11

Measures of Bivariate Association between Two Spatial Variables

11.1 Introduction

The calculation of measures of bivariate association among attribute values is often one of the first steps taken following the initial exploratory analysis of a data set. In this sense, measures of bivariate association form a sort of bridge between the exploratory and confirmatory stages of the analysis. When dealing with continuously valued data, two of the most commonly used measures of bivariate association are the Pearson product moment correlation coefficient r and the Spearman rank correlation coefficient r_s. These statistics measure *linear* association between two quantities. The exploratory phase of the data analysis should, however, have established a general working hypothesis about the form of the relationships between variables, and the computation of the correlation coefficient should be carried out if the results of this exploratory analysis indicate that a linear association is possible.

Consider the case of the wheat data from Field 1 of Data Set 4. The scatterplot matrix of Figure 11.1a indicates that there appears to be a more or less linear relationship between the variables *SPAD* and *LeafN*. *SPAD* refers to the reading on a Minolta SPAD® meter, which is an optical device that estimates the chlorophyll content of a leaf based on the relationship between its transparency to electromagnetic radiation in the near-infrared and in the red regions (Markwell et al., 1995). *LeafN* is the total leaf nitrogen concentration. The SPAD reading has been proposed as a measure of *LeafN* content (Bullock and Anderson, 1998), so one would expect that a relationship between these two quantities would exist. One might expect in turn that leaf nitrogen content would be related to grain protein content, which is a measure of grain quality. Figure 11.1b appears to indicate a relationship between grain protein content and SPAD meter reading, although there is a lot of variability. If this relationship were demonstrated to be significant and consistent, it would provide a relatively low-cost method to estimate the spatial variability of grain quality, which is an economically important quantity. Thus, it is of interest to study the relationship between SPAD reading and grain protein content.

When dealing with categorical data, the most common descriptor of association is the contingency table (Larsen and Marx, 1986, p. 442; Agresti, 1996, 2002). As a very simple example, we can consider the distribution of oak species in Data Set 2 as a function of elevation. According to Pavlik et al. (1991), the blue oak (*Quercus douglasii*) is found primarily at elevations less than about 1100 m. Using a similar approach to that taken in Section 7.3, we can construct a contingency table of the relationship between the presence or absence of blue oaks at a site and the relationship of the elevation of the site to 1100 m.

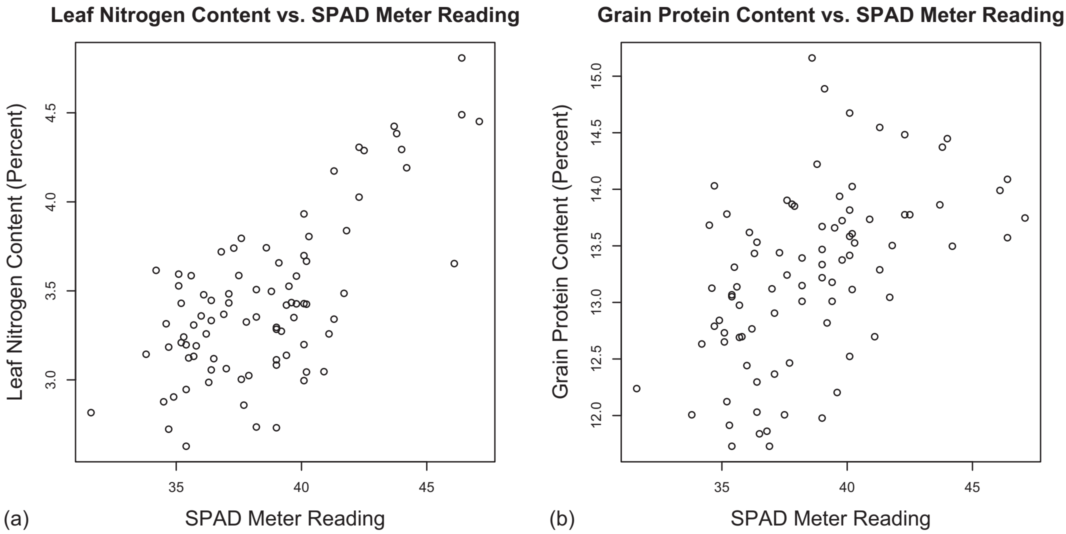

FIGURE 11.1
(a) Scatterplot of leaf nitrogen concentration vs. SPAD ® reading for the data of Field 4.1; (b) Scatterplot of SPAD ® reading vs. grain protein content.

```
> with(data.Set2,
+    matrix(c(sum(Elevation <= 1100 & QUDO == 1),
+    sum(Elevation <= 1100 & QUDO == 0),
+    sum(Elevation > 1100 & QUDO == 1),
+    sum(Elevation > 1100 & QUDO == 0)), nrow = 2, byrow = TRUE,
+    dimnames = list(c("Low", "High"), c("Pres", "Abs"))))
     Pres  Abs
Low  1609 1776
High   37  679
```

It is evident that blue oaks do indeed prefer elevations below 1100 m, but they are not ubiquitous at these elevations. There are about the same number of sites below 1100 m with and without blue oaks (recall that all sites in Data Set 2 contain some species of oak).

One does not need much of a statistical analysis to conclude that the fraction of sites at which a blue oak is present is higher below 1100 m than it is above that elevation. With some contingency tables, however, one may need to test a null hypothesis of randomness in order to be able to conclude that there is a pattern in the data. One way to accomplish this is to use a chi squared test, which is asymptotically valid as the sample size increases. A second method, asymptotically equivalent to the chi squared test, involves the maximum likelihood estimator. A third test, the Fisher test, is used in cases where the sample size is not large enough for the chi squared or likelihood tests to be approximately valid.

When one evaluates a contingency table or a statistic such as the correlation coefficient for spatial data, one faces the same sort of issue that arises with the hypothesis tests that were discussed in Chapter 10: the existence of spatial autocorrelation may reduce the effective sample size, inflating the apparent significance level determined by the test. The first two sections of this chapter discuss methods for addressing this problem. Section 11.2 addresses continuously varying data, and Section 11.3 addresses categorical data. There are other issues that arise in the evaluation of measures of bivariate association, however. One of these can be understood by considering again the relationship between SPAD reading and grain protein level. In trying to understand this relationship, one could simply compute the correlation coefficient between the former and the latter. It is reasonable to suppose, however, that SPAD reading is more directly related to leaf nitrogen level, and leaf nitrogen level to grain protein level, than SPAD is directly to grain protein. Nevertheless, it is grain protein and not leaf nitrogen that provides the farmer with economic yield, so this is the relationship we would like to understand. There may be other factors that strongly influence this relationship. The partial Mantel test, which is described in Section 11.4, provides one way to determine whether a third factor influences the behavior of the associated quantities.

The data describing the relationship between SPAD meter reading and leaf nitrogen level come from 86 sample points arranged 61 m apart in a grid in the field. In dealing with the relationship between, for example, NDVI and yield as measured by a yield monitor, there are thousands of measurements that are either contiguous or very close to each other. These data must be aggregated before they can be compared. The *scale*, as defined in Chapter 6, of this aggregation is not naturally imposed, but rather must be selected. It turns out that the scale at which the data are represented may play a major role in the statistical properties of these data, particularly those that describe the relation between two quantities. This phenomenon is called the *modifiable areal unit problem* (MAUP). Furthermore, the relationship between two quantities when they are aggregated may be different from that between these same quantities when compared as individuals. Using aggregated data to form a conclusion about a relationship between individuals is called the *ecological fallacy*. The MAUP and the ecological fallacy are very closely related, and indeed Cressie (1997) considers them to be manifestations of the same phenomenon, the “change of support problem” discussed in Section 6.4.3. Conceptually they are different, however, in the following sense. The MAUP deals with ecological systems for which there may be no meaningful geographical subdivision, such as an agricultural field. The ecological fallacy concerns systems where the concepts of an individual and a collection of individuals make sense. We will therefore treat the problems as distinct phenomena. They are discussed in Section 11.5.

Finally, but perhaps most importantly, there is one problem that does not necessarily have a statistical solution. This is the temptation to confuse association with causality. Although this problem arises in dealing with any data set, it is particularly pervasive in dealing with spatial data, where similar patterns in thematic maps can be particularly compelling. There is a well-known and easily conceptualized analogy in time series, where the term *nonsense correlation* was used by Yule (1926) to describe the phenomenon. A good example is the high correlation observed between the stork population and the human birthrate in the German city of Oldenberg (Box et al., 1978, p. 8). The stork population and human birthrate both increased due a growth in the human population and an associated increase in the construction of buildings. In the same way, in data arising from an observational study, spatial variables can exhibit a spurious correlation due to a common response to a third factor that has a geographic trend. Perhaps the best way to avoid this pitfall is to constantly be on the lookout for it.

11.2 Estimating and Testing the Correlation Coefficient

11.2.1 The Correlation Coefficient

The most common statistic measuring the linear association between two variables is the Pearson product moment correlation coefficient r. Let the two variables be denoted Y_1 and Y_2. The population correlation coefficient ρ is defined as (Larsen and Marx, 1986, p. 435)

$$\rho = \frac{\operatorname{cov}\{Y_1, Y_2\}}{\sigma_1 \sigma_2}, \tag{11.1}$$

where σ_1^2 is the population variance of Y_1 and $\sigma_1 = \sqrt{\sigma_1^2}$ is the standard deviation, and similarly for σ_2. An estimator of ρ is Pearson product moment correlation coefficient r, defined as (Larsen and Marx, 1986, p. 453; Wackerly et al., 2002, p. 568)

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n} (Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\sqrt{\sum_{i=1}^{n} (Y_{i1} - \bar{Y}_1)^2 \sum_{i=1}^{n} (Y_{i2} - \bar{Y}_2)^2}} \\ &= \frac{\operatorname{cov}\{Y_1, Y_2\}}{s_1 s_2}. \end{aligned} \tag{11.2}$$

where s_1 and s_1 are defined implicitly by the second equation as $s_1 = \sqrt{\Sigma(Y_{i1} - \bar{Y}_1)^2}$ and similarly for s_2. In this section, we use the symbols Y_1 and Y_2 rather than X and Y to represent the variables to emphasize the fact that we are only testing association. There is no implication of a direction of influence. A computational formula for r that will come in handy in Section 11.4 is (Larsen and Marx, 1986, p. 437)

$$r = \frac{n\sum_{i}^{n} Y_{i1}Y_{i2} - n^2\bar{Y}_1\bar{Y}_2}{s_1 s_2}. \tag{11.3}$$

A value of r can be computed for any pair of random variables, but statistical tests on r are only valid if the probability distribution of the random pair (Y_1, Y_2) is *bivariate normal*. The bivariate normal distribution has five parameters, μ_1, μ_2, σ_1^2, σ_2^2, and ρ.

We will work with the three data fields from Field 1 of Data Set 4 discussed in Section 11.1: SPAD meter reading, leaf nitrogen content, and grain protein content. These data can be used to address two questions, namely whether SPAD meter reading is linearly related to *LeafN* content, and whether meter reading is linearly related to grain protein content. As stated in Section 11.1, scatterplots of the relationship between these variables appear to indicate a positive linear association (Figure 11.1), but, particularly in the case of the relationship between grain protein content and SPAD meter reading, the significance of this relationship requires testing.

To formulate the problem in a general way, let $Y_1(x_i, y_i)$ and $Y_2(x_i, y_i)$ be two quantities measured at a set of locations with coordinates (x_i, y_i), $i = 1, \ldots, n$. If, conditional on Y_1, the values of Y_2 are independent and identically distributed, then the variance of the sampling distribution of the random variable r under the null hypothesis $r = 0$ is

$$\sigma_r^2 = 1/(n-1) \tag{11.4}$$

(Kendall and Stuart, 1979). If Y_1 and Y_2 are normally distributed, the null hypothesis $r = 0$ can be tested using the statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \tag{11.5}$$

which has a t distribution with $n-2$ degrees of freedom under the null hypothesis (Wackerly et al., 2002, p. 568). A preliminary visual inspection of Figure 11.1a casts some doubt on the issue of whether the quantities in the scatterplot are normally distributed (they appear skewed low). In case they are not, the question of a linear relationship between them can be addressed using the Spearman rank correlation coefficient r_s, which is computed by replacing the values for Y_1 and Y_2 in Equation 11.1 with their respective ranks. The statistic r_s does not follow a t distribution under the null hypothesis $r_s = 0$, but corresponding acceptance regions can be either computed directly or tested via a permutation test (Section 3.3).

The first question we address is whether the effect of positive spatial autocorrelation on a significance test of the correlation coefficient is to make the test more "anticonservative," that is, to artificially inflate the Type I error rate. This question can be addressed through Monte Carlo simulation in the same manner as was done in Section 3.4. To set up the simulation, we first run a single test. We create a 14 by 14 square grid of point pairs. Unfortunately, the same symbol ρ is traditionally used for both the correlation coefficient and the autocorrelation term, and this tradition is preserved in the R functions we will be using in this section, so we will follow it. The meaning will generally be clear from the context. Here we set the autocorrelation term ρ to 0.6 and generate the data.

```
> library(spdep)
> rho <- 0.6
> nlist <- cell2nb(14, 14)
> IrWinv <- invIrM(nlist, rho)
> set.seed(123)
> Y1 <- IrWinv %*% rnorm(14^2)
> Y2 <- IrWinv %*% rnorm(14^2)
```

Next, we eliminate the two outer cell layers to reduce edge effects.

```
> Y1samp <- matrix(Y1, nrow = 14, byrow = TRUE)[3:12,3:12]
> Y2samp <- matrix(Y2, nrow = 14, byrow = TRUE)[3:12,3:12]
```

Now we use the R function `cor.test()` to test the null hypothesis of zero correlation between Y_1 and Y_2, which we have constructed to be uncorrelated.

```
> cortest <- cor.test(Y1samp, Y2samp,
+    alternative = "two.sided", method = "pearson")
> print(r <- cortest$estimate, digits = 3)
  cor
0.205
```

```
> print(t.stat <- cortest$statistic, digits = 3)
   t
2.07
> print(p <- cortest$p.value, digits = 3)
[1] 0.0407
```

Setting the argument method in cor.test() to "pearson" causes the test to be carried out using the t statistic given in Equation 11.5. Assuming a significance level of $\alpha = 0.05$, a Type I error (rejecting the null hypothesis when it is true) occurs in this particular test. Table 11.1 contains the experimental Type I error rate generated via Monte Carlo simulation as the autocorrelation parameter ρ is increased (the table also contains similar results for two methods of correcting for spatial autocorrelation that are discussed below). As with the t test of Chapter 3, the Type I error rate of the test increases with increasing ρ, although it is moderate for small values of ρ. The next section describes a widely used method of correcting for spatial autocorrelation.

11.2.2 The Clifford et al. (1989) Correction

Clifford and Richardson (1985) and Clifford et al. (1989) have developed a correction method that reduces the effect of positive spatial autocorrelation on the outcome of the significance test. Recall (Section 3.4) that the intuitive explanation for the increased Type I error rate is that if the values of $Y_1(x,y)$ and $Y_2(x,y)$ are spatially positively autocorrelated, then each value provides some information about the values of its spatial neighbors, and therefore the actual number of degrees of freedom provided by the set of observations is less than the sample size n. Equation 10.8 is a formula for the estimated effective sample size $\hat{n}_e$ for the t test of the null hypothesis of the equality of two means. By analogy, an estimate $\hat{n}_e$ of the effective sample size n_e for the correlation coefficient could be obtained if one had an independent estimate $\hat{\sigma}_r^2$ of the variance of the sampling distribution of r. This estimate would be obtained by inverting Equation 11.4 to obtain

$$\hat{n}_e = 1 + \frac{1}{\hat{\sigma}_r^2}. \tag{11.6}$$

When $Y_1(x,y)$ and $Y_2(x,y)$ are positively autocorrelated the sampling distribution $\hat{\sigma}_r^2$ is generally larger than predicted under zero autocorrelation, so therefore $\hat{n}_e$ will generally be less than n in the case of positive spatial autocorrelation.

TABLE 11.1

Values of the Type I Error Rate for a Test of the Null Hypothesis $r = 0$ for the Uncorrected Test, the Correction of Clifford et al. (1989), and the Parametric Bootstrap Correction

ρ	0.0	0.2	0.4	0.6	0.8
Uncorrected	0.051	0.053	0.068	0.113	0.230
Clifford	0.052	0.052	0.057	0.073	0.111
Bootstrap	0.050	0.52	0.053	0.057	0.065

We can rewrite Equation 11.2 as

$$r = \frac{s_{12}}{s_1 s_2},$$
$$s_{12} = \frac{1}{n}\sum_{i=1}^{n}(Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2). \tag{11.7}$$

Using an ingenious Taylor series argument, Clifford et al. (1989) showed that under fairly general conditions the value of $\hat{\sigma}_r^2$ can be approximated by

$$\hat{\sigma}_r^2 \cong \frac{\text{var}\{s_{12}\}}{E\{s_1^2\}E\{s_2^2\}}. \tag{11.8}$$

They further showed that the numerator of this equation can be written as

$$\text{var}\{s_{12}\} = n^{-2}\sum_{i-1}^{n}\sum_{j=1}^{n}(Y_1(x_i, y_i) - \bar{Y}_1)(Y_2(x_j, y_j) - \bar{Y}_2)\text{cov}\{Y_1, Y_2\}. \tag{11.9}$$

Clifford et al. (1985, 1989) suggest using the covariogram (Section 4.6.2) to approximate the covariance term in this equation. Recall from that section (Equation 4.27) that the experimental covariogram of lag group k is given by

$$\hat{C}(k) = n_k^{-1}\sum_{i,j \in H(k)}(Y(x_i, y_i) - \bar{Y})(Y(x_j, y_j) - \bar{Y}), \tag{11.10}$$

where $H(k)$ is the k^{th} lag group and n_k is the cardinality (i.e., the number of pairs of data records) in $H(k)$.

Suppose there are a total of H_T lag groups, and that the covariance is isotropic. Of these H_T lag groups, the first H are used in the estimate of var$\{s_{12}\}$ in Equation 11.8. This implicitly assumes that the covariogram declines to zero as the lag distance k increases. Under these assumptions, Clifford et al. (1989) show that the variance var$\{s_{12}\}$ may be approximated by

$$\text{var}\{s_{12}\} \cong n^{-2}\sum_{h=1}^{H} n_k \hat{C}_1(k)\hat{C}_2(k). \tag{11.11}$$

where

$$\hat{C}_m(k) = \sum_{i,j \in H(k)_k} n_k^{-1}(Y_m(x_i, y_i) - \bar{Y}_m)(Y_m(x_j, y_j) - \bar{Y}_m) \tag{11.12}$$

is the experimental covariogram for Y_m, $m = 1, 2$. Clifford et al. (1985, 1989) suggest estimating the values of $E\{s_1^2\}$ and $E\{s_2^2\}$ in Equation 11.8 by the sample variances s_1^2 and s_2^2. Since the experimental correlogram $\hat{r}_1(k)$ of Y of the k^{th} lag group $H(k)$ is given by $\hat{r}_1(k) = \hat{C}_1(k) / s_1^2$, and similarly for $\hat{r}_2(k)$ (see Section 4.6.2), we may substitute Equation 11.11 into the numerator of Equation 11.8. Based on the definitions in Section 4.6.2 of the relationship between the covariogram and the correlogram, we can express Equation 11.8 in terms of the correlogram as

$$\hat{\sigma}_r^2 \cong n^{-2}\sum_{k=1}^{H} n_k \hat{r}_1(k)\hat{r}_2(k). \tag{11.13}$$

Since Equation 11.13 involves the product of the experimental correlograms for Y_1 and Y_2, if either Y_1 or Y_2 is spatially uncorrelated, then its correlogram at a positive lag is zero, and therefore the spatial autocorrelation of the other variable has no effect on the hypothesis test. Haining (1990, 1991) has extended the work of Clifford et al. (1989), testing the approximation under various combinations of values H_T and H of lag groups and also showing that the approximation may be used for Spearman's rank correlation coefficient as well as Pearson's r.

We can apply the correction of Clifford et al. (1989) to the hypothesis test carried out on the artificial data set in Section 11.2.1. First, we calculate the experimental correlograms $\hat{r}_1(k)$ and $\hat{r}_2(k)$ in Equation 11.13 using the function `correlogram()` of the R package `spatial` (Venables and Ripley, 2002). There are a number of packages that calculate the correlogram, but the function `correlogram()` of the `spatial` package is particularly useful for this application because it also returns the value of n_k in Equation 11.13. The calculation of the correlogram for Y_1 is accomplished by first using the function `surf.ls()` to express the data as a surface, and then passing this surface to the function `correlogram()`.

```
> library(spatial)
> coords.xy <- expand.grid(1:10,10:1)
> Y1.vec <- as.vector(t(Y1samp))
> Y1.surf <- surf.ls(0,coords.xy[,1], coords.xy[,2], Y1.vec)
> r1 <- correlogram(Y1.surf,20)
> Y2.vec <- as.vector(t(Y2samp))
> Y2.surf <- surf.ls(0,coords.xy[,1], coords.xy[,2], Y2.vec)
> r2 <- correlogram(Y2.surf,20)
```

The second argument of the function `correlogram()` specifies the number of lag groups, denoted H_T above. The object `r1` contains the correlogram values $\hat{r}_1(k)$ in Equation 11.13. The object `r2`, containing the values of $\hat{r}_2(k)$, is computed similarly. Next, we implement Equation 11.13 with $H = 10$ to compute $\hat{\sigma}_r^2$.

```
> nr1r2 <- r1$cnt * r1$y * r2$y
> print(sr.hat <- sum(nr1r2[1:10]) / 10^4, digits = 3)
[1] 0.0116
```

Finally, we compute $\hat{n}_e$ from Equation 11.6, substitute this for n in Equation 11.5 to compute a corrected t statistic t_{corr}, and then compute a p value, again using $\hat{n}_e$ to represent the number of degrees of freedom.

```
> print(ne.hat <- 1 + 1 / sr.hat, digits = 3)
[1] 87.2
> print(t.corr <- r * sqrt(ne.hat - 2) / sqrt(1 - r^2), digits = 3)
 cor
1.93
```

```
> print(p.corr <- 2 * (1 - pt(q = abs(t.corr),
+    df = ne.hat - 2)), digits = 3)
   cor
0.0564
```

In this particular case, the p value is increased from 0.040 to 0.056, and effective sample size is reduced from 100 to 87. The second row of Table 11.1 shows the results of a Monte Carlo simulation for increasing values of ρ.

11.2.3 The Bootstrap Variance Estimate

A possible alternative method for obtaining an estimate of σ_r^2 for use in Equation 11.6 is to compute a bootstrap estimate. Efron and Tibshirani (1993, p. 49) discuss the bootstrap estimate of the variance of the correlation coefficient for independent pairs of data values. This method can in principle be extended to autocorrelated data using the block bootstrap or parametric bootstrap in the same way that these methods were used to estimate the variance of the mean in Chapter 10. Here we will only discuss the parametric bootstrap. As with the test of the null hypothesis that the difference between two means is zero carried out in Chapter 10, one must fit a parametric model to the data. In comparing the means of two random variables $Y_1(x,y)$ and $Y_2(x,y)$ in Chapter 10, we set $Y = Y_1 - Y_2$ and applied the model

$$\begin{aligned} Y &= \mu + \eta \\ \eta &= \lambda W \eta + \varepsilon. \end{aligned} \tag{11.14}$$

with the null hypothesis $\mu = 0$. In the present application, we must fit Y_1 and Y_2 separately. The use of the model (Equation 11.14) for data such as those in Section 11.1, where we are determining the associations between SPAD meter reading, leaf nitrogen level, and grain protein content, would require the assumption that the deterministic component of each variable is constant over the entire region of measurement. This is generally not appropriate. One possibility to correct for this lack of stationarity would be to insert a trend $T(x,y)$ into the model, and a second would be to insert explanatory variables, as will be done in Chapter 13. A third, simpler possibility is to replace the spatial error model (Equation 11.14) with a different model that may provide a better empirical description of the behavior of the measured quantities. This is the approach we will take. Before going on, however *a word of caution is in order*. So far as I know, using the parametric bootstrap in this way has not been validated theoretically, and the conditions, if any, under which its use is justified have not been delimited. Until it receives further scrutiny, use it at your own risk.

The model we will use is the spatial lag model (Anselin, 1992). This model is discussed in greater detail in Chapter 13, but for the present we will write it as

$$Y = \mu + \rho W Y + \varepsilon. \tag{11.15}$$

where ρ is an autocorrelation parameter and again ε is an independent normally distributed random variable with mean zero and variance σ^2. The reason for the name "spatial lag model" should be evident: the lag term ρW is applied directly to the measured variable Y as opposed to applying it to the error η as in the spatial error model of Equation 11.14. We will not make any attempt to justify the use of the spatial lag model on biophysical grounds at this point (indeed, such a justification may be impossible), but rather simply use it as an empirical description of the observed process.

With this model, we can implement the parametric bootstrap in the usual manner. To apply the model to the artificial data analyzed in Section 11.2.2, we first estimate ρ in Equation 11.15 and generate the matrix $(I - \rho W)^{-1}$ based on this estimate. In the case that the estimated value for ρ is negative, we replace it with zero. The following code uses the same artificial data set as Section 11.2.2:

```
> nlist.10 <- cell2nb(10,10)
> W <- nb2listw(nlist.10) > Y1.vec <- as.vector(t(Y1samp))
> Y1.mod <- lagsarlm(Y1.vec ~ 1, data = data.frame(Y1.vec), W)
> IrWinv.1 <- invIrM(nlist.10, max(Y1.mod$rho, 0), style = "W")
> Y2.vec <- as.vector(t(Y2samp))
> Y2.mod <- lagsarlm(Y2.vec ~ 1, data = data.frame(Y2.vec), W)
> IrWinv.2 <- invIrM(nlist.10, max(Y2.mod$rho, 0), style = "W")
```

Next, we write a function to generate the bootstrap resamples.

```
> boot.samp <- function(Y1.mod,Y2.mod,IrWinv.1,IrWinv.2){
+   e1 <- sample(Y1.mod$residuals, replace = TRUE)
+   Y1 <- IrWinv.1 %*% e1
+   e2 <- sample(Y2.mod$residuals, replace = TRUE)
+   Y2 <- IrWinv.2 %*% e2
+   r.boot <- cor(Y1,Y2)
> }
> U <- replicate(200, boot.samp(Y1.mod,Y2.mod,IrWinv.1,IrWinv.2))
> sigmar.hat <- var(U)
```

Finally, we estimate the value of $\hat{\sigma}_r^2$ and use it to compute the corrected t statistic and corresponding p value.

```
> print(ne.hat <- 1 + 1 / sigmar.hat, digits = 3)
[1] 77
> print(t.corr <- sqrt(ne.hat - 2) * r / sqrt(1 - r^2), digits = 3)
 cor
1.82
> print(p.corr <- 2 * (1 - pt(q = abs(t.corr),
+   df = ne.hat - 2)), digits = 3)
   cor
0.0735
```

The parametric bootstrap is more conservative than the Clifford et al. (1989) correction, with an effective sample size of 77 and a p value of 0.07. As usual, we can test this in a Monte Carlo simulation.

Table 11.1 gives a comparison of the three methods (uncorrected, Clifford correction, and parametric bootstrap) for increasing values of ρ. The method of Clifford et al. (1989) provides a considerable improvement over the uncorrected data except at the highest ρ value, while for this particular example the parametric bootstrap method provides a relatively accurate Type I error rate over the range of ρ values. Once again, it must be emphasized that the parametric bootstrap method in this example has an unfair advantage in that it is being applied to data that are known to satisfy the parametric model. The application of this method to data for which the model is incorrectly specified has not been studied, and how the method would fare under these circumstances is an open question. Indeed, the method is, for the present, best used in conjunction with a Clifford correction as a check on this method.

11.2.4 Application to the Example Problem

The example discussed in Section 11.1 involves the correlation of Minolta SPAD meter reading with leaf nitrogen content and with grain protein content. The scatterplot of SPAD reading with *LeafN* content (Figure 11.1) suggests that the distribution of SPAD readings and/or *LeafN* might be skewed. Although we are carrying out a correlation test at this point, ultimately we will be interested in using the SPAD meter to predict leaf and grain N levels. Regression is more sensitive to data heteroscedasticity than to non-normality (Kutner et al., 2005, p. 110), so we will test for a need to stabilize the variance via a transformation. We will try the original data set and log transformed data.

One way to test the variance stabilizing effect of a log transformation is to carry out a linear regression of both the transformed and untransformed data and then conduct a homogeneity of variance test on the residuals of the linear regression (Kutner et al., 2005, p. 116). Here is the code for a Breusch-Pagan test (Section 9.3) of the two sets of residuals.

```
> spadn.lin <- lm(LeafN ~ SPAD, data = data.Set4.1)
> library(lmtest)
> bptest(spadn.lin)
        studentized Breusch-Pagan test
data:  spadn.lin
BP = 4.4284, df = 1, p-value = 0.03534
> spadn.log <- lm(log(LeafN) ~ log(SPAD), data = data.Set4.1)
> bptest(spadn.log)
        studentized Breusch-Pagan test
data:  spadn.log
BP = 1.0646, df = 1, p-value = 0.3022
```

The results of both the Breusch-Pagan test and a Levene's test (not shown) indicate that homogeneity of variance is considerably improved with the log transformed data. Therefore, these are what we will use. Note, by the way, that the R function for the natural logarithm is `log()`, not `ln()`.

```
> log.SPAD <- log(data.Set4.1$SPAD)
> log.LeafN <- log(data.Set4.1$LeafN)
> cor.test(log.SPAD, log.LeafN, alternative = "two.sided",
+    method = "pearson")
        Pearson's product-moment correlation
data:  log.SPAD and log.LeafN
t = 8.1567, df = 84, p-value = 2.945e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5271963 0.7684462
sample estimates:
     cor
0.6648121
```

The p value of the correlation coefficient, uncorrected for spatial autocorrelation, is $p = 2.945 \times 10^{-12}$.

We are now ready to carry out the correction of Clifford et al. The function `correlogram()` of library `spatial` is again used to compute the correlograms of the ln(*SPAD*) and ln(*LeafN*) data. The code for ln(*SPAD*) is the following:

```
> library(spatial)
> spad.surf <- surf.ls(0, data.Set4.1$Easting,
+    data.Set4.1$Northing,
+    log.SPAD)
> spad.cgr <- correlogram(spad.surf,10, cex.main = 2, # Fig. 11.2a
+    main = "Log SPAD Correlogram", cex.lab = 1.5,
+    xlab = "Lag Group", ylab = "Experimental Correlogram")
```

The code for ln(*LeafN*) is similar. The correlograms are shown in Figure 11.2. The failure of the correlograms to approach zero is an indication of the nonstationarity of the data (Section 4.6.1). Clifford et al. (1989) indicate that their method should still be effective with data such as these that display a global trend. Based on the figures, only the first three correlogram values are used to compute the estimated covariance. Here is the code to generate the value of $\hat{\sigma}_r$ from Equation 11.13.

```
> n <- nrow(data.Set4.1)
> cgr.prod <- spad.cgr$y * LeafN.cgr$y * LeafN.cgr$cnt
> sigr.hat <- sum(cgr.prod[1:3]) / n^2
```

Next, we generate the value of $\hat{n}_e$ from Equation 11.6 and plug this into Equation 11.5 to generate the t statistic.

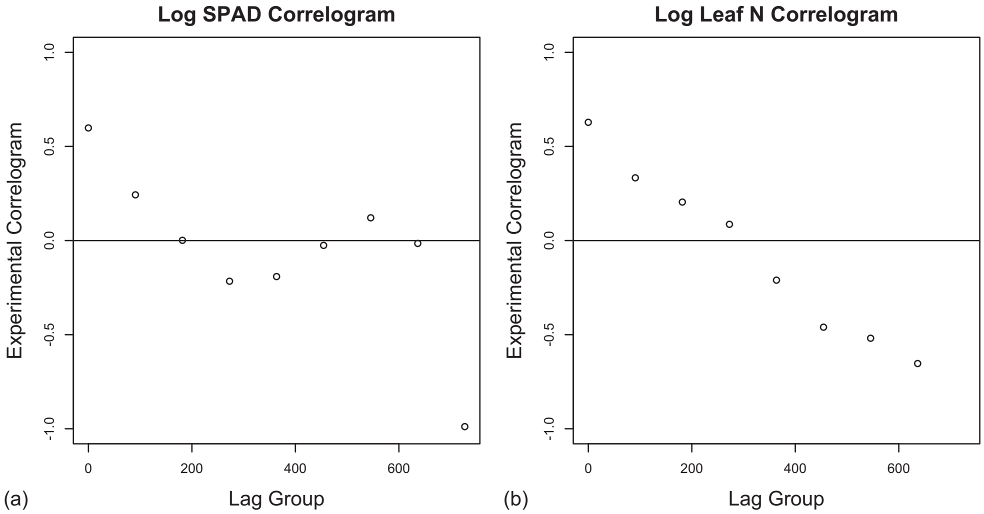

FIGURE 11.2
(a) Correlograms of log *SPAD*; (b) log *LeafN* for the data of Field 4.1.

```
> print(ne.hat <- 1 + 1 / sigr.hat, digits = 3)
[1] 43.3
> print(r <- cor(log.SPAD,log.LeafN), digits = 3)
[1] 0.665
> print(t.uncorr <- sqrt(n - 2) * r / sqrt(1 - r^2), digits = 3)
[1] 8.16
> print(t.corr <- sqrt(ne.hat - 2) * r / sqrt(1 - r^2), digits = 3)
[1] 5.72
> print(p.uncorr <- 2 * (1 - pt(q = t.uncorr, df = n - 2)), digits = 3)
[1] 2.94e-12
> print(p.corr <- 2 * (1 - pt(q = t.corr, df = ne.hat - 2)),digits = 3)
[1] 1.05e-06
```

Not surprisingly, the correlation is still highly significant. Perhaps the most interesting result is that the estimated effective sample size is 43.

Next, we will address the same problem using the parametric bootstrap. The first step is to use the spatial lag model of Equation 11.15 to generate the models for ln(*SPAD*) and ln(*LeafN*).

```
> library(spdep)
> coordinates(data.Set4.1) <- c("Easting", "Northing")
> nlist.w <- dnearneigh(data.Set4.1, d1 = 0, d2 = 61)
> W <- nb2listw(nlist.w)
> Y1.mod <- lagsarlm(log.SPAD ~ 1, data = data.Set4.1, W)
> Y1.mod$rho
      rho
0.7659298
> IrWinv.1 <- invIrM(nlist.w, max(Y1.mod$rho, 0), style = "W")
> Y2.mod <- lagsarlm(log.LeafN ~ 1, data = data.Set4.1, W)
> Y2.mod$rho
      rho
0.7497912
> IrWinv.2 <- invIrM(nlist.w, max(Y2.mod$rho, 0), style = "W")
```

Next, the bootstrap resampling function is constructed. The model is then executed to generate the bootstrap resample.

```
> boot.samp <- function(Y1.mod,Y2.mod,IrWinv.1,IrWinv.2){
+   e <- sample(Y1.mod$residuals, replace = TRUE)
+   Y1 <- IrWinv.1 %*% e
+   e <- sample(Y2.mod$residuals, replace = TRUE)
+   Y2 <- IrWinv.2 %*% e
+   r.boot <- cor(Y1,Y2)
+ }
```

Finally, the bootstrap sample is generated and used to create the results.

```
> set.seed(123)
> U <- replicate(200, boot.samp(Y1.mod,Y2.mod,IrWinv.1,IrWinv.2))
> print(sigmar.boot <- var(U), digits = 3)
[1] 0.0308
> print(r <- cor(log.SPAD, log.LeafN), digits = 3)
[1] 0.665
```

```
> print(ne.hat <- 1 + 1 / sigmar.boot, digits = 3)
[1] 33.5
> print(t.corr <- sqrt(ne.hat - 2) * r / sqrt(1 - r^2), digits = 3)
[1] 5
> print(p.corr <- 2*(1-pt(q=abs(t.corr), df = ne.hat - 2)),digits = 3)
[1] 2.09e-05
```

The value of $\hat{n}_e$ generated by the parametric bootstrap is in this particular example smaller than that generated by the Clifford correction. Repeating the parametric bootstrap for *GrainProt* yields an effective sample size estimate of 39 and a corrected p value of 0.0008 for a correlation coefficient of $r = 0.52$ (Exercise 11.1).

In summary, after correction for spatial autocorrelation of the two variables, the correlation between the logarithms of SPAD reading and leaf nitrogen level is still highly significant. Both the Clifford et al. (1989) and the parametric bootstrap correction yield estimated effective sample sizes considerably less than the nominal value of 86. In order to more fully understand the processes underlying the formation of grain protein, however, it is of interest to determine whether there are any other quantities associated with grain protein. This issue is taken up in Section 11.4.

11.3 Contingency Tables

11.3.1 Large Sample Size Contingency Tables

There are three commonly used methods of dealing with contingency table data: the chi-square test, the likelihood test, and Fisher's exact test (Larsen and Marx, 1986, p. 422; Agresti, 1996, p. 39). The chi-square test and the likelihood tests are more appropriate for large data sets and Fisher's exact test is more appropriate for small data sets (the terms *small* and *large* will be made more precise below). We will discuss the first two tests first. Our discussion is restricted to the 2×2 contingency table case, as in Table 11.2. The extension to the more general case can be found in the references.

Let X and Y represent two categorical variables. We use X and Y in this section because the notation is much simpler, with fewer subscripts to deal with. Also, it is often (although not always) the case that X is considered an explanatory variable for Y. Let $p_1 = \Pr\{X = X_1\}$, and $p_2 = \Pr\{X = X_2\} = 1 - p_1$, and let $q_1 = \Pr\{Y = Y_1\}$, and $q_2 = \Pr\{Y = Y_2\} = 1 - q_1$. Let n_{ij} be the number of observations in which $X = X_i$ and $Y = Y_j$, $i, j = 1, 2$. Assume that there are a total of n observations, that is, that $n_{11} + n_{12} + n_{21} + n_{22} = n$. Assume further that the observations

TABLE 11.2

A two by two Contingency Table

	Y_1	Y_2		
X_1	n_{11}	n_{12}	n_{1m}	$\hat{p}_1 = n_{1m} / n$
X_2	n_{21}	n_{22}	n_{2m}	$\hat{p}_2 = n_{2m} / n$
	n_{m1}	n_{m2}	n	
	$\hat{q}_1 = n_{m1} / n$	$\hat{q}_2 = n_{m2} / n$	1	

are mutually independent (this is the assumption we will drop when we take up the analysis of spatial data). Let the marginal sums n_{im} and n_{mj} be defined as $n_{im} = n_{i1} + n_{i2}$ and $n_{mj} = n_{1j} + n_{2j}$ (Table 11.2).

In the case of the blue oak presence/absence vs. elevation data discussed in Section 11.1, the column variable X represents the elevation of the site with X_1 = {less than or equal to 1100 m} and X_2 = {greater than 1100 m} (Table 11.2), and the row variable Y represents the presence (Y_1) or absence (Y_2) of oaks. In this case, one can think of X as an explanatory variable and Y as a response variable (Agresti, 1997, p. 17). The question of interest to us is whether X and Y are independent. If they are, then the probability that $X = X_i$ and $Y = Y_j$ is just the product $p_i q_j$ of the individual probabilities, and X has no explanatory value concerning Y.

If we let H_0 be the null hypothesis that X and Y are independent and H_a be the alternative that they are not, then under H_0 the statistic

$$C^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - np_i q_j)^2}{np_i q_j} \tag{11.16}$$

has asymptotically as $n \to \infty$ a χ^2 distribution with one degree of freedom (Larsen and Marx, 1986, p. 423; Agresti, 1996, p. 30). The maximum likelihood estimate for p_1 is $\hat{p}_1 = n_{1m} / n$, and that for p_2 is $\breve{p}_2 = n_{2m} / n$, and similarly for $\hat{q}_1$ and $\hat{q}_2$. Therefore, one way to test the null hypothesis is to compute the estimate $\hat{C}^2$ of the statistic C^2 in Equation 11.16 and compare this with the percentiles of the χ_1^2 distribution. Because Equation 11.16 holds only asymptotically as $n \to \infty$, it cannot be reliably applied to data sets with small sample sizes. Snedecor and Cochran (1989, p. 127) and Agresti (1997, p. 28) suggest the rule of thumb that this test can be used if $n > 20$ and if the smallest expected value $n\hat{p}_i \hat{q}_j$ is larger than 5.

The blue oak presence vs. elevation contingency table computed in Section 11.1 is not particularly interesting to analyze, since blue oak presence is clearly dependent on elevation. We will focus on another aspect of Data Set 2 that is more interesting and more helpful to the process of gaining an understanding of the data set. In Section 9.4.2, we constructed logistic regression models of the response of blue oaks to environmental variables. These environmental variables can be roughly divided into three classes: those related to temperature and solar radiation, those related to soil, and those related to precipitation. There is a high level of correlation between variables in these classes, particularly among temperature-related variables (Figure 7.9).

In the exploratory analysis of Section 7.3, we found that in the Coast Range in a multivariate frequency plot with *Precip*, mean annual temperature, or *MAT*, has a bimodal distribution for values of *Precip* less than 800 mm (Figure 7.13). The mean value of *QUDO* tends for a given value of *Precip* to be lower in the Coast Range (Figure 7.12a), so it makes sense to ask whether this could be due to a temperature effect. As a preliminary means of investigating this issue, we will consider the data set consisting of all sites in the Coast Range where $Precip \le 800$, that is, where *MAT* is roughly bimodal. Within this data set, we will subdivide the data records into those with values of *MAT* less than 15°C (`LoMAT = 1`), and those with values greater than or equal 15°C (`LoMAT = 0`). Let X_1 represent the property of a site having a value of `LoMAT` equal one, and let X_2 represent the property of a site having value of `LoMAT` equal zero. Again, let $Y = Y_1$ represent the presence of blue oak at the site and $Y = Y_2$ represent the absence.

First, we set up the data to determine the values of *X* and *Y* from the sf object coast.sf. These are reflected in the R variables LoMAT and QUDO

```
> coast.800 <- coast.sf[which(coast.sf$Precip <= 800),]
> coast.800$LoMAT <- 0
> coast.800$LoMAT[which(coast.800$MAT <= 15)] <- 1
```

Next, we compute the contingency table. With a two by two table, it is easy to do this by hand. With a larger table, we would use the function table().

```
> print(n.mat <- with(coast.800, matrix(c(
+    sum(LoMAT == 1 & QUDO == 1),
+    sum(LoMAT == 1 & QUDO == 0),
+    sum(LoMAT == 0 & QUDO == 1),
+    sum(LoMAT == 0 & QUDO == 0)),
+    nrow = 2, byrow = TRUE,
+    dimnames = list(c("LoMAT", "HiMAT"), c("Pres", "Abs")))))
      Pres Abs
LoMAT  508 523
HiMAT  293 182
> print(n.tot <- sum(n.mat))
[1] 1506
```

A total of 1,506 of the Coast Range sites have values of *Precip* less than or equal 800.

The estimated probabilities for use in Equation 11.16 are computed as follows:

```
> print(p1 <- (n.mat[1,1]+n.mat[1,2])/n.tot, digits = 2)
[1] 0.68
> print(p2 <- (n.mat[2,1]+n.mat[2,2])/n.tot, digits = 2)
[1] 0.32
> print(q1 <- (n.mat[1,1]+n.mat[2,1])/n.tot, digits = 2)
[1] 0.53
> print(q2 <- (n.mat[1,2]+n.mat[2,2])/n.tot, digits = 2)
[1] 0.47
```

Plugging these into Equation 11.16 leads to the following:

```
> print(C2.hat <- (
+    (n.mat[1,1]-n.tot*p1*q1)^2/(n.tot*p1*q1) +
+    (n.mat[1,2]-n.tot*p1*q2)^2/(n.tot*p1*q2) +
+    (n.mat[2,1]-n.tot*p2*q1)^2/(n.tot*p2*q1) +
+    (n.mat[2,2]-n.tot*p2*q2)^2/(n.tot*p2*q2)))
[1] 20.11941
```

Before one tests the null hypothesis of independence of *X* and *Y* using the normal approximation, one must make one last computation, which is called the Yates' correction for continuity (Snedecor and Cochran, 1989, p. 126). The continuity correction results from the fact that we are applying a normal distribution to data that are only defined for integer values. We will not go into the details here; they are provided by Snedecor and Cochran (1989), and the correction is computed automatically in R. The R function chisq.test()

provides two methods of carrying out a significance test, one based on the normal approximation and the other on a permutation test. The continuity correction is applied only to the normal approximation test.

```
> chisq.test(n.mat, simulate.p.value = FALSE)
       Pearson's Chi-squared test with Yates' continuity correction
data:  n.mat
X-squared = 19.624, df = 1, p-value = 9.428e-06
> chisq.test(n.mat, simulate.p.value = TRUE)
       Pearson's Chi-squared test with simulated p-value (based on 2000
replicates)
data:  n.mat
X-squared = 20.1194, df = NA, p-value = 0.0004998
```

Both tests result in a very low *p* value, although the simulated *p* value is by far the higher of the two despite the fact that the computed chi-squared statistics are very close.

Now we incorporate spatial autocorrelation on the data analysis. Based on what we have seen in earlier chapters, we would expect that spatial autocorrelation in the data would have the effect of reducing the effective sample size and thus increasing the *p* value of the significance test. In order to carry out this analysis, however, we need to first describe the second, asymptotically equivalent method for testing the null hypothesis of independence of X and Y. This is to use a maximum likelihood estimator (Agresti, 1996, 2002). The explanation is easiest to understand by using the concept of the *odds ratio* (Agresti, 1996, p. 23). We will use for intuition the interpretation of X as the temperature type in Coast Range (*LoMAT* = 0 vs. *LoMAT* = 1) and Y as the presence vs. absence of blue oaks. We denote by $p_{i,j}$ the *joint probability* $\Pr\{Y = Y_i, X = X_j\}$, that is, the probability that $Y = Y_i$ and that $X = X_j$. We then denote by $p_{i|j}$ the *conditional probability*, that is, the probability that $Y = Y_i$ *given* that $X = X_j$. Recall (e.g., Larsen and Marx, 1986, p. 42; Kutner et al., 2005, p. 1300) that the conditional probability $\Pr\{Y = Y_i \mid X = X_j\}$ satisfies $\Pr\{Y = Y_i \mid X = X_j\} = \Pr\{Y = Y_i, X = X_j\} / \Pr\{X = X_j\}$.

Define the *odds* of the outcome of Y in row j as the ratio of the two conditional probabilities: $odds_j = \Pr\{Y = Y_1 \mid X = X_j\} / \Pr\{Y = Y_2 \mid X = X_j\}$. In our example, $\Pr\{Y = Y_1 \mid X = X_1\}$ represents the probability that a site will contain a blue oak, given that the site is in a low *MAT* environment, and $\Pr\{Y = Y_2 \mid X = X_1\}$ represents the probability that a site will not contain a blue oak given that the site is in a low *MAT* environment. The quantity $odds_1$ represents the odds (ratio of probabilities) that a blue oak will be present in a low *MAT* environment, and $odds_2$ represents the odds that a blue oak will be present in a high *MAT* environment. Intuitively, if the presence or absence of blue oaks is independent of climate type, then these odds should be the same.

To continue, we define the *odds ratio* θ by $\theta = odds_1 / odds_2$. Then

$$\theta = \frac{\Pr\{Y = Y_1 \mid X = X_1\} / \Pr\{Y = Y_2 \mid X = X_1\}}{\Pr\{Y = Y_1 \mid X = X_2\} / \Pr\{Y = Y_2 \mid X = X_2\}}. \tag{11.17}$$

We can estimate θ as follows: The estimator for the joint probability $\Pr\{Y = Y_i, X = X_j\}$ is $\hat{p}_{ij} = n_{ij} / n$. The estimator for p_j is $\breve{p}_j = (n_{1j} + n_{2j}) / n$. Therefore, the estimator for the conditional probability $\Pr\{Y = Y_i \mid X = X_j\}$ is $\hat{p}_{i|j} = n_{ij} / (n_{1j} + n_{2j})$, where we make use of the fact that the n values cancel. Using these relationships, we define the estimator $\hat{\theta}$ as

$$\hat{\theta} = \frac{n_{11}/(n_{11}+n_{12})}{n_{12}/(n_{11}+n_{12})} \Bigg/ \frac{n_{21}/(n_{21}+n_{22})}{n_{22}/(n_{21}+n_{22})}. \\ = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{11.18}$$

Now, suppose X and Y are independent. In this case, we have $\Pr\{Y = Y_i \mid X = X_j\} = \Pr\{Y = Y_i\}\Pr\{X = X_j\}$. Therefore, under this condition Equation 11.17 becomes

$$\theta = \frac{\Pr\{Y = Y_1\}\Pr\{X = X_1\}/\Pr\{Y = Y_2\}\Pr\{X = X_1\}}{\Pr\{Y = Y_1\}\Pr\{X = X_2\}/\Pr\{Y = Y_2\}\Pr\{X = X_2\}} \\ = \frac{\Pr\{Y = Y_1\}\Pr\{Y = Y_2\}\Pr\{X = X_1\}\Pr\{X = X_2\}}{\Pr\{Y = Y_1\}\Pr\{Y = Y_2\}\Pr\{Y = Y_2\}\Pr\{X = X_1\}\Pr\{X = X_2\}}. \\ = 1 \tag{11.19}$$

Therefore, we can test for independence by testing the hypothesis $\theta = 1$ against the alternative $\theta \neq 1$ using the estimator $\hat{\theta}$ given in Equation 11.18. Agresti (1996, p. 24; 2002, p. 425) shows that if we define the statistic W as

$$W = \frac{n(\ln\hat{\theta})^2}{\hat{\sigma}^2}, \tag{11.20}$$

where

$$\hat{\sigma}^2 = \frac{n}{n_{11}} + \frac{n}{n_{21}} + \frac{n}{n_{12}} + \frac{n}{n_{22}} \tag{11.21}$$

then W is the maximum likelihood estimator for the standard error of $\ln\hat{\theta}$. Agresti further shows that (under the assumption that the X_i and Y_i are not spatially autocorrelated) W is asymptotically distributed as χ_1^2 as $n \to \infty$. Thus, the statistic W could be used in a chi-squared test instead of the statistic $\hat{C}^2$ defined earlier, provided that there is no spatial autocorrelation in the data.

The problem, of course, is that for our data the X_i and Y_i *are* autocorrelated, or at least they may be. Motivated by the work of Clifford et al. (1989) discussed in the previous section, Cerioli (1997) developed a means of estimating the effect of spatial autocorrelation using the covariogram. In a slight change of notation from Equation 4.27, we express the covariogram for X as

$$\hat{C}_X(k) = n_k^{-1} \sum_{i,j \in H(k)} (X(x_i, y_i) - \bar{X})(X(x_j, y_j) - \bar{X}). \tag{11.22}$$

and similarly for $\hat{C}_Y(k)$. Recall that both X and Y are binary variables. With these definitions, Cerioli (1997) showed the following: Define a quantity λ by

$$\lambda = \frac{2}{np_1q_1p_0q_0}\sum_{h=1}^{K} n_k\hat{C}_X(k)\hat{C}_Y(k). \tag{11.23}$$

Then the adjusted statistic W_{adj} defined by

$$W_{adj} = \frac{W}{1+\lambda} \tag{11.24}$$

is asymptotically distributed as a χ_1^2 random variable under the null hypothesis that X and Y are independent. In other words, W_{adj} is adjusted to take into account the effects of spatial autocorrelation. Cerioli (1997) also derives a second adjustment statistic based on the Moran correlogram. We will not use this second adjustment, but rather base our analysis on Equation 11.24.

As with the correction of Clifford et al. (1989) (Equation 11.13), Equation 11.24 involves the product of the experimental covariograms for X and Y, and therefore if either X or Y is spatially uncorrelated, then its covariogram is zero and the autocorrelation of the other variable has no effect on the significance test. Moreover, if both X and Y are positively autocorrelated, then since λ must be greater than zero, the value of W_{adj} as calculated in Equation 11.24 will be less than the value of W, so that the adjustment for autocorrelation reduces the value of the test statistic.

The computation of W based on Equations 11.20, 11.23, and 11.24 proceeds as follows. First, we compute W in Equation 11.20.

```
> print(theta.hat <-
+    (n.mat[1,1]*n.mat[2,2]) / (n.mat[1,2]*n.mat[2,1]), digits = 3)
[1] 0.603
> print(sig2.hat <- sum(1/(n.mat/n.tot)), digits = 3)
[1] 19.3
> print(W <- n.tot * log(theta.hat)^2 / sig2.hat, digits = 4)
[1] 19.96
```

The value of W is quite close to, and in between, those of the chi-square statistics computed above. To implement Equation 11.23, we must compute the covariograms of X and Y. The covariogram at lag k can be estimated from the variogram through the use of Equation 4.26, $\gamma(h) = C(0) - C(h)$, which, when written in terms of lag groups instead of lags, implies that $\hat{C}(k) = \hat{C}(0) - \hat{\gamma}(k)$. Computation of the nugget estimator $\hat{C}(0)$ ordinarily requires that we fit a model to the variogram. For this application, however, it is simpler, and probably sufficiently accurate, to assume that at large values of k the variogram has reached its sill and therefore $C(k_{l\arg e}) \cong 0$. This implies that

$$\hat{C}(0) \cong \hat{\gamma}(k_{l\arg e}) \tag{11.25}$$

is an adequate estimator, and we will use it. As in the case of the Clifford correction of the previous section, we choose the function `variogram()` of the `spatial` package for our variogram estimator. Here is the code.

```
> library(spatial)
> qudo.surf <- with(coast.800, surf.ls(2, Longitude, Latitude, QUDO))
> qudo.vgr <- variogram(qudo.surf,200)
```

The raw experimental variograms of the *QUDO* and *LoMAT* data are fairly ugly, as might be expected for a binary variable defined over a long, narrow domain. We are only interested in the variogram at the smallest lags, however, and fortunately these are relatively well behaved (Figure 11.3). We will sum over the first 20 terms in the experimental variogram and use the maximum over these terms as an estimate of the sill. Applying Equations 11.23, 11.24, and 11.25, we also check to make sure that the numbers of pairs in the lag groups of X and Y are the same.

```
> CX.hat <- max(qudo.vgr$y[1:20]) - qudo.vgr$y[1:20]
> nkX <- qudo.vgr$cnt[1:20]
> CY.hat <- max(lomat.vgr$y[1:20]) - lomat.vgr$y[1:20]
> nkY <- lomat.vgr$cnt[1:20]
> all.equal(nkX,nkY)
[1] TRUE
> print(lambda <-
+    (2 * n.tot* lambda.sum)/(n.mat[1,1]*n.mat[2,2]), digits = 3)
[1] 6.72
> print(W.adj <- W / (1 + lambda), digits = 3)
[1] 2.59
```

Since W_{adj} is asymptotically chi square, we can test its significance as follows:

```
> print(p <- 1 - pchisq(W.adj, df = 1), digits = 3)
[1] 0.108
```

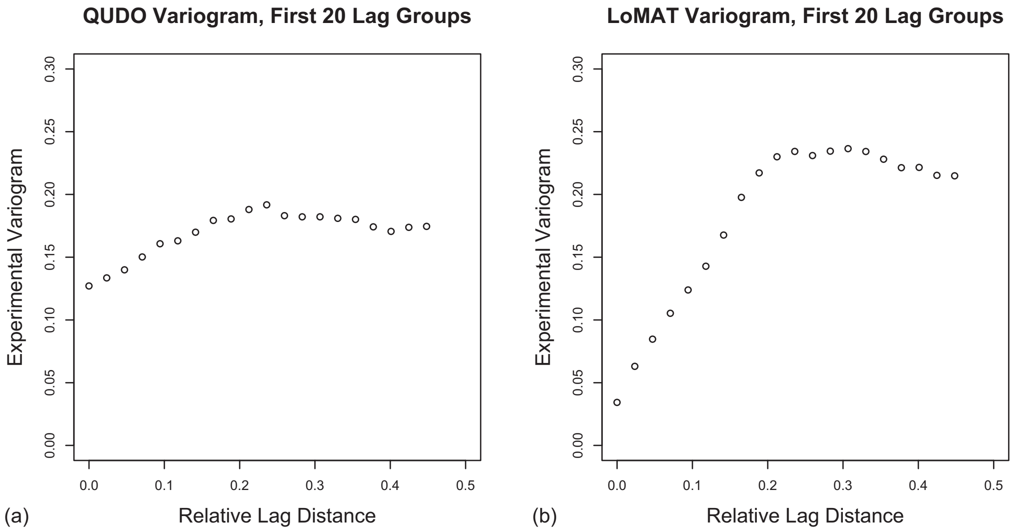

FIGURE 11.3
Experimental variograms of (a) *QUDO* and *LoMAT* from Data Set 2; (b) Only the first 20 lag groups are plotted.

The statistic is not significant, which calls into question whether there is a significant effect of mean annual temperature on the presence or absence of blue oaks in the Coast Range. We must remember, however, not to take the p value of these significance tests too seriously, since they are an approximation of a quantity that is somewhat questionable in the first place. The main conclusion is that we need to be cautious about interpretation of our results. Interestingly, when this analysis is repeated using *JuMax*, the maximum July temperature, the resulting contingency table is significantly different from random even when spatial autocorrelation is incorporated (Exercise 11.2). This may be meaningful, or it may just be data snooping, but we will pursue it.

11.3.2 Small Sample Size Contingency Tables

In addition to the contingency table developed in Chapter 7 for the distribution of blue oaks with elevation, a second contingency table was developed in Section 7.2, involving Data Set 1. The contingency table has as entries the relationship between X, representing suitability of habitat patches for the yellow-billed cuckoo as defined by the California Wildlife Habitat Relationships (CWHR) model, and Y, representing the observation or failure to observe at least one cuckoo in the patch. The contingency table has values $n_{11} = 5$, $n_{12} = 1$, $n_{21} = 2$, and $n_{22} = 12$. The number of observations in this contingency table does not fall into the range in which the asymptotic chi-square approximation is valid. One means of overcoming this problem is to use a Monte Carlo simulation. When it is called with the argument `simulate.p.value = TRUE`, the R function `chisq.test()` carries out a Monte Carlo simulation in which contingency tables are generated whose entries n_{ij} are sets of random numbers such that their values add up to the marginal values of the contingency table being tested and the fraction of the tables whose chi-square statistic is more extreme is recorded (Hope, 1968).

An alternative method, Fisher's exact test (Fisher, 1922), relies on the fact that, under the assumption of independence of X and Y and for fixed marginal values, the elements of the contingency table are distributed according to a hypergeometric probability distribution (Larsen and Marx, 1986, p. 91). For a given set of marginal values n_{im} and n_{mj}, the process of specifying the value of n_{11} in a two by two contingency table also specifies the others by the equations $n_{12} = n_{1m} - n_{11}$, $n_{21} = n_{m2} - n_{11}$ and $n_{22} = n - n_{11} - n_{12} - n_{21}$. The probability distribution for n_{11} under the independence assumption has the form (Agresti, 1995, p. 39)

$$\Pr\{n_{11} = k\} = \binom{n_{m1}}{k}\binom{n_{m2}}{n_{m1} - k}\bigg/\binom{n}{n_{m1}}, \tag{11.26}$$

which defines a hypergeometric distribution. If n is relatively small, the total number of possible enumerations (i.e., of values of k in Equation 11.26) will be small as well, and therefore it is relatively easy to compute the number of enumerations that are "more extreme" than the observed one, and thereby obtain a p value for the test of the null hypothesis of independence of X and Y. Agresti (2002, p. 93) points out that for small sample sizes n, the p values resulting from Fisher's exact test can only take on a relatively small number of discrete values, and for this reason using the test in a hypothesis test context with a fixed α value of, say, 0.05 will result in the test being conservative. Agresti (2002, p. 94), therefore, recommends simply reporting the p value. There are some other technical issues with the test, which are discussed by Campbell (2007).

In Section 7.2, we generated a contingency table for presence vs. absence of the yellow-billed cuckoo based on suitable vs. unsuitable habitat as defined by the model. The contingency table was based on an array `PresAbs` of bird observations (1 = observed, 0 = not observed), and a second array `HSIPred` (1 = suitable, 0 = not suitable). The data are loaded as described in Appendix B.1 and Section 7.2.

```
> Set1.corrected$HSIPred
 [1] 1 0 1 0 1 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0
> Set1.corrected$PresAbs
 [1] 1 0 1 0 1 0 0 1 0 0 0 0 1 1 0 0 1 0 0 0
> 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
```

Here are the results of a Monte Carlo based chi-square test, an asymptotic chi-square test, and a Fisher test for the cuckoo data. The R function `fisher.test()` returns the estimated value $\hat{\theta}$ of the odds ratio (Equation 11.17) as well as the p value.

```
> chisq.test(cont.table, simulate.p.value = TRUE)
X-squared = 8.8017, df = NA, p-value = 0.009995
> chisq.test(cont.table, simulate.p.value = FALSE)
X-squared = 6.0283, df = 1, p-value = 0.01408
Warning message:
In chisq.test(cont.table, simulate.p.value = FALSE) :
  Chi-squared approximation may be incorrect
> fisher.test(cont.table)
p-value = 0.007224
sample estimates:
odds ratio
  22.96080
```

The declared significance value of the table in both the permutation test and the Fisher test is about 0.01. Once again however, this data set, like Data Set 2, contains spatially autocorrelated data, which may invalidate the results of the test. With only 20 values, there are not enough data to compute a correlogram as would be required by the method of Cerioli used in the previous section.

Since an adjusted test based on the asymptotic properties of the distribution is not possible, it is natural to consider employing a permutation test. There is, however, an important difference between the application of a permutation test to this problem and the applications discussed in Chapter 4. As was the case with the bootstrap employed in Chapters 9 and 10, the process of randomizing spatially autocorrelated data breaks up the spatial pattern, which defeats the purpose of employing the randomization. Analogous to the bootstrap case, permutation tests carried out on spatially autocorrelated data must involve a restricted randomization that preserves the spatial structure of the data.

Fortin and Jacquez (2000) discuss restricted randomization for the purpose of carrying out permutation tests on spatial data. They list three approaches. One approach, generally the most effective when it can be used, is the application of conditional simulation (Cressie, 1991, p. 207; Burrough and McDonnell, 1998, p. 152) to spatially autocorrelated data. Conditional simulation is a form of restricted randomization that uses the kriged interpolation estimate of the quantity being simulated to generate random data arrangements having the same spatial autocorrelation structure as the original data set. Unfortunately, our data set has only 20 values, which is not sufficient to generate the variogram necessary to compute a kriged estimate.

A second approach discussed by Fortin and Jacquez (2000) is the *toroidal shift* (Upton and Fingleton, 1985, p. 253). In this approach the map is converted into a *torus*, that is, a donut-shaped object, by first connecting the east and west edges to form a cylinder and then looping the cylinder so that the north and south edges connect. Coordinates are then repeated shifted a randomly selected distance on the torus. Fortin and Dale (2005, p. 240) give an example of this method. However, this method works best when the data are continuously defined on a rectangular surface. Data Set 1 does not fit this criterion.

The third approach discussed by Fortin and Jacquez (2000) is to restrict the randomization to geographically defined subsets. This is considered the least powerful of the three methods, but it can be applied to our present data set. Its use is described by Manly (1997, p. 188). The data are first divided into geographic blocks (Figure 11.4). Randomization only takes place within these blocks. That is, one variable is randomly rearranged (we will use the presence-absence data) within each block and then the resulting rearrangement is compared with the original arrangement of the other variable, in this case, the habitat suitability index. We will run a permutation test using this restricted randomization according to blocks in the same manner as the permutation tests described in Chapter 4. We use the function `fisher.test()`

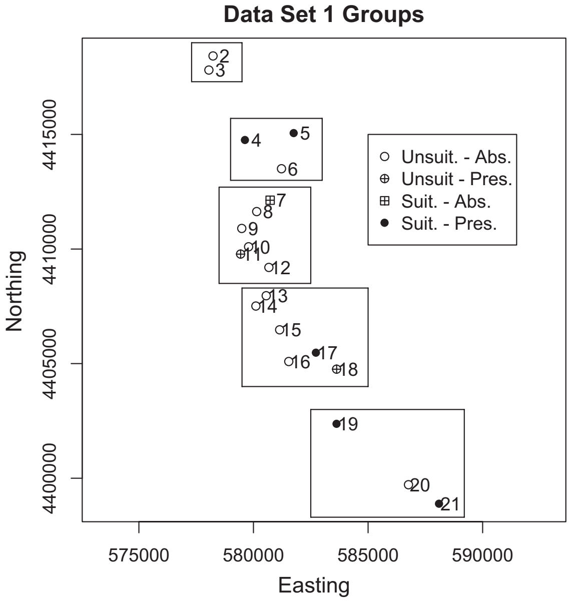

FIGURE 11.4
Subdivision into 5 blocks of the 20 observation points making up the contingency table for the data of Data Set 1.

to compute an odds ratio statistic, and determine where the observed odds ratio statistic lies along the spectrum of statistics computed from permuted data.

This method bears a superficial resemblance to the block bootstrap discussed in Section 10.4.1. There is, however, an important difference. In the case of the block bootstrap, the spatial arrangement of data is preserved within the blocks, and the randomization is accomplished by rearranging the blocks. In the case of the permutation test described in this section, the arrangement of the blocks is preserved and the randomization is accomplished by rearranging the data with each block. The reason for the difference is the following. The objective of the block bootstrap is to estimate the sample variance. Spatial autocorrelation disrupts the estimate because of the correlation of values of with their neighbors. This effect is indicated by Equation 3.28, which is reproduced here:

$$\mathrm{var}\{\bar{Y}\} = \frac{\sigma^2}{n} + \frac{2}{n^2}\sum_{i \neq j} \mathrm{cov}\{Y_i, Y_j\}. \tag{11.27}$$

The arrangement of the data values within blocks is preserved in the block bootstrap in order to preserve the covariance term on the right-hand side, which involves values of Y_i in relation to those at nearby locations. In the case of the permutation test discussed in this chapter, the effect of autocorrelation is indicated by Equation 11.23,

$$\lambda = \frac{2}{np_1q_1p_0q_0}\sum_{h=1}^{K} n_k \hat{C}_X(k)\hat{C}_Y(k). \tag{11.28}$$

The effect of spatial autocorrelation in this case depends on the spatial correlation structure of both X and Y. By rearranging the Y_i only within blocks, we do our best to preserve the correlation structure of the Y_i and its relation to that of the X_i.

We begin the implementation of block-based randomization by creating five blocks of observation data to match the blocks of Figure 11.4.

```
> obs.block1 <- Set1.corrected$PresAbs[1:2]
> obs.block2 <- Set1.corrected$PresAbs[3:5]
> obs.block3 <- Set1.corrected$PresAbs[6:11]
> obs.block4 <- Set1.corrected$PresAbs[12:17]
> obs.block5 <- Set1.corrected$PresAbs[18:20]
```

Next, we create a function `sample.blocks()` to sample without replacement within each block.

```
> sample.blocks<- function(){
+    s1 <- sample(obs.block1)
+    s2 <- sample(obs.block2)
+    s3 <- sample(obs.block3)
+    s4 <- sample(obs.block4)
+    s5 <- sample(obs.block5)
+    c(s1,s2,s3,s4,s5)
+ }
```

Next, we create a function called `calc.fisher()` that computes the estimate of the odds ratio statistic for the block permutation of sites.

```
> calc.fisher <- 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)
+    odds.ratio <- fisher.test(n)$estimate
+ }
```

To carry out the permutation test, we first recompute the observed odds ratio statistic for the data in their observed arrangement.

```
> obs.test <- fisher.test(cont.table)
> print(obs.stat <- obs.test$estimate)
odds ratio
  22.96080
```

Next, we sample 1,999 block permutations, letting the observed data serve as the two thousandth.

```
> set.seed(123)
> U <- replicate(1999,calc.fisher())
```

We now determine where the observed statistic lies in the ranks of all of the permutations.

```
> print(p <- sum(U >= obs.stat - 0.001) / 2000)
[1] 0.0725
```

If you are wondering what the 0.001 is doing in there, see Exercise 11.7.

The value $p = 0.073$ generated by the restricted permutation test is considerably larger than the value $p = 0.0072$ from the original Fisher test. Here we must be careful not to fall into the $p = 0.05$ trap, that is, we must not take these p values too seriously. The fact that we get a p value of 0.07 rather than, say, 0.04, really doesn't matter much. To get some perspective, the original value $p = 0.0072$ is roughly the probability of tossing a fair coin and getting heads seven times in a row. This is fairly unlikely. The value $p = 0.073$ is close to the probability of tossing the same coin and getting heads four times in a row. This is not all that unlikely. At this stage, all we can say that there may be a relationship between the habitat suitability index, as computed by multiplying the quantities in the model, and the presence or absence of cuckoos in a patch, but we cannot be too sure.

The significance test of independence of X and Y carried out in this section does not really satisfy the original objective laid out in Chapter 1, which was to test the CWHR model for habitat suitability. A simple test of independence provides a fairly weak test of the model. It simply shows that the model is at least as good as a random assignment of classes to observations. We will next see Data Set 1 in Section 17.2, where we draw our final conclusions using a different, and arguably more appropriate, statistic.

11.4 The Mantel and Partial Mantel Statistics

11.4.1 The Mantel Statistic

The autocorrelation statistics discussed in Sections 4.3 and 4.4, namely the join–count statistics, Moran's I, and Geary's c, are all special cases of the Mantel statistic Γ defined in Equation 4.1. The theory of these statistics was first studied by Mantel (1967) and Mantel and Valand (1970) in the more general context of comparing two matrices. Let the $n \times n$ matrices A and B have elements a_{ij} and b_{ij}, where $1 \le i, j \le n$. By analogy with the Pearson's product moment correlation coefficient (Equation 11.3), Mantel (1967) defined the correlation coefficient between these matrices as (Manly, 1997, p. 174)

$$r = \frac{\Sigma_{ij} a_{ij} b_{ij} - \Sigma_{ij} a_{ij} \Sigma_{ij} b_{ij} / m}{\sqrt{(\Sigma_{ij} a_{ij}^2 - \left[\Sigma_{ij} a_{ij}\right]^2 / m)(\Sigma_{ij} b_{ij}^2 - \left[\Sigma_{ij} b_{ij}\right]^2 / m)^2}} \tag{11.29}$$

where the symbol Σ_{ij} indicates summation over both the indices i and j and $m = n(n-1)$ is the number of off-diagonal elements. If A and B are symmetric, then the sums can be taken over only the lower triangular elements, setting $m = n(n-1)/2$. Mantel (1967) showed that the distribution of r is asymptotically normal. Although this is true, nevertheless, in practice the significance of r is generally tested using a permutation test. As described in Section 4.2, this involves permuting the ordinal relationship between the a_{ij} and the b_{ij} and computing the statistic for each permutation. The only term in Equation 11.27 that is affected by the order of the a_{ij} and b_{ij} is the term

$$Z = \Sigma_{ij} a_{ij} b_{ij}, \tag{11.30}$$

where again the sum is over both i and j, so this quantity may be used in place of r in a permutation test of the significance of the relation between A and B.

We will illustrate the permutation test process by working through a simple example. We examine the relationship between the logarithms of the leaf nitrogen level *LeafN* and SPAD reading *SPAD*, respectively, in the four points in the two by two square in the northwest corner of Field 4.1. We will use the package `vegan` (Oksanen et al., 2017). Let A be a matrix measuring the distance between ln(*LeafN*) values, where the distance between two values Y_i and Y_j is defined as $d_{ij} = |Y_i - Y_j|$, and let B be a distance matrix measuring the distance between corresponding ln(*SPAD*) values. By their definition, A and B are symmetric, and therefore we may restrict the computation of Z in Equation 11.30 to the lower triangular parts of A and B. The lower triangular part of A is given by

$$A = \begin{bmatrix} a_{12} & & \\ a_{13} & a_{23} & \\ a_{14} & a_{24} & a_{34} \end{bmatrix}, \tag{11.31}$$

and the lower triangular part of B is defined similarly.

Here is the computation of A using the function `vegdist()` from the `vegan` package. As usual, the data frame `data.Set4.1` holds the contents of the file *Set4.196Sample.csv* and is loaded using the statements in Appendix B.4.

```
> library(vegan)
> data.2by2 <- with(data.Set4.1,data.Set4.1[which((Row <= 2) &
+    (Column <= 2)),])
> print(log.LeafN <- log(data.2by2$LeafN), digits = 3)
[1] 1.28 1.14 1.14 1.23
> print(A <- vegdist(log.LeafN, method = "euclidean"), digits = 3)
       1      2      3
2 0.1335
3 0.1415 0.0080
4 0.0454 0.0881 0.0961
```

The matrix B of distances of log *SPAD* is computed similarly.

```
> log.SPAD <- log(data.2by2$SPAD)
> B <- vegdist(log.SPAD, method = "euclidean")
```

The function `mantel()` of the `vegan` package carries out a permutation test.

```
> mantel(A, B)
'nperm' >= set of all permutations: complete enumeration.
Set of permutations < 'minperm'. Generating entire set.
Mantel statistic based on Pearson's product-moment correlation
Call:
mantel(xdis = A, ydis = B)
Mantel statistic r: 0.305
     Significance: 0.25
Upper quantiles of permutations (null model):
  90%   95% 97.5%   99%
0.329 0.608 0.652 0.662
Permutation: free
Number of permutations: 23
```

Although the Mantel statistic can be used in this manner as a measure of association between two quantities, that is what the Pearson product moment correlation does, so the Mantel statistic is not usually used in this way. The distinguishing feature of the Mantel statistic is that it can serve as a means of measuring the association between two bivariate quantities, or between a univariate and a bivariate quantity. Thus, one can measure the association between a quantity Y_i measured at location (x_iy_i) and the same quantity Y_j measured at location (x_j, y_j). Here Y is the univariate quantity and (x, y) is the bivariate quantity. In this way, the Mantel statistic becomes a measure of spatial autocorrelation. Indeed, as mentioned in Chapter 4, both the Moran's I and the Geary's c can be considered as special cases of a Mantel statistic.

One of the primary uses of the Mantel statistic Z as defined in Equation 11.30 is in the *partial Mantel* test (Smouse et al., 1986). This test is used primarily to determine the relationship between two quantities after taking into account their spatial relationship. In order to describe the partial Mantel test, we first briefly review the concept of partial regression, which was introduced in Section 8.2.2 the context of partial regression plots, which in that section are called added variable plots.

Recall from that section that if we have a set of measurements of a response variable Y and corresponding sets of measurement of two explanatory variables, X_1 and X_2, then we can use partial regression to help determine whether the explanatory variable X_2 provides a substantial amount of information about Y independently of the information provided by X_1.

Specifically, partial regression measures the effect of adding X_2 to the linear regression model given that X_1 is already in the model. To carry out a partial regression analysis (Kutner et al., 2005, p. 268) one computes the residuals of a regression of the response variable Y on X_1, which we denote $Y \mid X_1$ and the residuals of X_2 on X_1, which we denote $X_2 \mid X_1$. One then carries out a linear regression of $Y \mid X_1$ on $X_2 \mid X_1$. Intuitively, the residuals $Y \mid X_1$ contain the information about the variability of Y not explained by X_1. The residuals $X_2 \mid X_1$ contain the information about the variability of X_2 not explained by X_1. If X_2 contains substantial information about Y that is not provided by X_1, then the coefficient of determination R^2 of the regression of $Y \mid X_1$ on $X_2 \mid X_1$ will be high. On the other hand, if most of the information that X_2 provides about Y is also provided by X_1, then the coefficient of determination R^2 of the regression of $Y \mid X_1$ on $X_2 \mid X_1$ will be low. In simple linear regression, there is a close relationship between the correlation coefficient and regression. This relationship was exploited by Smouse et al. (1986). They viewed the Mantel statistic as a measure of a linear regression and applied the partial regression concept to develop the partial Mantel test. In order to explain their work, we must first cover one further aspect of the theory of simple linear regression.

In Appendix A.2, mention is made of two types of simple linear regression. In the first type, which is discussed in the appendix, the explanatory variable X is a *mathematical variable*, that is, it is a deterministic quantity. A simple example of this is an experiment involving wheat yield vs. applied fertilizer, in which the experimenter applies fertilizer at fixed rates X of, say, 0, 50, 100, 150, and 200 kg ha^{-1} and measures the yield Y. In the second type of regression, both X and Y are random variables. An example of this would be an experiment carried out at many locations in which the explanatory variable X is total seasonal rainfall and the response variable Y is again wheat yield. Suppose X and Y are normally distributed random variables with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2. Changing the notation slightly from Equation 11.1, let

$$\rho = \frac{\text{cov}\{Y, X\}}{\sigma_Y \sigma_X}, \tag{11.32}$$

where $\text{cov}\{Y, X\}$ is the covariance between Y and X, σ_Y is the standard deviation of Y, and σ_X is the standard deviation of X. Then the correlation coefficient ρ is related to simple linear regression of Y on X through the equation (Larsen and Marx, 1986, p. 446)

$$E\{Y \mid X\} = \mu_Y + \frac{\rho \sigma_Y}{\sigma_X}(X - \mu_X). \tag{11.33}$$

That is, if the regression equation is written as $Y_i = \alpha + \beta X_i + \varepsilon_i$, then the regression coefficient β satisfies the equation $\beta = \rho\sigma_Y/\sigma_X$. The estimator b of β obtained by least squares satisfies a similar relationship,

$$b = r s_Y / s_X, \tag{11.34}$$

One further regression property simplifies our analysis. The regression of the residuals of Y on X_1 against the residuals of X_2 on X_1 is a simple linear regression, and therefore the correlation coefficient between these sets of residuals is the square root of the coefficient of determination (Kutner et al., 2005, p. 78). For this reason, we do not need to actually carry out the regression in order to compute the coefficient of determination. Instead, we can simply compute the correlation coefficient between the sets of residuals $Y \mid X_1$ and $X_2 \mid X_1$. This is called the *partial correlation coefficient* (Kutner et al., 2005, p. 270).

11.4.2 The Partial Mantel Test

Smouse et al. (1986) exploited the close relationship between correlation and regression statistics in their development of the partial Mantel test. They showed that the Mantel statistic of Equation 11.29 is related to a simple linear regression in the same way that r in Equation 11.2 is related to the regression model of Equation A.22. The most powerful aspect of their work, and the most widely used, is the partial Mantel test. Suppose one has three distance matrices as defined in Section 11.4.1, denoted A, B, and C. One can compute the residuals of the simple linear regression of the matrix A on C and of B on C, and then compute the Mantel statistic of these residuals. Since the Mantel statistic is a matrix correlation, this will provide information about the contribution of distance matrix B to the regression model for A given that matrix C is already included in the model. Smouse et al. (1986) refer to this as a *partial Mantel test*. Since the terms of the distance matrices are not independent, the Student t statistic cannot be employed to test significance. Therefore, Smouse et al. (1986) recommend testing for significance using a permutation test.

Although the partial Mantel test can be applied to any distance matrix, when dealing with spatial data one often carries out the test using a matrix C that represents actual geographic distances. Then a partial Mantel test of A and B on C tests whether or not there is spatial information not in the model that significantly contributes to explaining the variation in A. For example, in the relationship between ln(*SPAD*) and ln(*LeafN*) in Section 11.2.1, let A be the distance matrix of ln(*LeafN*), let B be the distance matrix for ln(*SPAD*), and let C be the Euclidean distance matrix between locations i and j. Then a significant value of the partial Mantel statistic indicates that there is spatial information besides that provided by ln(*SPAD*) that contributes to explaining the variation in ln(*LeafN*).

In our derivation, we will first consider only the "attribute distance" matrices A and B, and after we have worked through these, we will incorporate the geographic distance matrix C. For purposes of generality Smouse et al. (1986), consider matrices A and B that include upper as well as lower diagonal elements. In the case of symmetric distance matrices such as those in our example, this simply means including the upper as well as the lower triangular part. Let $\bar{A} = A/(n[n-1])$, the denominator being the total number of off-diagonal elements, and let $\bar{B}$ be defined similarly. Regarding the elements a_{ij} and b_{ij} as samples of random variables, the quantities $\bar{A}$ and $\bar{B}$ are the sample means. Smouse et al. (1986) define the *corrected sum of products* $SP(A, B)$ as

$$SP(A,B) = Z - n(n-1)\bar{A}\bar{B}, \tag{11.35}$$

where Z is the Mantel statistic defined in Equation 11.30, and they define the *corrected sums of squares* $SS(A)$ and $SS(B)$ as

$$\begin{aligned} SS(A) &= \Sigma_{ij}(A-\bar{A})^2, \\ SS(B) &= \Sigma_{ij}(B-\bar{B})^2. \end{aligned} \tag{11.36}$$

Smouse et al. (1986) point out that in a permutation test, $SP(A, B)$ is affected by a reordering of the values of the a_{ij} and b_{ij} in exactly the same way as is Z (since $n(n-1)\bar{A}\bar{B}$ is not affected by permutations of a_{ij} and b_{ij}), and $SS(A)$ and $SS(B)$ are not affected by permutations. Therefore, the results of a permutation test involving a function of these statistics will be the same as a result of a permutation test on the Mantel statistic Z.

Now let

$$b_{AB} = SP(A,B)/SS(A), \tag{11.37}$$

and consider the equation

$$b_{ij} - \mu_b = \beta(a_{ij} - \mu_a) + \varepsilon_{ij}, \tag{11.38}$$

where μ_a and μ_b are the population means of the distributions from which the a_{ij} and the b_{ij} are drawn, respectively, and the ε_{ij} are independent, identically distributed normal random variables with mean zero and variance σ^2 Taking an expectation over B conditional on A yields (recall that $E\{\varepsilon\} = 0$)

$$E\{B \mid A\} = \mu_b + \beta E\{A - \mu_a\}. \tag{11.39}$$

Comparing Equation 11.39 with Equation 11.33 indicates that they have the same form. Define the quantity r_{AB} by

$$r_{AB} = SP(A,B)/[SS(A)SS(B)]^{1/2}. \tag{11.40}$$

This has the same form as that of r in Equation 11.32. Based on Equation 11.37, we can write

$$b_{AB} = r_{AB}\sqrt{SS(B)}/\sqrt{SS(A)}, \tag{11.41}$$

which has the same form as Equation 11.34. Therefore, computing a Mantel statistic is equivalent to computing a simple linear regression of B on A.

Now, consider the case where there is a third distance matrix C. By analogy with the discussion of the partial regression coefficients in Section 11.4.1, we can compute the partial correlation coefficient between the residuals of the regression of A on C and the residuals of the regression of B on C. Assume the elements of C are the geographic distances between the (x, y) coordinates of the locations of measurement of A and B, that is, $c_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. The residuals of the regression of A on C contain all the information about A not predicted by a linear regression of A on location, and the residuals of B on C contain similar information about B. If the correlation coefficient between these two sets of residuals is approximately zero, then this is an indication that B provides little information about A beyond that which is due to how A and B vary together with location. If the correlation coefficient is large, then this is an indication that B provides more information about A than simply that due to their shared location. Finally, as Smouse et al. (1986) point out, because the relationship between these sets of residuals is one of simple linear regression, in which the correlation coefficient equals the square root of the coefficient of determination, it is not necessary to make any regression-related assumptions about the relationship between predictors and response variables. One can equally consider the partial correlation coefficient as measuring the association between the three distance matrices.

When C is a matrix containing geographic distances, it is often recommended that inverse distances be used in order to put greater weight on small distances rather than large distances (Manly, 1997, p. 177, cf. the discussion on the spatial weights matrix in Section 3.2.2). We will consider the cases where C represents a Euclidean distance matrix and an inverse Euclidean distance matrix between the sampling points at which the data for matrices A

and *B* are measured. Smouse et al. (1986) discuss the more general case where the matrices *A*, *B*, and *C* are arbitrary distance matrices. In the context of the example of the relationship between ln(*LeafN*) content and ln(*SPAD*), we are addressing the issue of whether these two quantities really are associated or whether they are both responding to some environmental variable that is associated with the measurement locations. We use functions from the `vegan` package. First, we use the function `vegdist()` to compute the distance matrices *A* for ln(*LeafN*) and *B* for ln(*SPAD*).

```
> LeafN.dist <- vegdist(log(data.Set4.1$LeafN), method = "euclidean")
> SPAD.dist <- vegdist(log(data.Set4.1$SPAD), method = "euclidean")
```

Next, we use the function `dist()` to compute the matrix *C* representing the Euclidean distance between sample points. Since *C* is symmetric, we compute only the lower triangular part to reduce the amount of computational complexity. Here is the code to compute the distance matrix and the matrix of inverse distances.

```
> dist.mat <- with(data.Set4.1,
+    dist(cbind(Easting, Northing), upper = FALSE))
> invdist.mat <- 1 / dist.mat
```

Next, we use the `vegan` function `mantel.partial()` to compute the partial Mantel statistic.

```
> mantel.partial(SPAD.dist, LeafN.dist, dist.mat)
Partial Mantel statistic based on Pearson's product-moment correlation
Call:
mantel.partial(xdis = SPAD.dist, ydis = LeafN.dist, zdis = dist.mat)
Mantel statistic r: 0.4556
     Significance: 0.001
Upper quantiles of permutations (null model):
   90%    95%  97.5%    99%
0.0708 0.0920 0.1141 0.1404
Permutation: free
Number of permutations: 999
```

The results of using the distance matrix are shown; results using the inverse distance (not shown) are very similar. The value of the Mantel statistic is highly significant. This indicates that the linear relationship between ln(*SPAD*) and ln(*LeafN*) is not based simply on their common reaction to a location effect. As with the partial regression coefficient described in Section 9.2.2, the partial Mantel statistic can be very difficult to interpret (Legendre and Legendre, 1998, p. 559). The value and significance of the statistic depends on the relationship between *A* and *B*, between *A* and *C*, and between *B* and *C*. One of the original goals of this chapter was to determine whether the relation between SPAD meter reading and grain protein content is based strictly on geographic location. You are asked to make this determination in Exercise 11.9. A number of other three-way distance matrix tests besides the partial Mantel test have been proposed (Manly, 1997, p. 181). Oden and Sokal (1992) compared several of these and concluded that all of the methods had problems dealing with spatially autocorrelated data, but that the method of Smouse et al. (1986) is conservative, that is, it is unlikely to falsely reject the null hypothesis of zero correlation.

11.5 The Modifiable Areal Unit Problem and the Ecological Fallacy

11.5.1 The Modifiable Areal Unit Problem

Spatial analysis often involves aggregating data that have been collected at different spatial scales, as discussed in Section 6.4. In addition, data collected at one spatial scale are often used to make inferences about processes at a different spatial scale. Sometimes this mixing of scales results in incorrect or misleading results. The main purpose of this section is to describe two ways this can happen, each of which is so prevalent that it has a recognized name. The first is the *modifiable areal unit problem* (MAUP), which has to do with the effect of scale on conclusions about relationships between quantities. The second is the *ecological fallacy*, which is the use of aggregated data to draw conclusions about the behavior of individuals.

Data Set 4 includes measurements from a yield monitor that are separated by a distance of about 1 m in the direction of travel of the harvesting equipment and about 7 m in a direction perpendicular to the direction of travel. The data set also includes aerial images made up of pixels, each of which represents an area of about 9 m^2. We will use these data to demonstrate the MAUP. This term itself was first used by Openshaw and Taylor (1979), who note that it "is in reality two separate but interrelated problems" (Openshaw and Taylor, 1979, p. 128). The first problem, which they call the *scale problem* was recognized as early as 1934 by Gehlke and Biehl (1934). Cressie (1997) calls it the "change of support problem." The scale problem is, again quoting Openshaw and Taylor (1979, p. 128), "the variation in results that may be obtained when the same areal data are combined into sets of increasingly larger areal units of analysis." The most commonly observed manifestation of the scale problem involves the Pearson product moment correlation coefficient (Equation 11.2). Long (1996) demonstrates the scale problem with field crop data. He shows that the correlation coefficient between aggregated NDVI data and aggregated yield data depends on the scale over which the data are aggregated. We can demonstrate this same effect with data from Field 4.1. Figure 11.5a shows the boundary of the field, together with a rectangular area measuring 700 m north to south and 350 m east to west over which the NDVI values and the yield data are to be compared. The figure shows a set of fifty square cells over which the data are aggregated. It also shows a subset of the yield monitor data, which are housed in a `SpatialPointsDataFrame` called `data.Yield4.1`. The square cells are established by creating Thiessen polygons in the `SpatialPolygons` object `thsn.sp` (Section 3.5), after first using the function `expand.grid()` to create a set of points to use to define the polygons. The function `over()` of package `sp` is then used to aggregate the data by computing the mean over each polygon (see Section 6.4.2 for a discussion). We could also aggregate the data by using block kriging over each Thiessen polygon (Isaaks and Srivastava, 1989, p. 323), but simply computing the mean is simple, effective, and, works even when each block contains a relatively small number of data records. Here the object `img.rect` is a `SpatialPointsDataFrame` created from the May 1996 aerial image. The attribute data field `img.rect$NDVI` contains the NDVI computed according to Equation 7.1. Here is the computation of the mean over each polygon.

```
> data.ovrl <- over(thsn.sp, data.Yield4.1, fn = mean)
> image.ovrl <- over(thsn.sp, img.rect, fn = mean)
```

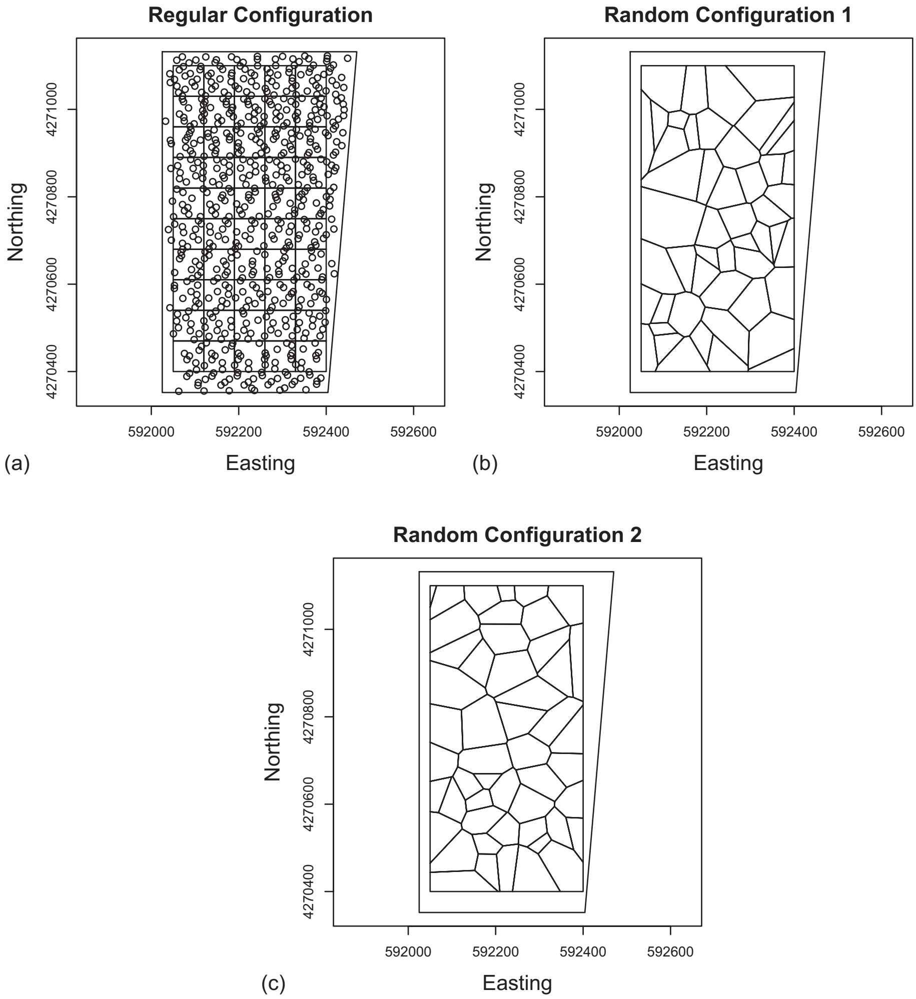

FIGURE 11.5
Illustrations of the modifiable areal unit problem: (a) aggregation of data into a regular array of square cells, illustrating the scale problem; (b) and (c) aggregation of data into two arrays of randomly arranged cells, illustrating the zoning problem. In Figure 11.5a, every 25th yield value is shown.

Finally, the correlation between the aggregated data is computed.

```
> print(cor(data.ovrl$Yield, image.ovrl$NDVI), digits = 3)
[1] 0.665
```

This code sequence is placed into a function `yield.vs.NDVI(cell.size)` to carry out the computations in sequence. The argument `cell.size` is expressed in terms of the width of the rectangle divided by the number of cells running from west to east. In addition to the correlation coefficient *r*, the function `yield.vs.NDVI()` returns the numerator and the denominator of the ratio that determines *r*. Here are the results.

```
> print(yield.vs.NDVI(350/20), digits = 2)
[1]   0.51  76.78 150.44
> print(yield.vs.NDVI(350/10), digits = 2)
[1]   0.59  79.71 135.95
> print(yield.vs.NDVI(350/5), digits = 2)
[1]   0.67  81.62 122.66
> print(yield.vs.NDVI(350/2), digits = 2)
[1]   0.86  85.96  99.71
```

The results indicate that as the cell size increases from 350/20 = 17.5 m to 350/2 = 175 m, the value of the correlation coefficient increases. This demonstrates a commonly observed phenomenon: as two data sets are aggregated over larger and larger areas, their correlation coefficient increases.

Wong (1995) gives a lucid explanation for the scale problem as it relates to the correlation coefficient. Referring to Equation 11.2, and examining the output just presented, we see that the covariance, which is the numerator of *r*, is relatively stable, increasing by about 11% over the range of areas. The denominator (the product of the standard deviations), however, decreases by about 50% over this range. Wong (1995) points out that it is often the case that as the aggregation area increases, a smoothing effect occurs in which the standard deviation decreases while the covariance remains relatively constant.

The second of Openshaw and Taylor's (1979) modifiable areal unit problems is called the *zoning problem*. This can also be demonstrated with our Field 4.1 data. Figure 11.5b and c show two Thiessen polygon sets generated for randomly placed locations in the data rectangle. The first polygon set was generated using the `spatstat` function `runifpoint()` following code as follows:

```
> E <- 592400
> W <- 592050
> S <- 4270400
> N <- 4271100
library(spatstat)
> ran1.ppp <- runifpoint(nrow(cell.ctrs), win = owin(c(W, E), c(S, N)))
> thsn.pp <- dirichlet(ran1.ppp)
> thsn.sp <- as(thsn.pp, "SpatialPolygons")
```

Carrying out the computation of the correlation coefficient, covariance, and product of the standard deviations yields the following:

```
> print(c(cor(data.ovrl$Yield, image.ovrl$NDVI),
+    cov(data.ovrl$Yield, image.ovrl$NDVI),
+    sd(data.ovrl$Yield) * sd(image.ovrl$NDVI)), digits = 2)
[[1]   0.76  97.82 128.42
```

Now we compute a second randomization.

```
> ran2.ppp <- runifpoint(nrow(cell.ctrs), win = owin(c(W, E), c(S, N)))
```

Repeating the computation yields the polygons of Figure 11.5c and the following coefficients:

```
> print(c(cor(data.ovrl$Yield, image.ovrl$NDVI),
+    cov(data.ovrl$Yield, image.ovrl$NDVI),
+    sd(data.ovrl$Yield) * sd(image.ovrl$NDVI)), digits = 2)
[1]   0.59  69.41 117.77
```

The correlation coefficients of the two configurations are 0.76 and 0.59. The covariance values are 97.82 and 69.41, and the denominator values are 128.42 and 117.77. Both the numerator and the denominator of Equation 11.2 are sensitive to the configuration of the polygons.

Now that the effects of the MAUP are apparent, the question arises as to what can be done about it. One possibility is to ignore it. Openshaw and Taylor (1981) consider three hypothetical justifications for this. The first is that it is insoluble, the second is that it is of trivial importance, and the third is that "to acknowledge its existence would cast doubts on the applicability of nearly all applications of quantitative techniques to zonal data" (Openshaw and Taylor, 1981, p. 67). They assert that the first two are incorrect, but that the third is true.

Wong (1995) discusses several approaches to a resolution of the MAUP. He divides these into three groups. The first approach, which is advocated by Moellering and Tobler (1972), involves computing the deviation from the mean of data aggregated at several scales. In the example of NDVI vs. yield in Field 4.1, these would be the regular Thiessen polygons. One then computes the sums of squares of these errors at each scale and assumes that the total of these sums of squares represents in some sense a total sum of squares. One then carries out significance tests for these sums of squares and considers the optimal scale to be the one which is most significant. Wong (1995) points out that the most serious problem with this method is that it does not rest on a very firm theoretical foundation. A second approach discussed by Wong (1995) is to attempt to construct a model, for example some form of parametric model that is immune to the effect of scale on the data relationships. The difficulty with this approach is that such a parametric model may not exist. The third approach is to construct a model that explicitly accounts for the spatial relationships among the errors. Once again, however, this approach may not be feasible.

Wong (1995) and Waller and Gotway (2004, p. 107) offer advice for dealing with the MAUP that we will summarize and interpret using the NDVI vs. yield example. Somewhat obviously, if appropriate scale invariant techniques are available to carry out the analysis, then these should be used. If such techniques are not available, then one must determine the objectives of the analysis and use statistics and a spatial scale that best fit these objectives. If the goal is to infer relationships at a fine scale, one should report the results of data analysis using fine scales. In our example, we would report the correlation coefficient between NDVI and yield as 0.51, the value at the smallest scale considered. This also has the virtue of being the most conservative. If the goal is to use NDVI to predict yield at the sub-field scale, it is appropriate to use a linear regression model over polygons of a size needed to generate a useful model. One should report the effects on the results of aggregating the data. In our example, this would involve reporting the values of the correlation coefficients at each of the scales. In dealing with the zoning effect, it would be reasonable, given a data set like that of Field 4.1, to analyze a large number of random arrangements of a fixed number of sample Thiessen polygon aggregation points similar to those in Figure 11.5b and c and report the mean and standard deviation of

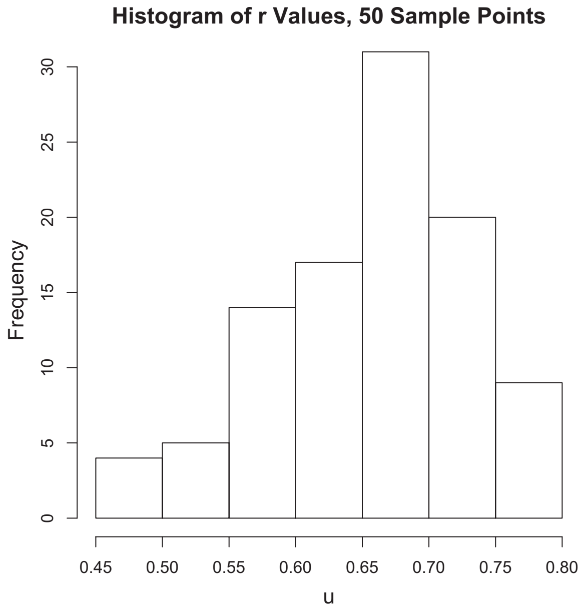

FIGURE 11.6
Histogram of values of Pearson's *r* between *Yield* and *NDVI* for 50 random arrangements of 50 polygons of the type shown in Figure 11.5b and c.

the correlation coefficient. Figure 11.6 shows a histogram of 100 random arrangements of 50 sample points for Field 4.1. The mean value of the correlation coefficient is 0.66, and the standard deviation is 0.075.

11.5.2 The Ecological Fallacy

The *ecological fallacy* is the fallacy that one may infer relationships at the individual level based on aggregated data collected about many individuals. Initial recognition of the ecological fallacy is attributed to Robinson (1950), although he did not use that specific term in the paper. Instead, he used the term *ecological correlation* to refer to a statistic describing a relationship that is defined for individuals but computed using aggregated data. An example from that paper can be used to illustrate the concept. Robinson (1950) presents statistics in the form of a contingency table in which the *X* variable is the number of foreign born vs. native born individuals in a sample of 97,272 US residents in 1930, and the *Y* variable is then number of literate vs. illiterate individuals in the sample. He shows that at the individual level, being illiterate is positively (albeit weakly) correlated with being foreign born. He then presents the same data, aggregated by region over nine regions in the US, and shows that when percent foreign born is related to percent illiterate, the correlation is negative. In the interest of full disclosure, it should also be pointed out that the relationship as presented in Figure 3 of Robinson (1950) is also highly nonlinear, so that the use of a correlation coefficient is probably not appropriate.

In any case, there are many instances where the process of ecological inference, which Haining (2003, p. 138) describes as the use of grouped data to infer individual relationships, breaks down. We have already seen an example of this phenomenon in Section 7.4 in the exploration the relationship between rice yield and silt content in Data Set 3 (Figure 7.20). This could be considered an example of the MAUP, but the distinction (possibly only a technical one) is that here there are identifiable spatial entities, the individual fields, whose data are aggregated. The ecological correlation between yield and silt content, that is, the correlation coefficient computed over the entire data set, has the value −0.61. Displaying the range of values of the individual field correlations provides the opportunity to demonstrate the use of the function by().

```
> Data <- with(data.Set3, data.frame(Yield, Silt))
> Field <- data.Set3$Field
> print(sort(by(Data, Field, function(x) cor(x)[1,2])[1:16]),
+ digits = 2)
Field
     8      7     15      2      1     14      9      4     13
-0.709 -0.630 -0.250 -0.246 -0.238 -0.183 -0.142  0.081  0.106
     5     12      6      3     16     11     10
 0.113  0.201  0.208  0.331  0.440  0.684  0.722
>
```

There is no evident pattern to the relationships between yield and silt content at the field scale.

The third line of the code sequence provides a good opportunity to review some R programming concepts. The first argument of the function `by()` is `Data`, a data frame whose columns are the quantities we wish to compare. The second argument, `Field`, is the index defining the groups of data. The function `cor()`, applied to these data, produces a two by two correlation matrix whose diagonal elements are unity and whose off-diagonal elements are the correlation coefficients. The third argument, `function(x) cor(x)[1,2]`, is a function whose argument is a two by two matrix and whose return value is one of the off-diagonal elements. Thus, the function `by()` returns an array of correlation coefficients of the data by field. The function `sort()` sorts this array from most negative to most positive.

The term *ecological bias* is also used to describe the difference between a statistical result based on data observed at one geographic scale and that based on the same data at a different scale (Greenland and Morgenstern, 1989). One reason that data may display an ecological bias is that individuals, or data measured at a small scale, may encompass less variability than the same data measured at a larger scale, and the variability at the larger scale may be necessary for it to manifest the relationship in question. For instance, Hill et al. (1992) describe a strong relationship in rice between grain moisture content at harvest and percent fractured grains. This observation is based on aggregated data over many fields. At the individual field scale, however, such variation is not necessarily seen (Marchesi, 2009). Thompson and Mutters (2006) have found that the relationship between grain moisture content at harvest and percent fractured grains in rice depends in a complex way on meteorological conditions during the period leading up to harvest. The difference in observations between the regional and field scale is probably due to the fact that the regional observations encompass a variety of meteorological conditions, whereas these conditions are more or less fixed for the harvest of an individual field. This is related to a statistical phenomenon called Simpson's paradox (Cressie, 1997).

11.6 Further Reading

Almost all statistics texts discuss the Pearson correlation coefficient, but the discussion by Snedecor and Cochran (1989) is particularly comprehensive and easy to understand. Haining (2003) provides an excellent discussion of the method of Clifford et al. (1989). Alternative methods have also been proposed (Dutilleul, 1993).

The books by Agresti (1996, 2002) are the standard references for categorical data analysis. Freeman (1987) also provides a good basic introduction to the subject. Thompson (2009) provides an excellent discussion of the use of R in the analysis of problems involving categorical data.

Cressie (1997) discusses the MAUP in geostatistical terms and provides an excellent discussion of its relationship to other statistical phenomena. Gotway and Young (2002, 2007) give an extensive discussion with many references of the MAUP and the ecological fallacy, and of approaches to its resolution. Openshaw (1984) provides an extensive discussion of the ecological fallacy and provides many references. Haining (2003) provides an excellent discussion of the analysis of aggregated data. For further discussion of the ecological fallacy, see Duncan et al. (1960), Alker (1969), and Langbein and Lichtman (1978).

For more on the Theory of the Stork, see Höfer et al. (2004).

Exercises

11.1 Use the Clifford correction and the parametric bootstrap to adjust the correlation coefficient between *LeafN* and *GrainProt* in Data Set 4, Field 1.

11.2 For the data of Data Set 2, using the definitions of Section 11.3.1, repeat the analysis carried out in that section, but instead of using $\{MAT \leq 15\}$ as the X variable, use $\{JuMax \leq 30\}$.

11.3 An alternative question relating to temperature to that posed in Section 11.3.1 is whether especially cool winters might impede blue oak establishment due to reduced seedling survival. As a preliminary means of investigating this issue, consider for the Coast Range only those sites that are on opposite sides of the median in their values of *JaMean*, the mean January temperature and *JuMean*, the mean July temperature. Let X_1 represent the property of a site having a value of *JaMean* below the median of all values of *JaMean* and a value of *JuMean* above the median of all values of *JuMean* (cool winters and hot summers), and let X_2 represent the property of a site having a value of *JaMean* above the median of all values of *JaMean* and a value of *JuMean* below the median of all values of *JuMean* (warm winters and warm summers). Again, let $Y = Y_1$ represent the presence of blue oak at the site and $Y = Y_2$ represent the absence. For the data of Data Set 2, using the definitions of Section 11.3.1, compute the contingency tables of WW-WS vs. CW-HS for the Coast Range and carry out a significance test both with and without accounting for spatial autocorrelation.

11.4 Construct scatterplot matrices of *Precip*, *JaMean*, *JuMean*, and *Elevation* for the Sierra Nevada and Coast Range subsets of Data Set 2. Do you see any striking differences?

11.5 Create a binomial variable in Data Set 2 that is 1 when *Precip* > median{*Precip*} and 0 otherwise. Create a contingency table for this variable and CW-HS. What does this tell you about the results of Section 11.3.1?

11.6 With the Sierra Nevada and Coast Range subsets of Data Set 2, construct scatterplots of *Precip* vs. *Elevation*, *MAT* vs. *Elevation*, and *Precip* vs. *MAT*, using different symbols for *QUDO* = 1 and *QUDO* = 0. The plots are easiest to visualize if you only plot every fourth point.

11.7 One might wonder why in Section 11.3.2 the factor 0.001 is subtracted from the computed value `obs.chisq` in computing the *p* value based on repeated evaluation of the observed chi-square statistics. Display the computed value of these statistics in a permutation test and use the function `unique()` to find all the possible values that the array `u` of permutation values takes on. Can you use this information to answer the question?

11.8 Repeat the block restricted randomization for the permutation test for the subsets of Data Set 1 determined in Exercise 7.4 to be important. What can you conclude about the effect of spatial autocorrelation on the conclusions drawn from the analysis?

11.9 Compute the partial Mantel statistic for the relationship between log SPAD meter reading vs. log Grain N. What can you say in an ecological context about the comparison between these results and those obtained for the relationship between log SPAD meter reading and log Leaf N?