4

Measures of Spatial Autocorrelation

4.1 Introduction

The previous chapter introduced the definition of positive spatial autocorrelation as a nonzero covariance between spatial proximity and attribute value proximity (i.e., nearby values are more similar than they would be if they were arranged randomly). This definition imposes the need for a means to measure, based on a sample of data values, the covariance between nearby points and to decide whether or not this covariance is consistent with a random spatial arrangement of values. We have to define a statistic and establish a null hypothesis concerning the value of that statistic when no spatial autocorrelation exists, and an alternative hypothesis concerning the value when spatial autocorrelation does exist. Consider the soil sand content data of Section 3.6. In that section, a model called the spatial error model was fit to the detrended sand content data of Field 4.1 of Data Set 4 (denoted Field 4.1). The parameter λ of the model was estimated in this process. One could measure and test autocorrelation using this parameter, and indeed we shall have occasion to do something like this in Chapter 13. However, the validity of this test requires that the data satisfy the spatial error model. In many applications there is considerable advantage to having a statistic that does not depend on a particular model of the autocorrelation structure. In this chapter, we introduce a collection of such statistics that are used to measure the strength of spatial autocorrelation. After some preliminary discussion in Section 4.2, Section 4.3 discusses tests for spatial autocorrelation of categorical data and Section 4.4 discusses tests for spatial autocorrelation of quantitative data. Section 4.5 discusses measures of autocorrelation structure within a data set. Each of these sections is concerned with areal data. Section 4.6 provides a brief discussion of measures of autocorrelation of continuous (geostatistical) data.

4.2 Preliminary Considerations

4.2.1 Measurement Scale

Figure 4.1a shows a plot of the numerical value of weed infestation level in Field 4.1. These data were obtained by having an expert agronomist walk through the field and at each sample location assess the weed infestation level on a scale of 1 = no infestation to 5 = complete weed cover. Figure 4.1b shows a map of detrended sand content in the same field. This is a modification of Figure 3.4b in which the actual numerical values of the data are shown.

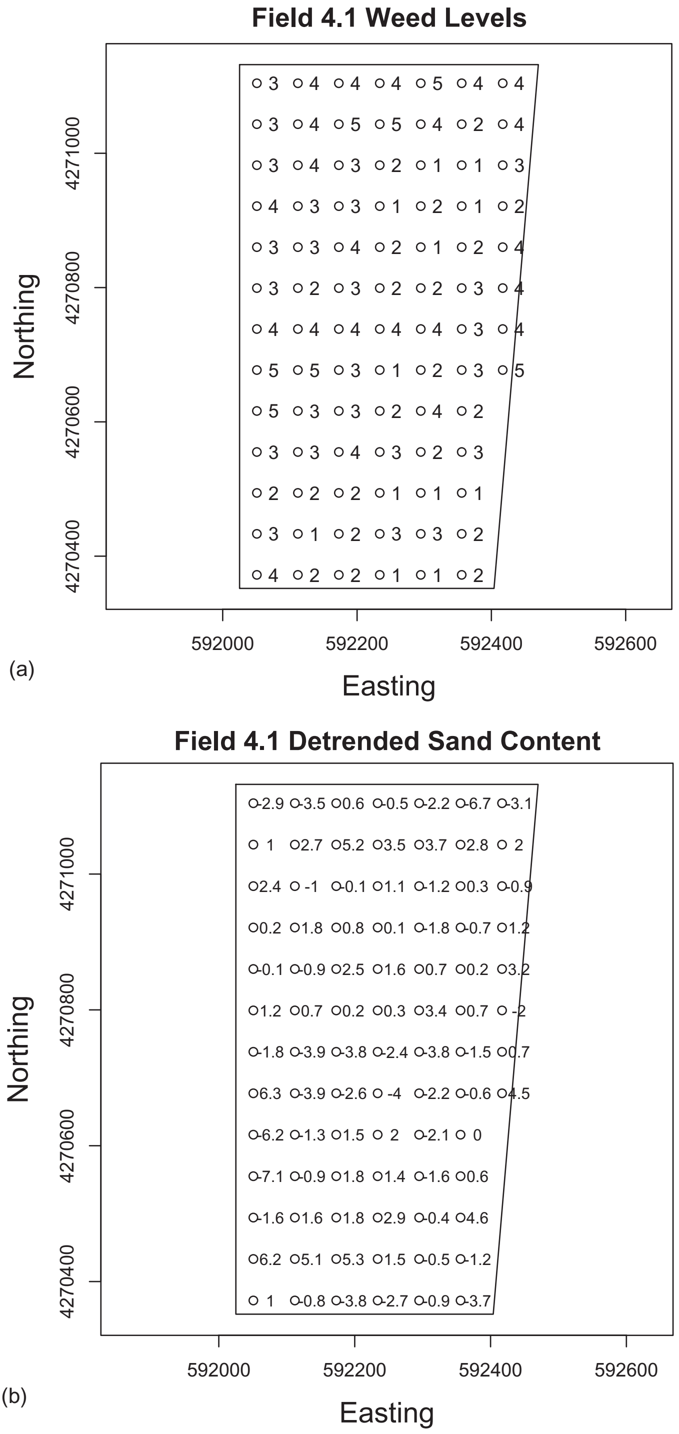

FIGURE 4.1
(a) Location map of numerical values of *Weeds* in Field 1 of Data Set 4; (b) Location map of numerical values of detrended sand content in Field 1 of Data Set 4.

These data sets are superficially similar, but there is an important difference. The sand content values in Figure 4.1b are represented by numbers, and some mathematical operations on these numbers are meaningful; for example, the difference between a detrended sand content of 2% and one of 1.5% is the same as the difference between 0% and −0.5%. The weed levels in Figure 4.1a are rankings, and although they, like the data representing sand content, are expressed as numbers, the interpretation of these numbers is quite different. The weed level rating (5, 4, 3, 2, 1) could also be represented, for example, with the sequence (10, 6.2, 0, −3, −25), and there would be no difference in interpretation. An analogous change in the sand content data, however, would completely change the interpretation.

The differences between these data types can be understood through the concept of *measurement scale* (not to be confused with spatial scale). A measurement can be defined as "the assignment of numerals to objects or events according to rules," and a scale in this context is, in the most general sense, a "means of dealing with aspects of these objects" (Stevens, 1946). Measurement scales may be classified according to whether or not operations such as equality comparison, ordering, and arithmetic are meaningful. The classification of measurement scales has been discussed by many statisticians, but the most generally recognized classification, which we will use in this book, is due to Stevens (1946). In this classification a measured quantity falls into one of the following four categories.

Nominal scale data. These are data for which the measurement defines membership in a category, but there is no implied relationship among categories. Familiar examples include soil type (as represented by set of numbers, one for each soil type), telephone number, and postal code. The only meaningful mathematical operation that can be carried out with nominal scale data is equality comparison: any two data values can be either equal or not equal. Thus, for example, if the number 27 represents the soil type *Rincon silty clay loam*, then all data records assigned the number 27 would be this soil type. A special case of nominal scale data is *binary data*, in which there are two categories. These categories can be represented by the number zero and one. There is an implied order in the relationship in that one is larger than zero, but because there are only two possible values the grouping of data by order is mathematically equivalent to grouping by equality comparison.

Ordinal scale data. These are data for which a rank order can be established, but differences between ranks do not have meaning. The weed infestation level data shown in Figure 4.1a are an example. In addition to equality comparison, an ordering can be established, but an operation on data values such as subtraction of one from another has no meaning. Thus, for example in the case of the weed level classes of Figure 4.1a, one cannot infer that the difference between a level 3 and a level 2 weed level is the same as the difference between weed levels 2 and 1.

Interval scale data. These are data for which the operations of addition and subtraction have meaning but, because no absolute datum is established, multiplication and division do not. Temperature measured on either a Celsius or Fahrenheit scale, is an example. We cannot say that 70°C is "twice" 35°C because if we convert the measurement units to °F, for example, the resulting temperatures (158°F and 95°F, in this case) do not have the same numerical relationship. The detrended sand content data are interval scale data.

Ratio scale data. These are data in which a datum exists, so that the operations of multiplication and division produce meaningful results. Population density is an example, as is the original sand content data set of Figure 3.4a. It is meaningful to say that if one location has a 20% sand content and a second location has a 40% sand content, then the sand content at the second location is twice that at the first.

From the perspective of measurement of spatial autocorrelation, there is no need to distinguish between interval and ratio scale data, and indeed we shall see that ordinal scale data are often, although not always, handled in the same way as well (this is most justifiable if the ordinal scale data take on values that are in some sense "analogous" to interval or ratio scale values). Thus, we will generally distinguish primarily between nominal (or "categorical") data and "quantitative" data. The term "quantitative" subsumes the categories of interval, ratio, and sometimes ordinal data.

4.2.2 Resampling and Randomization Assumptions

By carrying out statistical analysis of the data, we implicitly assume that these data represent a realization of a random field, that is, a set of random variables whose realized values are mapped onto (in our case) a two-dimensional surface. It turns out that the statistical properties of the sample depend on how this realization takes place. Following Cliff and Ord (1981, p. 14), we will allow four different possibilities. These are defined according to two dichotomous criteria. The first criterion is the method by which values would be assigned under the null hypothesis to the n locations (polygons or points) in successive hypothetical "experiments" (if you have forgotten why "experiments" is in quotes, see Section 3.3). The second criterion is the distribution of the population from which the values in the sample are drawn, which for all of our examples will be either binary or normal.

Regarding the first criterion, one way to carry out the assignment of values to locations is in each "experiment" to draw n random samples with replacement from the population and assign one of these values to each location. This is called the *resampling assumption*. The second way is to draw n samples from the population and then to randomly reassign these fixed numbers one to each location in each "experiment." This is called the *randomization assumption*. Let us look at each of these more closely.

Resampling assumption. Under this assumption, the n attribute values Y_i (e.g., the values of sand content in Figure 3.4a) in the sample are obtained from n independent drawings from a population having a given distribution. The null hypothesis states that each cell is assigned a value equal to a random number drawn from the population independently of the other cell assignments. For example, if time could be reset and another hypothetical drawing of sand content values in Figure 3.4a could be generated, they would be drawn from the same population as the first but would be independent of the values from the previous sample. If the data are assumed to be drawn from a normal population, they are sometimes said to follow a *normality* assumption (Cliff and Ord, 1981, p. 14). If the distribution is binomial, then the probability that a cell has the value one is p and the probability that a cell has the value zero is $q = 1 - p$. This case is sometimes called the *free sampling* assumption, and can be considered as sampling *with* replacement from a population with $p \times n$ ones and $q \times n$ zeroes.

Randomization assumption. If the data are assumed to follow a resampling assumption as defined in the previous paragraph, then each realization will in general produce a completely different attribute data set. Thus, the values assigned to the polygons or points will generally be different from one realization to the next.

> In the case of binary data under the resampling assumption, the number of ones and zeroes will generally be different in each realization since they are produced by n independent draws from a binomial distribution.

In contradistinction to this, the randomization assumption considers only the n attribute values observed in the sample. For example, in the hypothetical second drawing of detrended sand content data in Figure 3.4a, the values would be the same in each realization but the locations of these values would be randomly rearranged and independent of the locations from the previous drawing. In the case of normally distributed data the observed sample is considered to be one of the $n!$ (this is read as "n factorial" and is equal to $n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$) possible arrangements in which the n observed numbers could be assigned to the n cells. The randomization occurs in the assignment to the n geographical locations. In the case of binary data, if k of the n sites have the value one, then the null hypothesis is that the observed pattern is a random arrangement of k ones and $n-k$ zeroes, considering that all $\binom{n}{k}$ (this is read "n choose k" and is equal to $n!/k!(n-k)!$) possibilities are equally likely. This special binary case is sometimes called *nonfree sampling*, and can be considered as sampling *without* replacement from a population with $p \times n$ ones and $q \times n$ zeroes. Unfortunately, the terms used to characterize the assumptions mostly begin with N or R, and this can cause confusion. Table 4.1 can be used for ready reference.

One must decide which assumption makes more sense for a particular data set. Goodchild (1986, p. 23) expresses the opinion that the resampling assumption "is more reasonable in most contexts." In general, unless otherwise stated, we will assume that the data follow the resampling assumption. There is, however, one very important exception. The very useful permutation-based hypothesis tests described in Section 4.3 require the randomization assumption, and when we use a permutation test we will be implicitly assuming that the data follow this assumption. As usual, the most sensible course, where possible, is to carry out the analysis under both assumptions and, if they are substantially different, try to understand why.

4.2.3 Testing the Null Hypothesis

We are now in a position to address the question posed at the start of this chapter, namely, whether or not the data in a set of observations may be considered to be spatially autocorrelated based on their observed spatial distribution. This question is formulated in terms of a hypothesis test. The null hypothesis is that the data are randomly assigned to locations according to one of the two assumptions described in Section 4.2.2, the resampling assumption or the randomization assumption. We compute a test statistic and the probability of observing a value of this statistic at least as extreme as the one we actually observe (the p value). If this probability is very low, then we reject the null hypothesis that the data are randomly arranged by location. The evidence in this case indicates that the data display some level of spatial autocorrelation.

TABLE 4.1

Terminology Used to Characterize the Various Assumptions about the Behavior of the Data under the Null Hypothesis of Zero Autocorrelation

Assumption	**Resampling *N***	**Randomization *R***
Binary Data	Free Sampling	Nonfree Sampling
"Numerical" Data	Normality	Randomization

Referring to the notation given above, a very general form for a statistic to be used in a test of the null hypothesis of zero autocorrelation is the *Mantel* statistic (Mantel, 1967; Mantel and Valand, 1970; Anselin, 1995), given by

$$\Gamma = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}c_{ij} \tag{4.1}$$

Here w_{ij} is the ij^{th} element of the spatial weights matrix and represents the contiguity between locations i and j, and c_{ij} is a quantity that indicates how much statistical weight is assigned to the pair ij. Different ways of computing c_{ij} characterize different autocorrelation statistics. We will initially restrict ourselves to a few special forms that lead to widely used measures of autocorrelation. Cliff and Ord (1981) show that each of the forms of Γ we consider, under both the resampling and randomization assumptions, is asymptotically normally distributed as n becomes very large. Because of this, one method of testing the null hypothesis is to assume that the random quantities are normally distributed and use statistical tests based on this assumption, such as the *t*-test. However, there may be a concern that the actual sample size is not sufficiently large that asymptotic theory is justified. In the next sections, we will look at the application of various forms of Mantel statistics to specific cases.

4.3 Join–Count Statistics

Join–count statistics may be used to test the null hypothesis of no spatial autocorrelation for nominal data. The properties of the distribution of the join–count statistic were worked out by Moran (1948) and are discussed in detail by Goodchild (1986, p. 37). We restrict discussion to binary data, which is the most common case (for the more general, "multicolor" case, see Cliff and Ord, 1981, p. 19). Figure 4.2 shows the data of Figure 4.1a with weed levels 4 and 5 combined into one class called "High" and the remaining values combined into the class "Low." The data are represented by two colors, black (actually, dark gray so the borders are visible) and white (actually, light gray). Define a *join* to be a nonzero contiguity between spatial objects (points or polygons) i and j as defined by a nonzero spatial weights matrix element w_{ij}. In the case of a mosaic of polygons, a join is a boundary under the contiguity rule used in the analysis (rook's case or queen's case, see Section 3.5.2) between two cells. To avoid constant repetition, we will from now on refer to the objects as polygons, recognizing that what is said also applies to points if an appropriate spatial weights matrix is used.

Assume the two attribute values are denoted B, for black, and W, for white (be careful not to confuse this use of W with that of the spatial weights matrix). Define J_{rs} as the sum of weighted joins between color r and color s, where $r, s = B, W$, and the weight of the join between polygon i and polygon j is the value of the element w_{ij} of the spatial weights matrix W. We will use the statistics J_{BB}, J_{WW}, and J_{BW} to test the null hypothesis that the data have zero autocorrelation. A two-sided test would imply that the null hypothesis of zero autocorrelation is being tested against the alternative that the data are either positively or negatively autocorrelated, but the analyst does not know which. This situation is unlikely to arise in practice, since ordinarily the physical or biological characteristics of the data would favor either one alternative or the other. Therefore, a one-sided alternative is generally used.

Cliff and Ord (1981, p. 11) give formulas for computing J_{BB}, J_{WW}, and J_{BW}. Let $Y_i = 1$ if polygon i is black and 0 if it is white. Then the formulas for J_{BB} and J_{BW} are

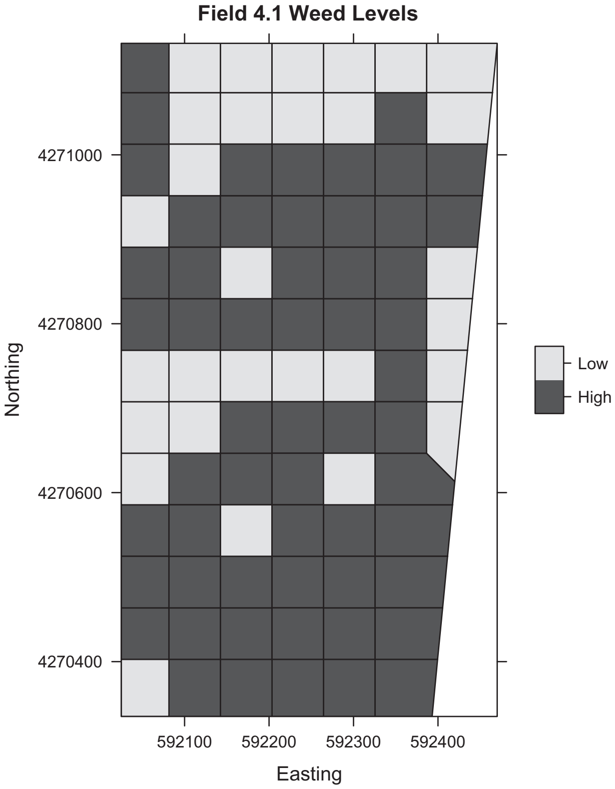

FIGURE 4.2
Thematic map of a binary interpretation of the weed levels in Field 1 of Data Set 4. Values of *Weeds* greater than or equal 4 are classified as High and values between 1 and 3 are classified as Low.

$$J_{BB} = \frac{1}{2}\sum_i \sum_j w_{ij} Y_i Y_j,$$
$$J_{BW} = \frac{1}{2}\sum_i \sum_j w_{ij} (Y_i - Y_j)^2, \tag{4.2}$$

where the identity $w_{ii} = 0$ is required for all i. The value of J_{WW} is $J - J_{BB} - J_{BW}$ where J is the total number of weighted joins. In order to test parametrically the null hypothesis of zero autocorrelation, one must compute the expected value and variance of the statistic under the resampling (free sampling in the case of binary data) or randomization (nonfree sampling) assumption. These are worked out by Cliff and Ord (1981, p. 19). The complexity of the general formulas greatly outweighs their usefulness, and so we will not show them but instead refer the interested reader to the cited reference.

The distribution theory of join–count statistics is discussed by Cliff and Ord (1981, p. 36). Under the free sampling model the probabilities that a cell has the value one or zero are distributed binomially, whereas under the nonfree sampling model they follow a hypergeometric distribution. The distributions are asymptotically normal as n approaches infinity (Cliff and Ord, 1981, p. 51–54). Cliff and Ord (1981, p. 63) carry out Monte Carlo simulations to show that the normal approximation is reasonable even for moderate values of n.

For any one map there are three join–count statistics that can be used to test the null hypothesis of zero autocorrelation, namely, J_{BB}, J_{WW}, and J_{BW}. We will focus on the *BB* and *WW* joins, since these are available in R. When invoking the asymptotic normality of the statistic, the join–count test is carried out as a *t*-test using the statistic

$$t = \frac{J_{rr} - E\{J_{rr}\}}{\sqrt{Var\{J_{rr}\} / n}}, \tag{4.3}$$

where $E\{J_{rr}\}$ and $\text{var}\{J_{rr}\}$ are the expectation and variance under the null hypothesis of zero autocorrelation using the formulas given by Cliff and Ord (1981).

Let us apply these formulas to the data represented in Figure 4.2. The contents of the file *Set4.196sample.csv* are loaded into the data frame `data.Set4.1` using the code in Appendix B.4. This creates a spatial feature file `thsn.sf`, which is then coerced in to a `SpatialPolygonsDataFrame` called `thsn.sp` (see Section 2.4.3). Thiessen polygons can be created as described in Section 3.6. The polygons in Figure 4.2, which are topologically equivalent but match the nonrectangular boundary of the field, were created in ArcGIS and are in the file of auxiliary data on the book's website.

The R function `joincount.test()` in the package `spdep` (Bivand et al., 2011) can be used to compute the expected values as well as the significance level under the normal approximation of the nonfree sampling model for *BB* and *WW* joins. We will take this opportunity to utilize two functions, although they make the code slightly more complex. For the test of the data in Figure 4.2, we first use the function `lapply()` to create a list (Section 2.4.1) in which the logical test $Weeds \geq 4$ is successively applied to each data record. The function `unlist()` then converts this list into a vector. The second statement creates a data field `HiWeeds` in the Thiessen Polygons, coverts the results of the logical test into a vector of factors, and creates labels for the factors.

```
> HiWeeds <- unlist(lapply(data.Set4.1$Weeds,
+     function(x) (as.numeric(x >= 4))))
> thsn.sp@data$HiWeeds <- factor(HiWeeds,
+   labels = c("High", "Low"))
```

(For practice, try implementing the two steps `lapply()` and `unlist()` separately and observing the output.) Next, we create a spatial weights matrix using the methods discussed in Section 3.5.2 and apply the function `joincount.test()`. In this implementation, we use a rook's case binary contiguity rule.

```
> nlist <- poly2nb(thsn.sp,
+   row.names = as.character(thsn.sp$ID), queen = FALSE)
> W <- nb2listw(nlist, style = "B")
> joincount.test(thsn.sp$HiWeeds, W)
             Join count test under nonfree sampling
data:  thsn.sp$HiWeeds
weights: W

Std. deviate for High = 2.9134, p-value = 0.001787
alternative hypothesis: greater
sample estimates:
Same colour statistic    Expectation          Variance
             70.00000       59.82271          12.20276
```

```
        Join count test under nonfree sampling
data:  thsn.sp$HiWeeds
weights: W

Std. deviate for Low = 3.4465, p-value = 0.0002839
alternative hypothesis: greater
sample estimates:
Same colour statistic     Expectation          Variance
            25.000000       15.218057          8.055317
```

The alternative tested is that the population parameter is greater than the observed value. The function returns the results for *BB* joins (value 1) and *WW* joins (value 0), and both indicate rejection of the null hypothesis, although with slightly different p values.

Permutation tests of join–count can be carried out using the spdep function joincount.mc(). The full output is not shown, only the line with the p values.

```
> set.seed(123)
> joincount.mc(thsn.sp$HiWeeds, W,1000,
+ alternative = "greater")
      Monte-Carlo simulation of join-count statistic
Join-count statistic for High = 70, rank of observed statistic = 997,
p-value = 0.003996

Join-count statistic for Low = 25, rank of observed statistic = 999,
p-value = 0.001998
```

Simulations with different random number seeds will produce slightly different results, but in general they should closely approximate the results obtained with joincount.test(). In this case, the permutation tests of both J_{BB} and J_{WW} once again indicate rejection of the null hypothesis. Table 4.2 shows the results of t tests and permutation tests for the data in Figure 4.2 for binary and row normalized spatial weights matrices. There is quite a bit of difference among the tests, with the binary queen's case contiguity rule being the most conservative.

For any data set, three different join–count statistics, J_{BB}, J_{WW}, and J_{BW}, are available to test a single null hypothesis, and it is very possible to draw completely different conclusions depending on which statistic is tested. This is especially true with relatively small data sets (Exercise 4.1). Which statistic should be used to interpret the significance level? Cliff and Ord (1981, p. 63) report on extensive Monte Carlo simulations under a wide variety of conditions. As is generally the case in the comparison of test statistics, the comparison by Cliff and Ord focuses on the power of the test using each statistic (Section 3.4).

TABLE 4.2

Results (p values) of Tests of the Null Hypothesis of Zero Autocorrelation for the Maps of Figure 4.2

	Rook's Case				Queens's Case			
	Binary		Row Norm.		Binary		Row Norm.	
	t test	permut.	*t* test	permut.	*t* test	permut.	*t* test	permut.
J_{WW}	0.002	0.004	<0.001	0.001	0.037	0.054	<0.001	0.001
J_{BB}	<0.001	0.002	<0.001	0.001	0.007	0.011	<0.001	0.001

The conclusions of Cliff and Ord may be summarized as follows. Both the randomization test and the normal approximation appear to apply reasonably well, so an observed large difference in the results from the functions `joincount.test()` and `joincount.mc()` should be interpreted as a sign of possible problems with the data, such as a failure to satisfy the assumptions of normality, or possibly an indication that `joincount.mc()` should be evaluated again with a larger number of permutations. Second, the color with the larger number of cells generally produces the more accurate result, so this *p* value should be used to interpret significance. This would be J_{WW}, the statistic for low weed level, in our case. A final conclusion (Cliff and Ord 1981, p. 63) is that the statistic J_{BW} measuring different colored joins often has a higher power than statistics measuring joins of the same color. There is a more complex `spdep` function `joincount.multi()` that permits the testing of *BW* joins as well as spatial objects with more than two colors. Finally, the functions `joincount.test()` and `joincount.mc()` both have the capability to extend the join–count statistic for same color joins to maps with more than two colors.

4.4 Moran's *I* and Geary's *c*

The Moran's *I* (Moran, 1948) and Geary's *c* (Geary, 1954) statistics were both developed to test the null hypothesis of zero autocorrelation with interval or ratio data. Although not developed for use with ordinal data, these statistics often work better than the alternatives with this measurement scale as well. As with the join–count statistic, the distribution theory of *I* and *c* is extensively discussed by Cliff and Ord (1981) as well as by Goodchild (1986) and Upton and Fingleton (1989). Both *I* and *c* can be expressed as special cases of the general Mantel statistic introduced in Equation 4.1 and defined by the particular choice of the attribute similarity measure c_{ij}. Moran's *I* is written

$$I = \frac{n}{S_0}\frac{\sum_i\sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2}, \tag{4.4}$$

where $S_0 = \sum_i \sum_j w_{ij}$. The similarity measure c_{ij} in Equation 4.1 can therefore be written

$$c_{ij} = \frac{n}{S_0}\frac{(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2}, \tag{4.5}$$

showing that Moran's *I* bears a strong conceptual resemblance to the Pearson correlation coefficient (Appendix A). Geary's *c* may be written

$$c = \frac{n-1}{2S_0}\frac{\sum_i\sum_j w_{ij}(Y_i - Y_j)^2}{\sum_i (Y_i - \bar{Y})^2}, \tag{4.6}$$

so that the similarity measure in this case is

$$c_{ij} = \frac{n-1}{2S_0}\frac{(Y_i - Y_j)^2}{\sum_i (Y_i - \bar{Y})^2}. \tag{4.7}$$

Thus, the similarity measure underlying Geary's c is a squared difference, analogous to the variogram measure, which will be discussed in Section 4.6.1, for data on a spatial continuum.

The similarity of appearance of Moran's I to the Pearson correlation coefficient r might lead one to believe that the I statistic would display similar behavior to r. However, while the behavior of I is somewhat analogous to that of r, there are important differences. The primary difference is that the expected value of I under the null hypothesis of zero spatial autocorrelation is not zero but rather $-1/(n-1)$ (Cliff and Ord, 1981, p. 42). Nevertheless, increasingly positive values of I are associated with increasingly strong positive autocorrelation, and negative values with negative autocorrelation. The expected value of Geary's c under the null hypothesis of zero spatial autocorrelation is one. A value of c between zero and one indicates positive autocorrelation, and a value $c > 1$ indicates negative spatial autocorrelation (note from the form of Equation 4.6 that c cannot be negative).

Moran's I and Geary's c are both asymptotically normally distributed under the normality assumption as n increases (Cliff and Ord, 1981, p. 47). Thus, as with the join–count statistics, the null hypothesis of zero autocorrelation may be tested by computing the expected value and variance of the statistic under normality and then testing via a t-test. Alternatively, one may employ a permutation test under the randomization assumption.

In summary, the possible ways that we have discussed that one may carry out a test of the null hypothesis of zero spatial autocorrelation using the Moran's I or the Geary's c statistic are as follows. First, that data may be modeled as points or polygons. Second, the data may be assumed to be distributed under the null hypothesis according to the normality assumption or the randomization assumption. Third, one may compute the value of the statistic and its variance under either assumption and compute the p value according to asymptotic normality, or one may carry out a permutation test (in which case the randomization assumption is implicitly made). Finally, one can carry out any of these calculations using any one of the forms for the spatial weights matrix W. Thus, a number of p values can be computed.

We can compare some of these methods using calculations involving the percent silt data discussed in Field 4.1. We use the silt data because it has little or no trend (this will be shown in Section 4.6.1) and there are issues (these will be discussed at the end of this section) with applying these tests to detrended data. Here is the code for the Moran test and the Geary test for distance based, rook's case point data under the normality assumption.

```
> nlistk <- knn2nb(knearneigh(data.Set4.1, k = 4))
> W.kW <- nb2listw(nlistk, style = "W")
> W.kB <- nb2listw(nlistk, style = "B")
> Silt <- data.Set4.1@data$Silt
> moran.test(Silt, W.kW, randomisation = FALSE, alternative = "greater")
         Moran's I test under normality
data:  Silt
weights: W.kW
Moran I statistic standard deviate = 2.2547, p-value = 0.01208
alternative hypothesis: greater
sample estimates:
Moran I statistic          Expectation            Variance
      0.371990105         -0.011764706         0.006482012
```

TABLE 4.3

Values and Variances of Moran's I and Geary's c for the Field 4.1 Silt Data Under the Resampling Assumptions Using Row-Normalized (W) and Binary (B) Spatial Weights Matrices for I and c Using Rook's Case and Queen's Case Contiguity Rules

	Point Data			**Polygon Data**			
	Binary			**Rook's Case**		**Queen's Case**	
	Value	***p.***		***Value***	***p.***	***Value***	***p.***
I_d	0.372	0.000	I_B	0.364	0.000	0.326	0.000
I_k	0.369	0.000	I_W	0.371	0.000	0.330	0.000
c_d	0.624	0.000	c_B	0.599	0.000	0.636	0.000
c_k	0.606	0.000	c_W	0.623	0.000	0.670	0.000

Note: For the point data, the subscript d refers to d distance based weights and the subscript k refers to k nearest neighbor weights and only the binary data are shown. Results under the randomization assumption are virtually identical and are not shown.

Table 4.3 shows results of the test of the null hypothesis of zero autocorrelation against the alternative of positive autocorrelation using the point location model. Results are shown for the four combinations of neighbor relationship (distance = 61 m and four nearest neighbors) and weight coding (binary coded and the row-normalized). Permutation tests displayed very similar results (not shown). All of the tests for point data indicate significant spatial autocorrelation in the data.

The I and c statistics may be used for ordinal as well as ratio and interval scale data. Cliff and Ord (1981, p. 15) point out that for ordinal data either the multicolor join–count statistic or the I or c statistic may be used, but that latter are preferable because they preserve the ordering of the relationship. Moreover, neither statistic is very sensitive to departures from normality.

Since two statistics are available, Moran's I and Geary's c, the natural question arises as to which should be preferentially used. Upton and Fingleton (1989, p. 170) point out that because of the form of Equations 4.4 and 4.6, c would be expected in general to be more sensitive to differences in values of Y while I would be more sensitive to extreme Y values. Cliff and Ord (1973, p. 45) suggest that I is less affected by the distribution of the data than is c. Cliff and Ord (1981, Chapter 6) carry out a detailed comparison of I and c focusing on two methods of comparison: an analytical method and Monte Carlo simulation.

The null and alternative hypotheses considered by Cliff and Ord (1981, p. 167) in their comparison of the two statistics are $H_0 : I = -1/(n-1)$ versus $H_a : I \neq -1/(n-1)$ and $c = 1$ versus $c \neq 1$. For these tests, Cliff and Ord compute the *asymptotic relative efficiency*, which is (roughly speaking) the limit as n approaches infinity of the power of two tests. They show that for the binary weights matrix the asymptotic relative efficiency c is always less than that of I. Perhaps somewhat more convincingly, Cliff and Ord (1981, p. 175) report on a series of Monte Carlo simulations that indicate that the statistical power of I is greater than that of c. They conclude that I is to be preferred over c. Griffith and Layne (1999, p. 15) provide additional comparison and also conclude that the Moran's I is generally preferable, although corroboration of Moran's I results by the computation of Geary's c is desirable. By displaying the mathematical relationship between the two, Griffith and Layne (1999, p. 15) show that a difference between the two might be observed in cases of highly irregular lattice cell sizes and outliers in the Y_i values. As always, the wise course is to compute both statistics. If they indicate similar properties, one breathes a sigh of relief and moves on. If they indicate dissimilar properties, one must try to determine the reason for this, and, ultimately, report the dissimilarity in any discussion of the analysis.

Two final notes of caution. First, if there is a substantial trend in the data, then the method chosen for detrending the data may have an effect on the results of a test for autocorrelation (see Exercise 4.3). In this context, referring to Equation 3.1, $Y(x,y) = T(x,y) + \eta(x,y) + \varepsilon(x,y)$, with detrended data, the test for autocorrelation in Equation 4.4 is not made on the values of the Y_i but rather on $Y_i - \hat{T}(x_i, y_i)$, where $\hat{T}(x,y)$ is an estimate of the trend $T(x,y)$. This may disrupt the results. Second, the tests described in this section do not apply to data that have been transformed via linear regression because the residuals of the regression are inherently correlated (see Section 13.2). This also means that if the trend estimate $\hat{T}(x,y)$ has been obtained by regression, then its residuals are autocorrelated. In this context, the issues associated with testing detrended data and transformed data are similar. Valid tests for such data do exist and will be described in Section 13.2. In general, in dealing with data that has been subjected to some form of transformation, it is wise to carry out tests using both the functions described in this section and those in Section 13.2 and compare the results.

4.5 Measures of Autocorrelation Structure

4.5.1 The Moran Correlogram

The autocorrelation statistics described in Sections 4.3 and 4.4 describe the overall level of autocorrelation of a spatial data set, but they say nothing about the structure of this autocorrelation. The *spatial correlogram* (Cliff and Ord, 1981, Chapter 5) describes the manner in which spatial autocorrelation changes with increasing lag distance between locations. For the polygon and point data with which we are concerned, distance is measured via the spatial weights matrix *W*. Locations *i* and *j* that are directly contiguous, as reflected in a nonzero term w_{ij} in the spatial weights matrix *W*, are said to have *first-order contiguity*. If locations *i* and *j* are first-order contiguous and a location *k* is not first-order contiguous with location *i* but is so with location *j*, then *i* and *k* have *second-order contiguity*. Higher orders of contiguity are defined similarly. To compute a spatial correlogram one must first fix the manner in which one models the spatial relationship between data locations with higher levels of contiguity.

One simple way to define higher-order contiguity is to base it on the elements of the spatial weights matrix (Bivand et al., 2013b, p. 269). We can define (somewhat informally) a path between two locations to be a sequence of nonzero elements of the spatial weights matrix that connects them. For example, suppose again that locations *i* and *j* are first-order contiguous and that location *k* is not first-order contiguous with location *i* but is with location *j*, so that *i* and *k* have *second-order contiguity*. If the locations are polygons and the spatial weights matrix is constructed using the rook's case contiguity rules, then the path is a sequence of polygons, each of which share a common nonzero boundary. One can then say that two locations are at a *lag distance h* if the shortest rook's path between these cells passes through $h-1$ intervening cells (Haining, 2003, p. 79). For example, consider the map of nine cells in Figure 3.7, and assume rook's case contiguity. Polygons 1 and 2 are at a lag distance of one, since they share a common boundary. Polygons 1 and 5 are at a lag distance of two, since they do not share a nonzero boundary, and they do share a boundary with polygon 2.

Polygons 1 and 6 are separated by a lag distance of three, since these are not at lag distance two and they share a boundary with at least one polygon at lag distance 2.

For each lag distance *h*, determined either using a distance or a contiguity rule, one can define a spatial weights matrix *W*(*h*) using binary weights, row-normalized weights, or any other type of weights, and then compute a Moran's *I* statistic *I*(*h*) via *W*(*h*). This function *I*(*h*) is called the *Moran correlogram*. It gives a measure of the change in correlation structure as distance between lattice cells is increased.

We will compare the Moran correlograms for three different methods of construction of *W*(*h*) using the detrended sand content data from Section 3.6. First, we create a distance based *W*(*h*) using the function `dnearneigh()` with lag spacings of 61 m as described above. We generate the Moran correlogram using the `spdep` function `sp.correlogram(neighbours, var, order, method, style,...)`. The argument `method = "I"` specifies a Moran correlogram, and `style = "W"` specifies a row normalized spatial weights matrix.

```
> nlist <- dnearneigh(data.Set4.1, d1 = 0, d2 = 61)
> I.d <- sp.correlogram(nlist, Silt, order = 5,
+   method = "I", style = "W")
```

Next, we construct a correlogram based on Thiessen polygons and a rook's case spatial weights rule. The Thiessen polygons in this section are constructed with the functions in the R package `spatstat` using the same coding as is described in Section 3.5.3. As such, they are rectangular rather than trapezoidal in shape, but they are topologically equivalent to the trapezoidal figure shown in Figure 4.2.

The attribute data are then added using the following statement.

```
> thsn.spdf <- SpatialPolygonsDataFrame(thsn.sp,
+   data.Set4.1@data, FALSE)
```

As always, we must check to see that each attribute data record is correctly matched with the appropriate polygon. We do this by first comparing the attribute data *ID* values.

```
> thsn.sf <- st_as_sf(thsn.spdf)
> all.equal(1:length(thsn.sf$ID), thsn.sf$ID)
[1] TRUE
```

See Section 3.6.2 for a visual check for misalignment.

Having verified that the attribute data is correctly aligned with the spatial data, we can construct the correlogram.

```
> nlist <- poly2nb(thsn.sp, queen = FALSE)
> I.r <- sp.correlogram(nlist, Silt, order = 5,
+   method = "I", style = "W", randomisation = TRUE)
```

Using similar code, we also construct a correlogram with queen's case contiguity. Figure 4.3 shows three Moran correlograms. They are similar, and all display a weak negative correlation at more than one spatial lag. This is a common occurrence and is due to the patchiness of the autocorrelated data (Sokal and Oden, 1978; Fortin and Dale, 2005, p. 127). The value of the spatial lag at which *I* is no longer significantly positive can be used as an indication of the range of autocorrelation of the data. Cliff and Ord (1981, Ch. 5) provide further discussion of the interpretation of the correlogram.

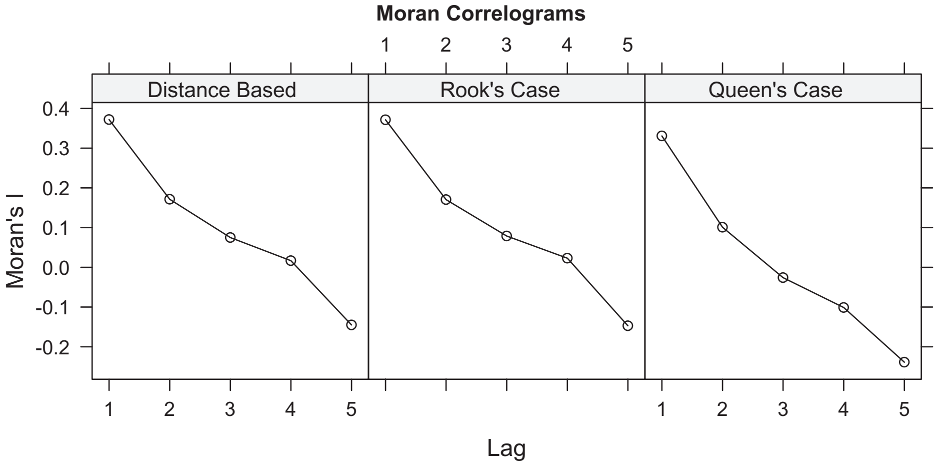

FIGURE 4.3
Moran correlograms for the detrended sand data of Field 1 of Data Set 4. Left to right are a correlogram of point data; a correlogram of Thiessen polygon data under rook's case contiguity, and a correlogram of Thiessen polygon data under queen's case contiguity.

4.5.2 The Moran Scatterplot

Another very useful graphical tool based on the Moran's I statistic is the *Moran scatterplot* (Anselin, 1996). This is a bivariate plot that takes advantage of the fact that the Moran's I statistic can be represented as the slope of a regression line. By plotting this regression line together with the coordinate pairs that it regresses, one can identify potential outliers and gain insight into the local structure of the data. To derive the regression equation, we observe that if W is row normalized then $S_0 = n$ so Equation 4.4 becomes

$$I = \frac{\sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2}. \tag{4.8}$$

We can rewrite Equation 4.8 as

$$I = \frac{\sum_i (Y_i - \bar{Y}) \sum_j w_{ij}(Y_j - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2} \tag{4.9}$$

or

$$I = \frac{\sum_i (Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum_i (Y_i - \bar{Y})^2} \tag{4.10}$$

where:

$$Z_i = \sum_j w_{ij} Y_j \tag{4.11}$$

Equation 4.11 may be written in matrix form as $Z = WY$, and comparison of Equation 4.11 with Equation 3.19 indicates that Z represents a spatial lag in Y.

Equation 4.10 represents the slope of the regression line of Z on Y (Equation A.25 in Appendix A) with a zero intercept term. This indicates that the Moran's I may be considered as the slope of the regression of the spatial lag WY on Y. It follows that a scatterplot of WY against Y should be a very useful way to visualize the spatial distribution of Y. Anselin (1996) calls such a plot a *Moran scatterplot*. In particular, the Moran scatterplot reveals the extent to which the global I statistic effectively summarizes the spatial structure.

We can construct a Moran scatterplot using the `spdep` function `moran.plot()`. As with any plotting function, the act of creating the plot is a side effect of the evaluation of the function. Many plotting functions do not return a meaningful value, but `moran.plot()` does. It returns a matrix whose columns are the results of diagnostic tests by the function `influence.measures()` for each data record (X_i, Y_i). These diagnostics are discussed in Appendix A.2.3. We will create an object to hold this matrix, and, as a side effect, print the Moran scatterplot (Figure 4.4). We do not run into the same issues using detrended data that we did in Section 4.4, so for purposes of comparison with Figure 3.4, we will use the detrended sand content.

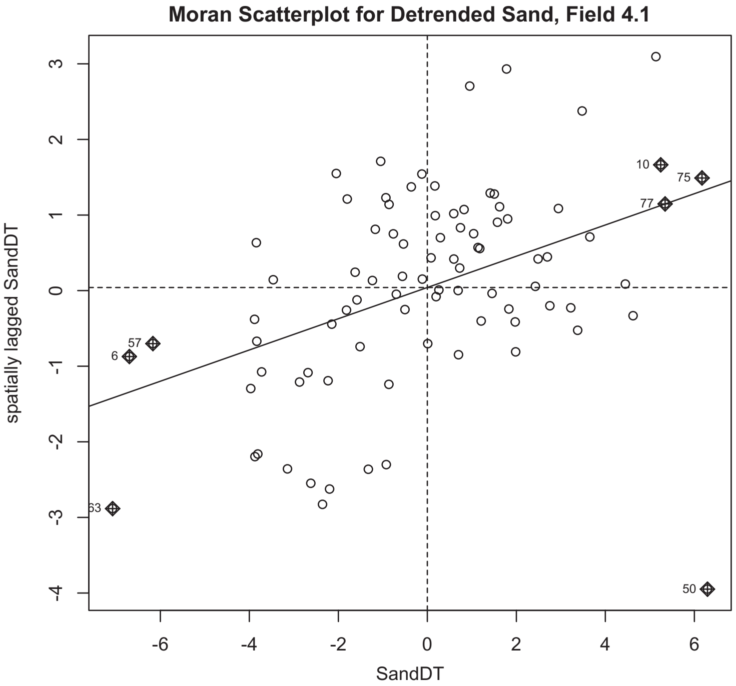

FIGURE 4.4
Moran scatterplot for the detrended sand content data of Figure 3.4b.

```
> trend.lm <- lm(Sand ~ x + y + I(x^2) +
+   I(y^2) + I(x*y), data = data.Set4.1)
> data.Set4.1$SandDT <- data.Set4.1$Sand - predict(trend.lm)
> SandDT <- data.Set4.1@data$SandDT
> SandDT.mp <- moran.plot(SandDT, W) # Fig. 4.4
> title(main = "Moran Scatterplot for Detrended Sand, Field 4.1")
```

The matrix of influence measure tests is in the slot `is.inf`. We can first determine which data records test positive for at least one influence measure by summing across the rows of the matrix. Any row with a positive sum indicates at least one positive test.

```
> print (inf.rows <- which(rowSums(SandDT.mp$is.inf) > 0))
 6 10 50 57 63 75 77
```

Seven data records test positive for at least one influence measure. These are also identified in Figure 4.4. Let's see how many measures are violated by each record.

```
> SandDT.mp$is.inf[inf.rows,]
   dfb.1_ dfb.x dffit cov.r cook.d   hat
6   FALSE FALSE FALSE  TRUE  FALSE  TRUE
10  FALSE FALSE FALSE  TRUE  FALSE FALSE
50  FALSE  TRUE  TRUE  TRUE   TRUE  TRUE
57  FALSE FALSE FALSE  TRUE  FALSE  TRUE
63  FALSE FALSE FALSE  TRUE  FALSE  TRUE
75  FALSE FALSE FALSE  TRUE  FALSE  TRUE
77  FALSE FALSE FALSE  TRUE  FALSE FALSE
```

Examining Figure 4.4 indicates that of these potentially influential points, data record 50 is the most distinctive. This data record has an exceptionally high value relative to its neighbors (see Figure 4.5a below). The location of point 50 in the extreme lower right of the plot is consistent with the following interpretation of the Moran scatterplot (Anslein, 1996, p. 117). Points in the upper right and lower left quadrants correspond to positive autocorrelation, with the upper right quadrant containing high-valued points with high-valued neighbors and the lower left quadrant containing low-valued points with low-valued neighbors. The upper left quadrant (low-valued points with high-valued neighbors) and the lower right quadrant (high-valued points with low-valued neighbors) correspond to pockets of negative spatial autocorrelation. Data record 50 is an extreme case of a high-value with low-valued neighbors.

4.5.3 Local Measures of Autocorrelation

The Moran scatterplot discussion in Section 4.5.2 indicates that the Moran's *I* itself summarizes the contribution of many spatial lag pairs in the same way that a regression line summarizes many pairs of response and predictor variables. The local autocorrelation structure may, however, be quite different in some areas from that described by a global statistic such as the Moran's *I*. For example, the region around data record 50 in the detrended sand data of Field 4.1 is characterized by negative spatial autocorrelation, although the overall detrended sand content displays a high level of positive spatial autocorrelation. Spatial structure exists at many scales and one might expect that subregions of a greater whole could exhibit a local autocorrelation structure very different from that characterized by the single statistic that describes the entire region.

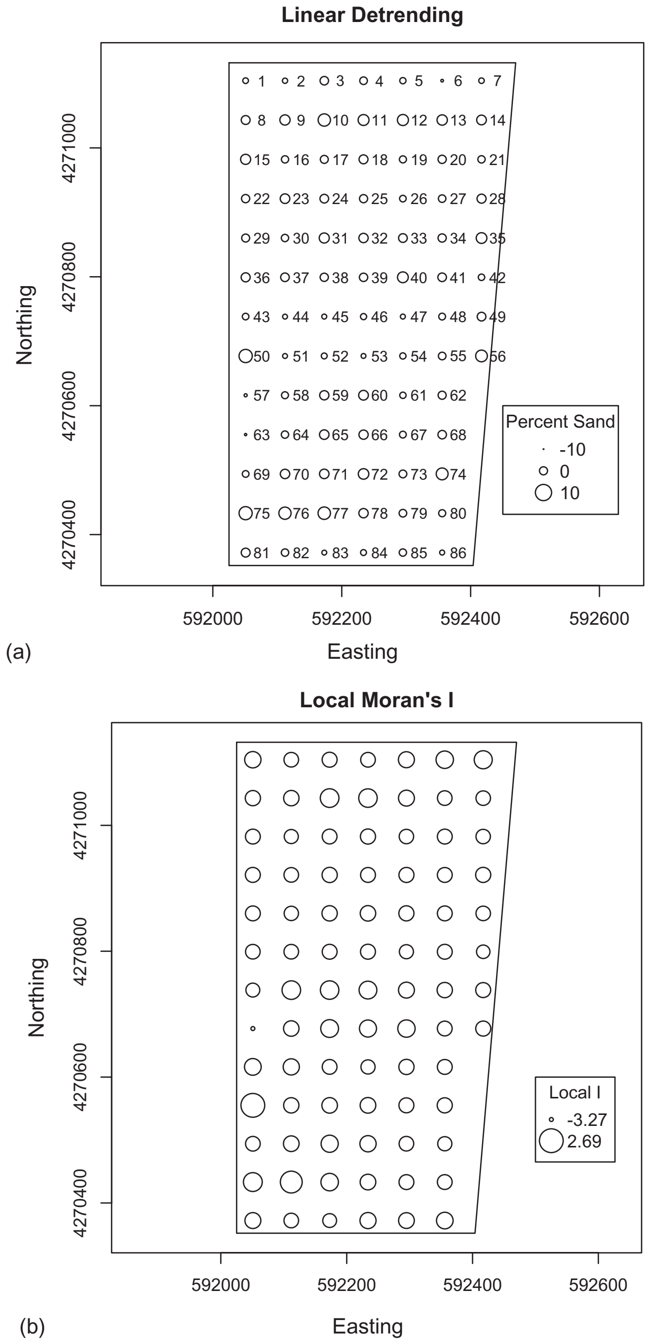

FIGURE 4.5
(a) Bubble plot of detrended percent sand content of Field 1 of Data Set 4 (same data as Figure 3.5b), shown with sample location ID numbers; (b) Local Moran's I for the data of Figure 4.5a.

The first widely used explicitly local measures of spatial autocorrelation statistics are attributed to Getis and Ord (1992, 1996), although in a prior paper Chou et al. (1990) actually proposed a special case of these statistics. Getis and Ord (1992) discussed a pair of statistics they called $G_i(d)$ and $G_i*(d)$. Here i is the index of the polygon element and d is a measure of distance from element i. In the original paper, d is assumed to be a Euclidean distance, but it could equally be a path distance similar to that used in computing the correlogram. The statistic $G_i(d)$ is defined as

$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}(d) Y_j}{\sum_{j=1}^{n} Y_j}, \tag{4.12}$$

where the sums explicitly exclude the $j = i$ term. In the original paper, Getis and Ord (1992) only consider a binary weights matrix, but Ord and Getis (1995) discuss other forms. The special case discussed by Chou et al. (1990), which they called a spatial weights index, consists of the G_i statistic with the first-order binary spatial weights matrix. The statistic $G_i*(d)$ is defined identically to $G_i(d)$ except that the summation explicitly includes the $j = i$ term. Both of these statistics apply only to ratio scale quantities (i.e., quantities possessing a natural zero value, see Section 4.2.1). In the later paper, however, Ord and Getis (1995) extend the statistics to interval and, potentially, ordinal scale quantities.

Anselin (1995) defines another class of statistics that he calls *local indicators of spatial association*, or LISA statistics. Like the Getis-Ord statistics, LISA statistics are defined for each cell of an areal data set and serve as an indication of spatial clustering of similar values. In addition, however, LISA statistics have the property that their sum, when taken over all locations, equals the value of the corresponding global statistic. The most important LISA statistic is the local Moran's I, defined as

$$I_i = (Y_i - \bar{Y}) \sum_{j=1}^{n} w_{ij}(Y_j - \bar{Y}). \tag{4.13}$$

Anselin (1995, p. 99) shows that $\Sigma_i I_i = I$, where I is given in Equation 4.4. LISA statistics, therefore, provide an indicator of how areas in the region contribute to the value of the global statistic.

The `spdep` package contains functions to compute both the G statistics and the local I. We will focus on the local I. We can again use the detrended sand data since we are considering local relationships. The function `localmoran()` computes the local I at each location.

```
> SandIi <- localmoran(SandDT, W)[,1]
```

Anselin (1995, p. 102) points out that the local I statistic is positive in regions of clustering of similar values and negative in regions of clustering of dissimilar values. We can plot the local I, but we should be careful to properly scale the plot. Section 2.6.2 contains an extensive discussion of the use of the function `plot()` for spatial data, but we give here a brief discussion of scaling. The argument `cex` of the function `plot()` is used to control the size of point symbols. The "nominal" value is 1, and we will use a range of 0 to 3 in our map. We can scale the plotting of the local I values as follows. Let cex_{min} and cex_{max} be the minimum and maximum values we wish to plot, and let I_{min} and I_{max} be the minimum and maximum local I values. Then the appropriate scaling is

$$cex = cex_{\min} + \frac{I_i \times (cex_{\max} - cex_{\min})}{I_{\max} - I_{\min}}. \tag{4.14}$$

Figure 4.5a shows the linearly detrended sand content data of Field 4.1. This is the same map as Figure 3.4b, but this time the data location ID numbers are shown. Figure 4.5b shows a bubble plot of the corresponding local Moran's I values. As was the case with the Moran scatterplot analysis of the previous section, the data at location 50 stands out as being very different from its neighbors. Removal of location 50 changes the value of the global Moran's I from 0.207 to 0.317.

Anselin (1995) describes two primary uses of local statistics, to detect "hot spots," or local areas of clustered extreme values, and to identify individual outliers. The trend analysis described in Section 3.2.1 detects nonstationarity due to trends, but may not be effective in detecting other forms of nonstationarity. It is possible in principle to test the null hypothesis of zero local autocorrelation for each of these statistics (Ord and Getis, 1995, Anselin, 1995), but probably their greatest use is exploratory rather than confirmatory.

4.5.4 Geographically Weighted Regression

The ordinary least squares (OLS) regression model with a single explanatory variable is (Appendix A.2, Equation A.22)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1,\ldots,n. \tag{4.15}$$

Geographically weighted regression (GWR) (Fotheringham et al., 1997, 2002) modifies Equation 4.15 to let the regression coefficients be functions of position:

$$Y_i = \beta_0(x_i, y_i) + \beta_1(x_i, y_i) + \varepsilon_i,\ i = 1,\ldots,n. \tag{4.16}$$

The method can easily be extended to more than one explanatory variable, in which case it becomes more convenient to write the regression equations in matrix notation (Appendix A). In particular, the equation for the least squares estimate of the OLS regression coefficients $b = [b_0\ \ b_1\ \ \ldots\ \ b_{p-1}]'$, where the prime denotes transpose, is Equation A.35

$$b = (X'X)^{-1}X'Y, \tag{4.17}$$

where Y is the vector whose components are the Y_i and X is a matrix, called the design matrix, whose elements are values of the explanatory variables X_i (see Appendix A.2).

In this notation, the geographically weighted regression model incorporates local variation in the estimated regression coefficients $b_k(x_i, y_i)$ by incorporating a geographic weighting matrix $G(x_i y_i)$. This is a diagonal matrix (see Appendix A.1) whose diagonal elements incorporate the spatial distance d_{ij} between the location (x_i, y_i) of the explanatory variable X_i and the location (x_j, y_j) of the response variable Y_j.

The modified form of Equation 4.17 used in GWR is written

$$b(x_i, y_i) = (X'G(x_i, y_i)X)^{-1}X'G(x_i, y_i)Y. \tag{4.18}$$

This is not a single equation but rather a system of matrix equations that is evaluated at each location.

As d_{ij} increases, the explanatory variable is expected to have less influence over the response variable, so the elements of G should reflect this. There is more than one form available for G, but a very common one is the *Gaussian* form, in which case the elements are $g_{jj}(x_i, y_i) = \exp(-[d_{ij} / h]^2)$. The parameter h, which determines the width of the normal shaped curve, is called the *bandwidth* and must be estimated from the data. One way to estimate the bandwidth is by using cross validation, in which one iteratively solves the regression equations and chooses h to minimize the quantity

$$Q = \sum_{i=1}^{n} \left[Y_i - \hat{Y}_{-i}(h) \right]^2, \tag{4.19}$$

where $\hat{Y}_{-i}(h)$ is the predicted value of Y_i resulting from a regression where the date record (X_i, Y_i) is not used.

GWR is particularly useful for the exploration of stationarity in the data. If there is a large variation in the estimated regression coefficients $b_k(x_i, y_i)$, then this can be taken as indicating that the data are not stationary. The method is implemented in R in the package `spgwr` (Bivand and Yu, 2017). We first illustrate it using an artificial data set on a 20 by 20 grid.

```
> data.df <- expand.grid(x = seq(1,20),
+   y = seq(20,1, by = -1))
```

We create two variables X and Y and convert the grid into a `SpatialPointsDataFrame.`

```
> set.seed(123)
> data.df$X <- rnorm(400)
> data.df$Y <- data.df$X + 0.1*rnorm(400)
> coordinates(data.df) <- c("x", "y")
```

Taking advantage of the fact that we can use `data.df` as if it were a data frame, we can compute the OLS regression coefficients.

```
> coef(lm(Y ~ X, data = data.df))
(Intercept)            X
0.0004473601 1.0028106263
```

We know that data are stationary because there is no dependence on position. Let's see what we get from a geographically weighted regression analysis. The first step is to compute the bandwidth via cross-validation using Equation 4.19. This is done using the `spgwr` function `gwr.sel()`.

```
> library(spgwr)
> Y.bw <- gwr.sel(Y ~ X, data = data.df)
```

Next, we plug this bandwidth plus the formula and data source into the function `gwr()`.

```
> Y.gwr <- gwr(Y ~ X, data= data.df, bandwidth = Y.bw)
```

One of the data fields of the object Y.grw is denoted SDF. This is a SpatialPointsDataFrame with, among other things, the estimated regression coefficients. There are 400 of them, one for each location. Let's look at the range.

```
> range(Y.gwr$SDF$X)
[1] 0.9932662 1.0107629
```

The range is very small.

Fotheringham et al. (2000, p. 114) recommend the following as an indication of nonstationarity. If the data are stationary, then the regression coefficients should not be affected by random rearrangements of the data. The standard deviation of the regression coefficients is one measure of this effect. Therefore, one can carry out a permutation test of (say) 99 random rearrangements of the X_i and compute the standard deviation of each rearrangement. Using the original data as the hundredth rearrangement, it should not be an extreme value. To carry out this test, we embed the code sequence above into a function and replicate it 99 times.

```
> demo.perm <- function(){
+    data.test <- data.df
+    data.test@data$Y <- sample(data.df@data$Y,
+       replace = FALSE)
+    Y.bw <- gwr.sel(Y ~ X, data = data.df)
+    Y.gwr <- gwr(Y ~ X, data = data.test,
+       bandwidth = Y.bw)
+    return(sd(Y.gwr$SDF$X))
+ }
> set.seed(123)
> U <- replicate(99, demo.perm())
```

You will get a lot of output! You can ignore it, because the important thing is the object U that comes at the end (after a long time). Here is the result of the test.

```
> print(Y.sd <- sd(Y.gwr$SDF$X), digits = 3)
[1] 0.00384
> length(which(U >= Y.sd)) / 100
  [1] 0.98
```

Ninety-eight percent of the permutations have a standard deviation at least as large as the data. This is a strong indication that the data are stationary.

Now let's try this same procedure with the sand data (not detrended) from Field 4.1. We expect gwr() to produce an indication of nonstationarity. This uses the interpolated yield data file *set4.1yld96ptsidw.csv* that contain 1996 yield data interpolated to the locations at which the data in the file *data.Set4.1* were sampled. It is created in Exercise 6.5, but is also available in the *created* folder on the book's website. We will use this to regress yield against sand concentration, which, due to its strong trend, we expect to be nonstationary. The code for the regression is analogous to that given above and is not shown.

```
> Sand.perm <- function(){
+    data.test <- data.Set4.1
+    data.test@data$Sand <- sample(data.Set4.1@data$Sand,
+       replace = FALSE)
+    Sand.bw <- gwr.sel(Yield ~ Sand, data = data.test)
+     Sand.gwr <- gwr(Yield ~ Sand, data = data.test,
+        bandwidth = Sand.bw)
+     return(sd(Sand.gwr$SDF$Sand))
+ }
> set.seed(123)
> U <- replicate(99, Sand.perm())
> print(Sand.sd <- sd(Sand.gwr$SDF$Sand), digits = 3)
[1] 143
> length(which(U >= Sand.sd)) / 100
[1] 0
```

We do indeed see a strong indication of nonstationarity. None of the permutations have a standard deviation as large as the observed data.

It is important to remember (Section 3.2.2) that stationarity is not a property that can be tested statistically, and that one person's nonstationarity is another person's correlated error structure. Dealing with apparent nonstationarity and trends is one of the most subjective and tricky aspects of spatial data analysis. The best practice is to try different alternative models for the trend $T(x, y)$ of Equation 3.1 and try to gain an understanding of how these alternatives influence the outcome of the analysis. Whatever you do, don't just rely on GWR as an easy way out of the problem.

4.6 Measuring Autocorrelation of Spatially Continuous Data

4.6.1 The Variogram

Although ecological landscape data is almost always more accurately modeled as varying continuously in space, all of the statistics discussed so far have assumed an areal model. This is mostly a matter of convenience because many of the statistical methods described in the following chapters are based on this model. In attempting to characterize the data as completely as possible, however, it makes sense to try to represent them using the model that most accurately characterizes their natural structure. The *variogram* is the most common means of characterizing the spatial structure of data that vary continuously across the landscape. In this section, we describe the construction of the variogram. Variogram analysis is associated with kriging, a widely used interpolation method for spatial data, and in this context it has been discussed in some excellent texts at both the intermediate (e.g., Isaacs and Srivastava, 1989; Panatier, 1996) and advanced (e.g., Cressie, 1991) level. Chapter 6 includes a brief introductory discussion of kriging interpolation.

The discussion of the variogram will use the soil sample data from Field 4.1. We begin by taking up again the *spatial lag*, which represents the difference between the locations of two measurements. In geostatistics, the spatial lag is traditionally denoted h. The spatial lag h is a vector quantity since it has both a magnitude and a direction. Figure 4.6 shows a plot of the sample points of Field 4.1 together with three different lag vectors. The distance between sample points is 61 m in both the north–south and east–west directions. Therefore, the vectors shown in the figure are $h_1 = (61,0)$, $h_2 = (61,61)$, $h_3 = (0,183)$. Throughout this discussion, we assume that the data are isotropic (i.e., independent of direction; see Section 3.2.2). The theory can be extended to more general cases (Cressie, 1991, p. 61; Isaacs and Srivastava, 1989, p. 96; Panatier, 1996, p. 53). In the isotropic case, the variogram properties

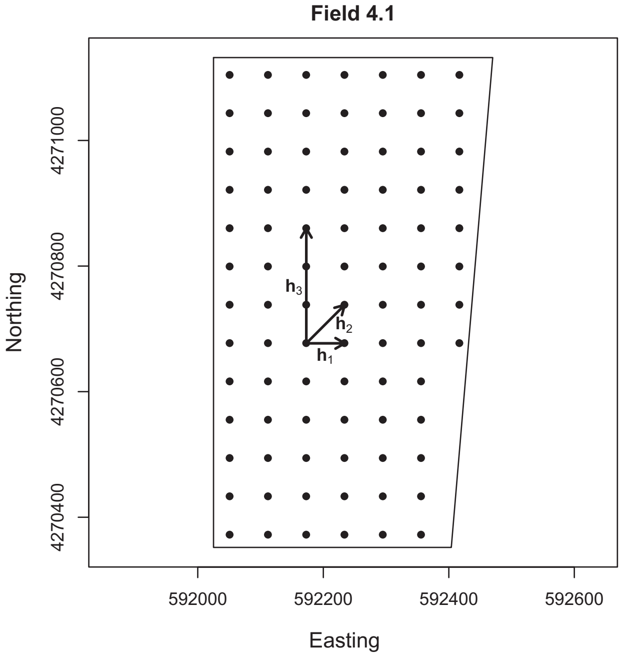

FIGURE 4.6
Sample points of Field 4.1, showing three different spatial lags.

depend only on the magnitude of *h*, and not on its direction. Because of this, we use the symbol *h* to denote the magnitude of the vector *h* as well as the vector itself.

In the isotropic case, we define the *variogram* to be the following function of the spatial lag (Cressie, 1991 p. 58):

$$\gamma(h) = \frac{1}{2}\mathrm{var}\{Y(x+h) - Y(x)\}, \tag{4.20}$$

where *Y* is the measured quantity and *x* is a position vector with coordinates (x, y). Because of the factor ½, the correct term for $\gamma(h)$ is the *semivariogram*. However, because the R functions for computing this quantity generally use the term "variogram," so will we. Whether or not one includes the ½ will turn out not to matter too much. The variogram is generally estimated by the *experimental variogram* $\hat{\gamma}(h)$, which is defined as follows. Let *m*(*h*) be the number of lag vectors that have a magnitude *h*. Then (Isaacs and Srivastava, 1989, p. 60; Panatier, 1996, p. 38),

$$\hat{\gamma}(h) = \frac{1}{2m(h)}\sum_{i=1}^{m(h)}[Y(x_i + h) - Y(x_i)]^2, \tag{4.21}$$

where the sum is taken over all sample points separated by a lag vector of magnitude *h*. Cressie (1991, citing Matheron, 1963) points out that $\hat{\gamma}(h)$ of Equation 4.21 is a method of moments estimator (Larsen and Marx, 1986, p. 267) for $\gamma(h)$ of Equation 4.20. The reason for the 2 in the denominator is that the sum counts each vector pair twice, once in each direction of the vector *h*.

In the case of a regularly spaced grid of data locations, the grid spacing is often referred to as the *fundamental lag*. For example, in Figure 4.6, if the fundamental lag is 61 m, then the sample points shown as separated by lag vector h_1 are in the group separated by one lag, and so are the same two points separated by the vector in the opposite direction. Although some of the lag vectors, such as h_1 and h_3 in Figure 4.6, have integer magnitude, others, such as h_2, do not. As the magnitude of h increases one arrives at a situation where there are relatively few lag vectors that have the same magnitude (i.e., $m(h)$ becomes small relative to the number of lags), but there are groups of lag vectors that have almost the same magnitude. For this reason, the calculation in Equation 4.21 is generally modified by dividing the lag vectors into *lag groups*, each of which contains all of the lag vectors whose magnitude lies within a specified interval, and carrying out the summation over these lag groups (Panatier, 1996, p. 30). When we need to represent these lag groups explicitly, we will denote the k^{th} group by $H(k)$, defined by those points separated by a distance $kh_0 \le h < (k+1)h_0$ for some fixed h_0. For example, in the case of the data of Field 4.1 we might set $h_0 = 61\,\text{m}$.

There are several R packages containing functions that may be used to compute the variogram. The computations described here use the package `gstat` (Pebesma, 2004). We begin by computing and plotting the variogram of the silt content of Field 4.1. The code to compute a variogram is as follows.

```
> library(gstat)
> Silt.vgm <- variogram(Silt ~ 1, data.Set4.1, cutoff = 600)
```

The argument `Silt ~ 1` in the function `variogram()` is a *formula* and indicates that the variogram of `Silt` is to be computed without reference to any other variables. The second argument provides the data set. The third argument, `cutoff`, has a couple of uses, but the main one is to exclude any spatial lags larger than the value provided. If no value of `cutoff` is provided, `variogram()` will use a default (see `?variogram`). Figure 4.7a plots the experimental variogram of the soil silt content of Field 4.1. This is almost a "textbook" plot of spatial data. The solid curve is the variogram model, computed using the function `fit.variogram()` using the following statement.

```
> Silt.fit <- fit.variogram(Silt.vgm, model = vgm(1, "Sph", 700, 1))
```

The first argument of this function provides the experimental variogram. The second argument specifies the model. This model is specified by the function `vgm()`. Let's see what each of the arguments of `vgm()` represents, starting with the second, the class of model (we will get to the first argument below). The argument `"Sph"` signifies a spherical model. The spherical variogram model has the form (Isaaks and Srivastava, 1989, p. 374; Pebesma, 2001, p. 38)

$$\gamma_{sph}(h) = \begin{cases} 1.5\dfrac{h}{a} - 0.5\left(\dfrac{h}{a}\right)^3 & : h \le a. \\ 1 \quad \textit{otherwise} & \end{cases} \tag{4.22}$$

The spherical model is one of a class of variogram models that are *positive definite*. These models are commonly used in kriging interpolation because they ensure that the kriging variance will be positive (Isaacs and Srivastava, 1989, p. 372). The fitting parameter a is called the *range*, and represents the value of h at which the $\gamma_{sph}(h)$ reaches a constant value

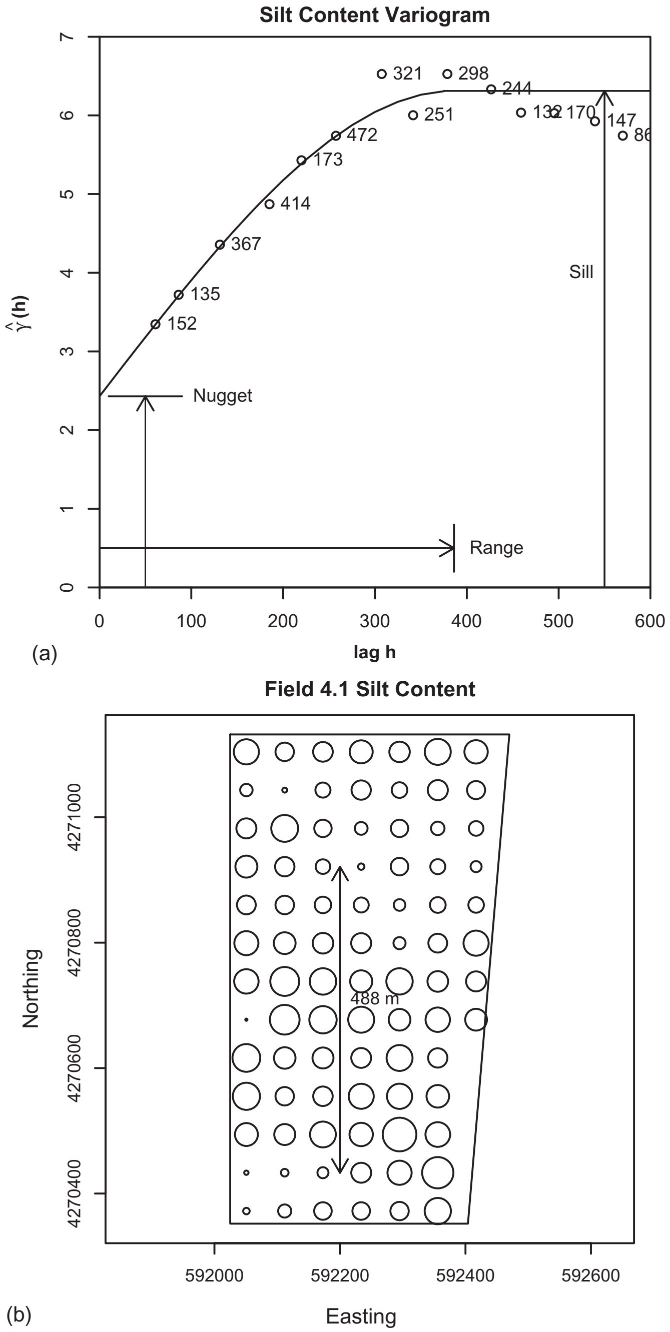

FIGURE 4.7
(a) Experimental variograms of percent silt content data of Field 4.1; (b) thematic map of percent silt content of Field 4.1, indicating that low values at the north and south ends are separated by a distance of about 450–500 m.

(i.e., "flattens out"). The range represents the distance at which attribute values cease to be spatially autocorrelated. Figure 4.7a indicates that the range for silt content is about 350 meters. Since $\gamma_{sph}(0) = 0$ and $\gamma_{sph}(h) = 1$ for $h \ge a$, the function actually fit by nonlinear least squares involves two other fitting parameters and has the form

$$\gamma(h) = b + (c - b)\gamma_{sph}(h),\ h > 0. \tag{4.23}$$

From the defining Equation 4.20, $\gamma(0)$ must equal zero. The parameter b is called the *nugget* and represents $\lim_{h \to 0} \gamma(h)$. The nugget represents the combination of sampling error and short-range variability that causes two samples apparently taken from the same location to have different values. If the nugget b is nonzero, then the variogram is discontinuous at zero. The parameter c is called the *sill*. This represents the long-range variability of the samples. These three quantities, the range, nugget, and sill, are defined not only for the spherical model, but also for many other variogram models. The arguments of the function `vgm()` are respectively the initial guess for the sill, the class of model, the initial guess for the range, and the initial guess for the nugget.

The numbers to the right of each experimental variogram point in Figure 4.7a are the number of points in the lag group used to compute the corresponding value of $\hat{\gamma}(h)$. The greater the number of pairs of points in a lag group, the more confidence can be placed in the computed experimental variogram value. Webster and Oliver (1990, p. 222) suggest that one hundred sample data values should be regarded as the minimum to compute an accurate variogram estimate. By this standard, the data set used in this section, with 86 values, falls just short. Webster and Oliver (1992) point out that in some cases as many as 200 data values may be necessary.

One way in which the experimental variogram of soil silt content in Figure 4.7a does not display "textbook" behavior is that it declines for very large values of the lag h. There are two reasons why this might occur. One is that the number of points in the lag group for very large lags becomes smaller, which can lead to increased variability in the variogram at these lags. Ordinarily, however, this variability includes values above as well as below the sill, while in Figure 4.7a the values of $\hat{\gamma}(h)$ are all below the sill for large h. Figure 4.7b shows that in Field 4.1 the low values of silt content are at the north ends of the field, separated by a distance of about 450–500 m. Therefore, values at this distance are actually more highly correlated than values from sample sites located closer to each other. Experimental variograms that decline with distance or oscillate in magnitude can be indicative of a patchy spatial structure.

Figure 4.8a shows experimental variogram for the "raw" (i.e., not detrended) sand content data. This variogram does not level off. The failure to reach a sill is a common manifestation of a trend. In the presence of a trend, the squared difference $(Y(x_i + h) - Y(x_i))^2$ in Equation 4.21 never reaches a constant value as h increases but rather continues to increase. The detrended variogram can be computed using the data field `SandDT` developed in Section 3.5, and it can also be computed directly using the function `variogram()` (see Exercise 4.5b). Figure 4.8b shows the experimental variogram of the detrended sand content. This experimental variogram displays an unfortunate property not uncommon with real data, namely, it does not resemble a "textbook" variogram. There is some indication of a residual trend, and the value of $\hat{\gamma}$ for the smallest lag group is not much smaller than the values for larger lags. A few attempts to fit this variogram with an automatic method were not successful. The computer program Variowin (Panatier, 1996), which is out of print but available on the internet, contains software that allows the user to easily fit a model to an experimental variogram manually.

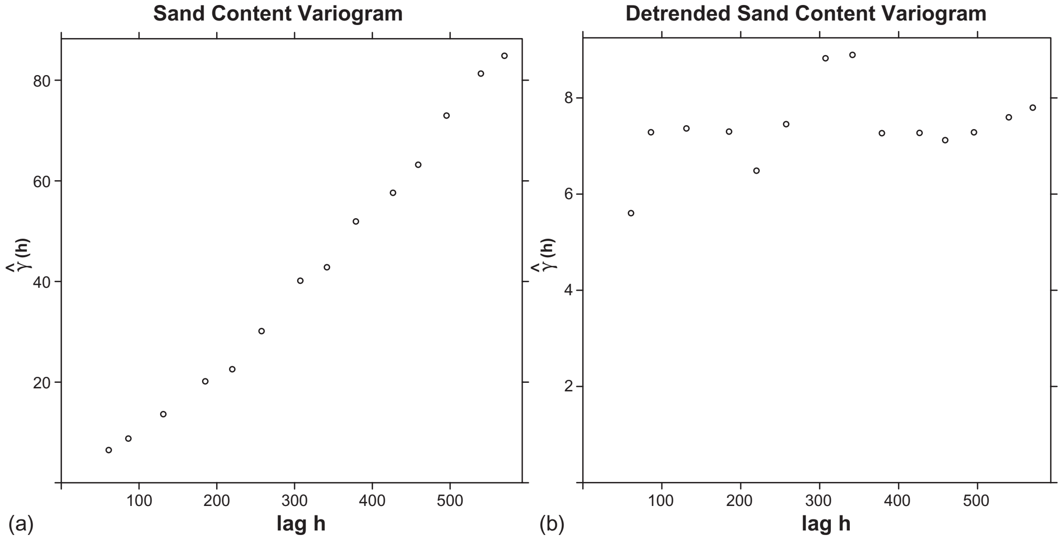

FIGURE 4.8
Experimental variograms of percent sand content data of Field 4.1. (a) Sand content data without detrending, (b) detrended sand content data.

4.6.2 The Covariogram and the Correlogram

Rather than using the variogram, it is sometimes more convenient to work with the *correlogram* $\rho(h)$. This quantity is defined in terms of the *covariogram*. For a second-order stationary random field Y, the covariogram $C(h)$ is defined as (Cressie, 1991, p. 53)

$$C(h) = \text{cov}\{Y(x), Y(x+h)\}, \tag{4.24}$$

and, provided $C(0) > 0$, the correlogram is then defined as (Cressie, 1991, p. 67).

$$\rho(h) = \frac{C(h)}{C(0)}. \tag{4.25}$$

The following discussion will gloss over some technical stationarity assumptions. Banerjee et al. (2004) provide a good discussion of these. We do not discuss them because for ecological data sets they are unlikely to be violated.

From the definition of $\gamma(h)$ (Equation 4.20), it follows that $C(0)$ is equal to the sill of the variogram $\gamma(h)$ and $\lim_{h\to\infty} C(h)$ is equal to the nugget. In other words, for a second-order stationary random field Y, the covariogram $C(h)$ and the variogram $\gamma(h)$ are related by the equation

$$\gamma(h) = C(0) - C(h). \tag{4.26}$$

As with the variogram, the covariogram and the correlogram are generally estimated over lag groups $H(k)$, where k is the index of the lag group. To avoid unnecessary confusion we explicitly note that the symbol h represents a spatial lag, and the symbol k represents the index of a lag group. The experimental covariogram is given by (Isaaks and Srivastava, 1989, p. 59; Ripley, 1981, p. 79)

$$\hat{C}(k) = n_k^{-1} \sum_{i,j \in H(k)} (Y(x_i, y_i) - \bar{Y})(Y(x_j, y_j) - \bar{Y}), \tag{4.27}$$

that is, it is the covariance of data values whose separation is in lag group k. The experimental correlogram $\hat{r}(k)$ can then be estimated by dividing by the sample variance as $\hat{r}(k) = \hat{C}(k) / s^2$.

4.7 Further Reading

Griffith (1987) provides an excellent introduction to autocorrelation statistics. Cliff and Ord (1981) provide a very thorough yet readable discussion of the resampling and randomization assumptions and of the derivation and properties of the join–count, Moran, and Geary statistics. Good entry level sources for permutation tests are Good (2001), Manly (1997), and Sprent (1998). Books at a somewhat higher level include Good (2005), Rizzo (2008), and Edgington (2007). The terminology distinguishing permutation tests is very inconsistent among authors. Some authors use the terms *randomization test* and *permutation test* synonymously, others distinguish between permutation tests in which all possible permutations are examined and randomization tests in which a random sample is drawn without replacement from the permutations, and still other authors use the term *randomization test* for something else entirely (Sprent, 1998, p. 26; Edgington, 2007, p. 20).

Haining (1990, 2003) is an excellent source for discussion of global and local measures of autocorrelation. There are many excellent geostatistics texts. Isaaks and Srivastava (1989) is universally considered to be a great resource. Good sources for Monte Carlo simulation include Ripley (1981), Bivand and Clifford (1989), Manly (1997), and Rubinstein (1981). Efron and Tibshirani (1993) is also very useful, and Rizzo (2008) provides R code. The eigenvectors and eigenvalues of the spatial weights matrix W can provide information about the spatial structure of the data. See Griffiths (1988, Ch. 3) for a discussion. It is considerably easier to understand the application of matrix theory and linear algebra to statistical problems by making use of the concept of the *subject space*, which is discussed by Wickens (1995).

Exercises

4.1 For very small samples, the join–count test can give meaningless results. Do a join–count test using k nearest neighbors with $k = 4$ on the *QUDO* data of Data Set 2 for longitude between –121.6° and –121.55° and latitude between 36.32° and 36.45°. Plot the QUDO data using `pch = QUDO` to see what you are testing.

4.2 Using the *Weed* data from Field 1 of Data Set 4, test the null hypothesis of zero spatial autocorrelation for the full range of levels (i.e., without combining levels). Use both a multicolor join–count statistic and a Moran's I test. How do the results compare?

4.3 (a) Carry out a significance test of spatial autocorrelation using Moran's I for both the "raw" sand content data and the detrended sand content data of Field 1 of Data Set 4. What do you think causes the difference? This is an example of "spurious autocorrelation" discussed in Section 13.1. (b) Compare the results of the test of spatial autocorrelation using Moran's *I* for the detrended data using the linear and median polish detrending.

4.4 Create a Moran scatterplot for mean annual precipitation of Data Set 2 for longitude between –121° and –119° and latitude between 37° and 39°. Identify two points with extreme values. What is the interpretation of these extreme values? Use the function `plot()` to plot the location of these points. Now plot the elevation and discuss whether there might be a physical reason for the extreme local *I* values.

4.5 Using the data of Field 2 of Data Set 4, create Thiessen polygons of different sizes and average the *Yield* data over these polygons. Compute the Moran's *I* for these polygons. What is the effect of polygon size on the Moran's *I*?

4.6 (a) Compute using the function `variogram()` the variogram of detrended sand content using the `SandDT` data field computed via the call to function `lm()` in Section 3.5.

(b) Compute the variogram for sand content directly with the function `variogram()` using the formula `Sand ~ x + y + I(x*y) + I(x^2) + I(y^2)` as the first argument. Compare the result with that of part (a).