in the scatterplot matrices of the two regions of the field. Although inclusion of an interaction between *SoilP* and *SoilpH* would be biochemically justifiable, there is no evidence that such a term would be significant in a regression model (Exercise 8.6a).

The classification tree analysis in Chapter 15 of the spatiotemporal data clusters indicated that clay content was important throughout the four years (Figure 15.15a). The areas of higher clay content, coupled with either low soil P or high weed levels in 1996, had the lowest yield. Yields in the lower and higher yielding area over the four years did, however, tend to become more uniform (Figure 15.2), and by the fourth year, soil P level had apparently ceased to become as dominant in determining yield. This will be addressed below. There was some indication that soil pH and soil K content were associated with yield in the high clay region in 1999, with areas of lower pH and higher K being associated with a higher yield.

The evidence points to a tentative conclusion that clay content is a dominant factor in yield, that soil phosphorous content and possibly soil pH also play a role, and that weed level is important in certain areas and years, particularly in the first year along the eastern boundary of the field and in a small region along the south central part of the western boundary. In the northern part of the field, where clay content is high, aeration stress presumably limits yield. The evidence indicates that soil phosphorous level is also limiting in the far north. In the southern part of the field, the negative association between yield and soil phosphorous indicates that the crop is mining this immobile nutrient. In the southwestern corner, where the soil acidity is closer to neutral, presumably increasing P availability, the yield is highest. There is some indication that disease and either soil potassium level or soil nitrogen level or both may play a role in limiting yield, but the evidence is not conclusive. In this context, it is important to recall a comment that was made in Chapter 9 regarding the difference between regression methods and partitioning methods. This is that regression methods summarize process over the geographical extent of the region being studied, while partitioning methods as the partitioning advances increasingly restrict the analysis to separate geographical areas.

Now let's examine the data from Field 4.2 and see how it relates to these conclusions. Although the two fields are similar in location, management, and presumably properties, it would be naïve to simply plug the data from Field 4.2 into the model and see what comes out. First, we must compare the fields a bit to see how their response to the various yield limiting factors might differ. We begin with the soil. Throughout the analyses of the previous chapters, we have found evidence that soil texture plays an important role in Field 4.1. What is the role of texture in Field 4.2? The maximum and minimum values of *Clay* in the two fields are almost identical.

```
> range(data.Set4.1$Clay)
[1] 23.20 46.91
> range(data.Set4.2$Clay)
[1] 23.2 46.1
```

The distributions, however, are very different.

```
> stem(data.Set4.1$Clay)
  22 | 279
  24 | 68
  26 | 559934578
  28 | 0338
  30 | 5
  32 | 2
  34 | 483456
  36 | 245670011
  38 | 1145678935
  40 | 1225567801244555999
  42 | 44456677790355678
  44 | 0135
  46 | 9
```

```
> stem(data.Set4.2$Clay)
  22 | 2
  24 | 016
  26 | 06
  28 | 50144
  30 | 0227883344555567778 99
  32 | 11238813679
  34 | 13338804557
  36 | 03552689
  38 | 4689911258
  40 | 4493
  42 | 9
  44 |
  46 | 1
```

The *Clay* distribution in Field 4.1 is bimodal, whereas in Field 4.2 it is unimodal and positively skewed. Also, wheat yield in the 1996 harvest is much lower in the high clay area of Field 4.1 than it is in Field 4.2.

There is also a considerable difference in the range of *SoilP* levels between the fields.

```
> range(data.Set4.1$SoilP)
[1] 2.48 18.14
> range(data.Set4.2$SoilP)
[1] 13.5 52.3
```

There is nowhere in Field 4.2 even remotely close to the P application threshold of 6 ppm. The one variable among those in which we are interested with similar distributions between the two fields is *SoilpH*.

```
> stem(data.Set4.1$SoilpH)
  The decimal point is 1 digit(s)
to the left of the |
  55 | 6
  56 | 001
  56 | 555889
  57 | 000013344444
  57 | 5555556777788999999
  58 | 00222
  58 | 44444677888899
  59 | 11112
  59 | 5555667888889
  60 | 0223
  60 | 499
  61 | 2
```

```
stem(data.Set4.2$SoilpH)
  The decimal point is 1 digit(s)
to the left of the |
  54 | 4
  54 | 9
  55 | 000111234444
  55 | 5567777888899
  56 | 00000012222222333334444
  56 | 55566788889
  57 | 111111234
  57 | 55689
  58 | 1
  58 | 7
  59 | 2
```

Field 4.2 is somewhat more acidic than Field 4.1.

During the actual study, standing water was observed in the north end of Field 4.1, which is a clear indication that the high level of association of clay content with yield was due to aeration stress, which was caused by poor drainage. There is only a very small isolated area of high clay content in Field 4.2, and the entire field is composed of well-drained soils (Andrews, 1972), so aeration stress should not be not a problem. Similarly, the high soil P levels in Field 4.2 likely preclude a relationship of yield with soil P. Therefore, we must conclude that the model of yield response to explanatory variables developed with the data of Field 4.1 cannot be extended to Field 4.2. Unfortunately, we were not privy to all of the cooperator's management decisions over the years, so we cannot tell whether the increased uniformity of yield is due to changed management practices. It may have also been due to the slightly higher temperatures in later years, as well as to the greater control that the farmer was able to exert over summer crops. A classification tree for the two clusters in Field 4.2 splits first (unsurprisingly) on *Weeds* ≥ 4.5 and then (surprisingly) on *SoilpH* $<= 5.665$. This is surprising because the more neutral soils are in the lower yielding cluster. Acidic soils are more commonly associated with reduced wheat yields. The area of reduced wheat yield associated with low soil pH is at the eastern end of the field. Unlike Field 4.1, it is unlikely that there is any substantial effect of acidity on P availability because P levels are so high. A scatterplot of wheat yield against soil pH reveals no trend, so it is likely that the association of these two quantities identified by the classification tree is spurious. There is definitely a pattern of lower wheat yields in the eastern end of the field, but this is not associated with anything measured in our study.

In the actual event, after he saw our first year's results, the cooperating farmer performed his own phosphate trial in the northern part of Field 4.1. He found that there was indeed a substantial yield response to added phosphate fertilizer. Increased P application in later years may have contributed to the increased uniformity. In that sense, the methodology worked as it was supposed to in Field 4.1. Field 4.2 turned out to be fairly uniform in its properties and is probably one of those fields for which the null hypothesis of precision agriculture (Whelan and McBratney, 2000) is valid, and uniform application of inputs is appropriate in this field.

17.6 Conclusions

The most important general conclusion that can be drawn from this text is that one should never rely on a single method to analyze any data set, spatial or otherwise. For each of the data sets we studied, every method applied to that data set provided some information, and no one method was sufficient by itself to provide all the information we needed to reach a conclusion. Data analysis is a nonlinear process. This is often not apparent in the publication of the results, and indeed the linear order of the chapters of this book may serve to reinforce the incorrect idea that data analysis proceeds from sampling to exploration to confirmation to conclusions. This is not the way a data set should be analyzed. Exploration should motivate modeling, and modeling should motivate further exploration. This is why you have to live with a data set for a long time to really come to know it.

Data from an observational study can never be relied on by themselves to produce a model of cause and effect. The best that can be accomplished with such data is to produce alternatives that can be parsed using biophysical arguments or controlled experiments. Sometimes, however, observational studies provide the only data that can be obtained. In the case of the Data Set 1, the yellow-billed cuckoo is an endangered species with a very small remaining habitat, and so it is doubtful that any replicated experiment could be undertaken that would contribute much to understanding the bird's behavior. In the case of Data Set 2, there have been numerous, highly informative replicated experiments that helped to understand blue oak seedling and sapling development. The distribution of mature blue oaks is more difficult to study in this way because trees of this species are extremely slow growing, and can remain in the seedling or sapling stage for decades before maturing (Phillips et al., 2007). This is a good example of a problem in which experimental data and observational study data must be combined to extend the scope and interpretation of the experiments.

Chamberlain (2008), in an editor's message that every ecologist (and every reviewer of ecological papers) should read, exhorts scientists not to "sacrifice biology on the altar of statistics." A research project always starts with a scientific question. Data are analyzed in a search for patterns that may help to provide an answer. One question (but not the only question) to ask about these patterns is how likely they are to have appeared by chance. This question is addressed via a probabilistic model. One must always keep in mind, however, that the most important objective is *not* to estimate the probability that the pattern in the data occurred by chance. It is to answer the scientific question that motivated the data collection in the first place. The problem with the use of a probabilistic model is that it lends an aura of credibility to the result. If the data do not satisfy the assumptions of the model, then this credibility may not be deserved. A more mathematically sophisticated

model provides a greater aura of credibility and therefore a greater risk if its credibility is undeserved. A simpler model, on the other hand, is easier to understand and use but often has more restrictive underlying assumptions that are likely to be incorrect. For this reason, in selecting the most appropriate statistical model, the data analyst must balance the advantages of the simpler model against those of the more complex one. To do this correctly requires an understanding of the theory behind both models. In the words of Benjamin Bolker (2008, p. 220), "you have to learn all the rules, but also know when to bend them."