CHAPTER 5

# Spatial Statistics

In contrast to land-cover or habitat classes, some kinds of landscape data are recorded as continuous numbers rather than discrete categories. Such data include vegetation density or height; net aboveground primary production; nutrient mineralization rates; percent of biomass killed by a disturbance; and distances from lakes, roads, or other features of interest. Most landscape metrics covered in Chap. 4 are not appropriate for quantifying the spatial pattern of continuous variables, and a different set of methods is required. Spatial statistics, including *geostatistics*, are used to quantify the spatial structure of continuous data, and they are widely applied in landscape ecology. Spatial statistics and geostatistics use point data for some property that is spatially distributed across the landscape; they do not require categorization of the landscape nor do they assume a patchy structure or the presence of boundaries. Observations, conventionally labeled as $z$, are made at specific $x$, $y$ locations and referred to as *regionalized variables* (Palmer and McGlinn 2016). Spatial statistics then quantify *spatial dependence* in the regionalized variable, or the tendency of $z$ measured at one $x$, $y$ location to be correlated with, or depend on, values of $z$ measured at another $x$, $y$ location. If there is spatial dependence in $z$, then information about $z$ at one place allows you to infer information about $z$ at another place. Spatial statistics quantify the magnitude of variance in the data, the proportion of that variance that is spatially dependent (i.e., spatially autocorrelated), and the scales, or distances, over which variables are spatially dependent. These methods are powerful, but the terminology and methods can be daunting for those new to the subject.

Landscape metrics and spatial statistics are distinct but complementary methods of analysis. Gustafson (1998) illustrated this point nicely by considering two


M.G. Turner, R.H. Gardner, *Landscape Ecology in Theory and Practice*,
DOI 10.1007/978-1-4939-2794-4_5

different ways of describing the spatial distribution of aspen (*Populus tremuloides*), a widely distributed tree species in North America. If one converts measurements of aspen tree density to categorical data by delineating patches where aspen is present above some minimum density threshold, the resulting map of aspen distribution can be analyzed using patch-based metrics. Alternatively, if one uses the actual measures (continuous numbers) of aspen density at locations throughout the landscape, then the spatial structure of aspen density would be analyzed using spatial statistics. Both approaches characterize the spatial pattern of aspen on the landscape, but they do so in different ways. Quantifying patterns of disturbance in a landscape offers another example (Fig. 5.1). Areas in a landscape affected by fire can be categorized as discrete patches of burned vs. unburned areas (Fig. 5.1b) and analyzed using landscape metrics, or represented by continuous measures of fire severity (Fig. 5.1a) and analyzed using spatial statistics (see Turner and Simard 2016). Both approaches assess the spatial structure of fire effects on the landscape, but they allow different forms of analysis and inference.

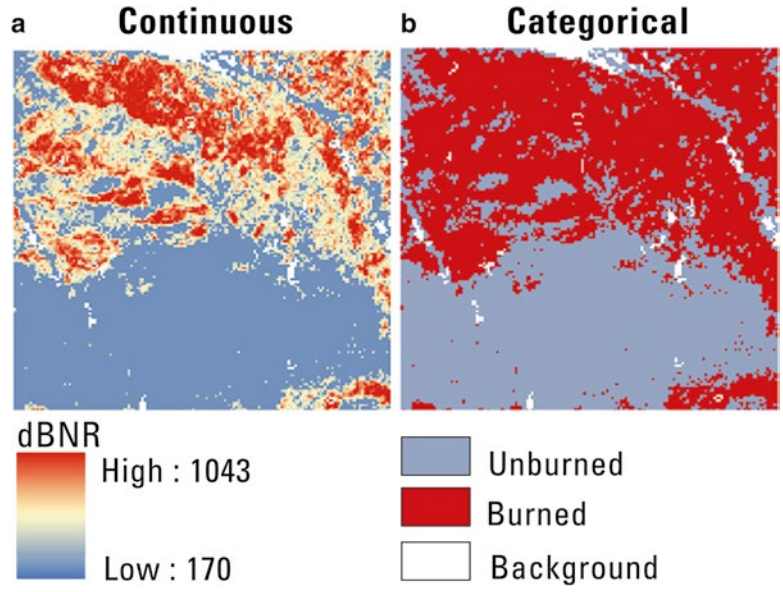


Figure 5.1.

Example of two ways to depict fire patterns in a 5 km × 5 km area of the Greater Yellowstone Ecosystem, Wyoming, USA. The left map shows continuous estimates of fire severity based on the Differenced Normalized Burn Ratio, with warmer colors indicating higher fire severity. The right map classifies these data into burned vs. unburned cells. Although both depict fire patterns, these maps would be analyzed with different methods.

Maps generated by Martin Simard.

The origin of spatial statistics is often traced to the South African mining engineer, D. W. Krige, who developed methods for locating ores within geologic formations and for whom the spatial interpolation method of *kriging* is named. In the 1950s, Krige pioneered statistical methods to predict the location of economically valuable ores (e.g., gold, uranium) using a limited number of boreholes. This work spurred development of methods for spatial evaluation of mineral resources based on geographically referenced point measurements and led to the field of geostatistics. These methods still provide a basis for *spatial interpolation*, which uses measurements at particular points to predict values in locations that lack empirical measurements and is discussed in detail below.

We begin this chapter by explaining several key uses of spatial statistics in landscape ecology, present important caveats for their use, and then provide a brief overview of major approaches. The literature regarding spatial statistics is large, diverse, and technical (Law et al. 2009), making a complete review well beyond the scope of this text (see Legendre and Legendre 1998; Fortin and Dale 2005 for broad coverage of spatial statistics). Our goal is to demystify the elementary jargon of spatial statistics and to provide illustrative examples emphasizing how these methods can be used in landscape ecology. We do this by focusing on two techniques that have been widely applied and that illustrate the general principles (and pitfalls) of the use and application of spatial statistics in landscape ecology. These two methods are (1) point pattern analysis, which analyzes observed "events" (e.g., nest locations, fire starts, etc.) and (2) spatial autocorrelation and variography, which use many spatially distributed measurements of continuous variables. We conclude the section on variography with a short discussion of efficient sampling designs that reduce the cost of acquiring sufficient data for using spatial statistics in landscape studies. But first, let's think about why these methods are useful and note key caveats to consider and pitfalls that may be encountered. The methods themselves are rapidly evolving as spatial data and analysis software become more available; therefore, as we did for landscape metrics, we first emphasize general considerations that should apply to most problems of landscape analysis.

## WHY DO LANDSCAPE ECOLOGISTS USE SPATIAL STATISTICS?

The use of spatial statistics in ecology has become widespread and much more sophisticated since the first edition of this book, and access to computation software has increased. Why have these methods become so important, and why are they used in landscape ecology?

### *Spatial Independence*

As encapsulated in Tobler's first law of geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970). However, independence among data values is one of the most common assumptions of standard parametric statistics. As Tobler's law suggests, when data are collected within a spatial framework, values near one another are often correlated; this violates the assumption of independence and complicates data analysis (Ripley 2005). What is the primary statistical danger? When measurements are spatially autocorrelated (i.e., near things are related to each other, so the data are not independent), there is an increased risk of Type I error—finding a statistically significant difference when none actually exists (i.e., erroneous rejection of the null hypothesis). A simple example illustrates this problem. Imagine a transect along which measurements of plant height are made, and you wish to determine whether the mean plant height is 20 cm. Assume that the variance is known and equal to 5, and you calculated a mean height of 21.5 cm from ten measurements made at 3-m intervals along the transect. A computed $Z$ statistic will reject the null hypothesis with $p = 0.034$, leading you to conclude that the measured mean plant height differs from 20 cm. Now assume that the adjacent sampling points were not independent, but rather they were correlated with $r = 0.40$ (i.e., spatially dependent). A correction to the $Z$ test yields $p = 0.14$, leading to the opposite conclusion that plant heights are not significantly different from 20 cm. Thus, spatial dependence in the data can cause the conclusions of a statistical analysis to change qualitatively. An important application of spatial statistics in landscape ecology is to determine the magnitude and scale of spatial dependence and to adjust either the sampling design or the statistical models to account for this lack of independence.

The plant-height example is a simple one, but spatial autocorrelation in landscape data is the rule rather than the exception. From a practical standpoint, reviewers of proposals or manuscripts routinely expect authors to have tested for spatial dependence and to have adjusted the analysis for spatial effects by using appropriate statistical methods and associated tests. When pilot data are available, spatial statistics can be used to describe the change in dependence with distance between samples and, from this analysis, set a separation distance beyond which measurements are spatially independent. This distance defines the minimal spacing required for future sampling locations that will allow the assumption of independence among measured values to be valid. For example, Pearson et al. (1995) used spatial statistics (semivariograms) to determine the scale over which measurements of winter grazing intensity by elk (*Cervus elaphus*) were spatially dependent, then subsampled data at distances beyond that scale to assure independent samples for analysis of how grazing intensity varied with environmental characteristics. When the degree of spatial autocorrelation in a dataset is not known a priori, spatial dependence can be evaluated in the raw data or from the residuals of a statistical

analysis. When spatial autocorrelation is detected in the residuals, the statistical model must be adjusted to account for this dependency (see Legendre 1993; Fortin and Payette 2002; Ishihama et al. 2010).

Even when it may not have been welcomed, detection of spatial dependence can be a source of new insights about a focal variable. Spatial dependence in the residuals of a statistical model should alert the investigator to the potential importance of a spatial process that was not previously considered. For example, Anderson et al. (2013) examined the spatial distribution of a nonnative invasive plant across a large mountainous landscape. After accounting for the local and landscape-level variables that predicted presence of the invasive plant, spatial dependence was still present in the residuals up to a scale of 3 km. The residual spatial structure suggested a hierarchical process of invasion. The spatial pattern was consistent with infrequent, long-distance dispersal events resulting in new nascent subpopulations that subsequently spread via shorter-distance dispersal (Anderson et al. 2013). Thus, residual spatial autocorrelation, after accounting for other covariates, can be informative and not simply problematic.

### *Nature of Spatial Structure*

Spatial statistics are also used in landscape ecology when an explicit goal is to understand the nature of the spatial structure of a particular variable, or set of variables, and to test hypotheses about that spatial structure. Quantifying variability in ecological measures over space (and time) is complementary to estimating average values or central tendency (e.g., Benedetti-Cecchi 2003; Fraterrigo and Rusak 2008). Variability can be highly sensitive and capture effects that are obscured by averaging. Although variance estimates are used less frequently than categorical landscape metrics for hypothesis testing in landscape ecology, use of measures of spatial variation are increasing (Legendre et al. 2002, 2004). For example, spatial statistics have been widely used to evaluate hypotheses about how disturbances or land-use history alter the magnitude or scale of variability in an ecological response (e.g., Gross et al. 1995; Lane and BassiriRad 2005; Mayor et al. 2007; De Jager and Pastor 2009, among others). For instance, Fraterrigo et al. (2005) found that historic agriculture in the Southern Appalachian Mountains was associated with altered spatial variability in soil resources, even though average soil variables were comparable with undisturbed areas. Soil resources varied over very fine spatial scales in undisturbed forests, but that spatial variability was homogenized in forests with historical land use such that soil resources were correlated over broader spatial scales (Fraterrigo et al. 2005). The magnitude and scale of spatial dependence in one variable also may be used as predictors for another variable, as illustrated by Gundale et al. (2006), who hypothesized that the spatial structure of variability in soil nitrogen would predict plant responses.

### *Spatial Interpolation*

Another use of spatial statistics in landscape ecology is spatial interpolation. Some variables simply cannot be measured everywhere, yet we may wish to estimate values at locations that were not sampled and/or produce a continuous surface, or map, of expected values. If the spatial structure of response variables is known, information about their autocorrelation can be used to predict their expected values at unmeasured locations—as Krige did to predict the likely locations of ore deposits. Landscape ecologists have developed spatial interpolations for a wide range of phenomena and scales. Bolstad et al. (1998) used kriging (among other methods) to predict vegetation patterns in a 2185-ha forested landscape in the Southern Appalachians. Smithwick et al. (2012) used kriging at much finer scales, predicting postfire rates of net nitrogen mineralization, abundance, and composition of soil microbes and aboveground vegetation within four 0.25-ha plots in Greater Yellowstone (USA) from point measurements (Fig. 5.2).

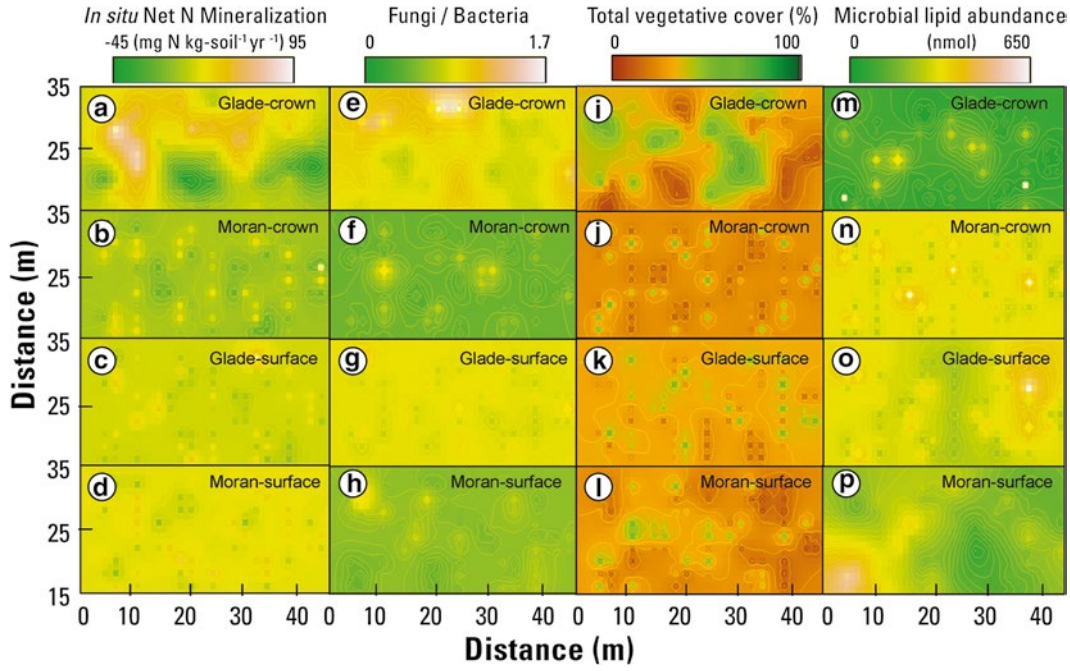


FIGURE 5.2.

Kriged maps of (**a–d**) in situ net nitrogen mineralization rate, (**e–h**) fungi:bacteria ratio, (**i–l**) total vegetative cover, and (**m–p**) microbial lipid abundance in four postfire study plots in Greater Yellowstone. Data were sampled at points, and semivariogram analysis was used for spatial interpolation.

FROM SMITHWICK ET AL. (2012).

Thus, in landscape ecology, spatial statistics provides important quantitative analysis tools to appropriately analyze continuous spatial data and to test new hypothesis regarding the magnitude, scale, and patterns evident in these data. Consequently, knowledge and use of spatial statistics are essential for studying pattern-process relationships in landscape ecology.

## CAVEATS FOR USING SPATIAL STATISTICS, OR "READ THIS FIRST"

As we did for landscape metrics in Chap. 4, we begin by providing guidance about what to consider *before* embarking on an analysis using spatial statistics. As always, it is critical that the question or rationale for analysis be clearly specified at the outset. The importance of beginning with a good scientific question can never be overstated!

### *#1. The Spatial Dependence in Landscape Data Must Be Characterized and Considered*

The need to recognize and test for spatial autocorrelation in data may be the most important take-home message from this chapter. The advances in spatial statistics and the understanding of how spatial dependence can affect conclusions drawn from data require a "heads-up" approach to this issue. Data recorded at intervals along transects and/or grids, especially along topographic or other environmental gradients, will often display spatial dependence over considerable distances (e.g., Everson and Boucher 1998; Nelson et al. 2005). In fact, it is difficult to imagine any landscape data that will not display spatial dependencies at some scale. As we showed with the simple example of plant heights, failure to characterize these dependencies can lead to serious biases in analysis and incorrect conclusions. It is incumbent upon the scientist to realize that spatial dependence is a potential issue, to appropriately test the data or models for spatial dependence, and to adjust the analysis accordingly. Such data exploration should be a routine component of statistical analyses, just like examination of other estimates of model strength and goodness of fit. If there is no spatial dependence, there should be a sentence in the manuscript stating that conclusion (e.g., Simard et al. 2012 do this). The old adage, "ignorance of the law is no excuse," will apply; be assured that reviewers and editors will follow up.

### *#2. Spatial Autocorrelation Is Not Always a Problem*

Although spatial autocorrelation is often considered a nuisance that interferes with testing interesting relationships between predictors and response, it is important to recognize that spatial dependence can provide ecologically relevant information.

Changes or differences in the spatial scale of autocorrelation can indicate changes or differences in the processes that have generated the observed patterns. Early spatially explicit studies of soils demonstrated that nutrient pools and transformations varied tremendously over small distances (e.g., Robertson et al. 1988), and that spatial variability could change through time (Gross et al. 1995; Cain et al. 1999). A study of postfire succession and soils illustrates hypothesis testing using spatial dependence as a response. Turner et al. (2011) asked how the variability and structure of aboveground vegetative cover and soil nitrogen availability changed during the first 4 years following stand-replacing fire. They laid out a set of expectations that could be tested with spatial data. For example, they expected little initial spatial structure in soil nitrogen variables because fire effects were likely to be spatially random, but that spatial structure would develop with vegetation recovery during early postfire succession (Turner et al. 2011). Results revealed a surprising absence of spatial structure in soil nitrogen transformations at the scales sampled. For biotic cover, the scale of autocorrelation was expected to increase over time, but it remained similar during the first 4 years postfire.

Nested scales of variability can suggest that the environmental factors structuring variability operate at different scales (e.g., Franklin and Mills 2003). Processes that homogenize local variation at a particular scale can induce spatial autocorrelation that then emerges at coarser scales (as shown by Fraterrigo et al. 2005). Furthermore, the spatial scale over which two variables are correlated may indicate an underlying process that is worth exploring. For example, analyses by Keitt and Urban (2005) using *wavelets* showed how topographic variation in physical drivers (e.g., sunlight, water availability) interacted to produce complex, scale-dependent patterns in vegetation growth (Keitt and Urban 2005). In short, the analysis of spatial data by autocorrelation methods can provide important insights into the spatial structure of both response and driver variables.

### *#3. Coincidence of Scales of Spatial Dependence Among Multiple Variables Does Not Prove Causality*

The coincidence of scales of variability of different ecological features, such as plants and soil nutrients (Grieg-Smith 1979) or seabirds and their prey (Schneider and Platt 1986) may indicate linkages worth exploring. However, it is important to remember that coincidence of the spatial structure does *not* prove causality, but rather suggests reason to test for causal mechanisms. As is always true when correlations exist among variables, statistical dependencies imply but do not prove pattern-process dependencies. In applying spatial statistics, it is tempting to surmise that variables having similar spatial structure may be responding to similar processes or that the variables may interact with one another. Here, the art of analysis will be instrumental for relating a coincidence of scales to potential mechanistic relationships.

The grain and extent of the data used in spatial statistics will influence the results, much as they influence landscape metrics. For continuous data, *grain* is usually the minimum or characteristic distance between sample points, whereas *extent* refers to the total area sampled, typically the linear dimension defining the sampled area. The characteristic distance between the points must be equal to or less than the scale at which the analyst wants to detect spatial structure. The overall linear extent of the dataset should be at least twice the maximum distance the analyst wishes to examine. Thus, to explore spatial dependence over scales of 1–100 m, sampling points should be separated by ≤1 m, and the overall data dimension should be at least 200 m. In general, fine-grained spatial structure cannot be determined from coarse-grained data, and broad-scale patterns cannot be adequately detected with data of limited spatial extent.

As for patch-based metrics, boundary effects are an important consideration when determining scale-dependent effects. All landscapes are of finite size and, because of this, edge effects dominate estimates derived from data that lie near boundaries. When these boundaries are artificial and the landscape is relatively small, truncation effects occur and may dominate estimates. Because these truncation effects are especially serious for point pattern analysis (discussed below), special correction factors may be required. Many of the scale issues associated with data collection can be reduced or eliminated if an optimal sampling plan can be devised ahead of time. Fortin and Dale (2005:14) provide guidance for developing landscape sampling designs, and we discuss some approaches later in the chapter. The purpose should be to obtain a match between the goal of the study, the spatial and temporal grain and extent of the data, and the validity of the statistical methods to be used for analysis.

Practically speaking, it is critical to check the analysis scales used when computing spatial statistics, especially when using readily available software. For ease of computation and display, software programs may automatically "bin" the data into 10–20 classes representing different separation distances among the data points. If the extent of the data is large, the characteristic bin sizes will also be large and may fully contain smaller separation distances that may be ecologically relevant. The unsuspecting analyst might obtain results using overly large bins that obscure, rather than detect, the finer-scaled variation. Therefore, the analyst must know the finest resolution needed for the analysis and the maximum separation distance that should be considered, then confirm these settings before results are determined. The bottom line is that you must carefully select the scales needed for data collection and analysis, then fit the models accordingly.

### *#5. Stationarity Is an Important Assumption in Many Spatial Statistical Analyses*

The assumption of *stationarity* states that the mean and variance of a particular variable will not change with the location of measurements. This assumption—technically referred to as first- and second-order stationarity—often catches those new to spatial statistic unawares. Stationarity is likely to be violated when there is an unrecognized gradient in the data (e.g., increasing precipitation or soil nutrient availability), when the underlying sampling methods are of limited extent and fail to measure changes in variables with scale, or when patterns are *anisotropic* (a marked directionality in the data). When non-stationarity exists and is unaccounted for in a statistical analysis, essential statistical parameters (such as mean and variance) will not be universal but location dependent. In landscape data, topographic relief can impose gradients and directionality on spatial data, both of which can violate stationarity assumptions. The usual response to this issue is either to detrend the data (commonly done for time series analysis via linear regression) or to account specifically for these gradients in the statistical model. Spatial analysis programs often assist with this adjustment when correcting for simple, linear trends. However, the removal of nonlinear trends remains difficult and is a continued area of investigation (Fortin and Dale 2005). Some spatial analysis programs provide the option for either isotropic (all-directional) or anisotropic (directional) analyses; the latter allows the user to evaluate spatial dependence and compare model parameters across a range of compass directions.

### *#6. Interpreting Spatial Statistics Is Both a Science and an Art*

As with other kinds of statistical methods, many analysis decisions are up to the user and can be perplexing. For example, some methods (including semivariograms, which are discussed in detail below) involve fitting a theoretical curve to the empirical data and then estimating parameters from this curve to describe the spatial structure in the data. There are a number of different standard curves that may be fit to the data (e.g., spherical, exponential, sinusoidal, linear models), but each may provide strikingly different parameter estimates for characterizing the spatial structure. Selecting the most appropriate model can be done in several ways, but the user ultimately must choose and justify her/his choice. If a study entails estimates of spatial dependence for multiple variables or among multiple landscapes, different variables or plots may achieve their best fit with different models. The analyst must then decide whether to use the best model each time, knowing that the models will vary; or to use the same model for consistency, knowing that the fit will vary. Either way, the quantitative estimates derived from the analysis can be considerably different.

Another issue of interpretation is that empirical data can be very noisy and, thus, not conform nicely to the theoretical curves. How to interpret apparent cycles in the data, or points that fall outside broad confidence intervals, can be challenging. The results of spatial statistical analyses with real landscape data are often much more ambiguous than results of an ANOVA that are based on *F*-ratios with set probability limits. If you encounter such situations, you are not alone. Do not be deterred, but do seek statistical advice from someone experienced with the application and interpretation of these analyses. Because these methods have and will continue to advance, these tasks will soon become more convenient, efficient, and reliable. However, there is no single prescription or cookbook approach—rather each new problem may require novel methods with their associated sets of assumptions and restrictions. Therefore, the user must take care when applying statistical techniques and drawing conclusions from results and remember that the question or objective of the analysis must be unambiguously specified ahead of time.

## POINT PATTERN ANALYSIS

The data for point pattern analysis methods are composed of records of event-based spatial phenomena, such as the location of individuals of an invasive species (Schreiber and Lloyd-Smith 2009); the presence of kangaroo rat nest mounds (Schooley and Wiens 2001); or occurrence of lightning caused fire ignitions (Podur et al. 2003). Point data are irregularly distributed in space and characterized by $x$, $y$ coordinates (the points) with variable supplemental information (the marks) to identify the type of event and relevant biological or physical attributes associated with that point (e.g., species, age, size, soil type, etc.). Once acquired, these spatial point pattern data allow hypotheses linking spatial pattern (the points) to ecological process (the marks) to be examined. Special techniques have been developed to describe these data and to test their association with relevant biological and physical attributes of the landscape (Perry et al. 2006).

Point pattern analysis has a long history of study in ecology. Clark and Evans (1954) first suggested that the mean distance to a nearest neighbor divided by the mean distance of randomly distributed points provided a normalized measure of the departure of spatial patterns from those expected by simple random processes. These calculations, when done by hand in the ‘50s, were computationally difficult, especially when data sets were large. A different approach was suggested by Greg-Smith (1952) who aggregated point data into quadrats and then used the differences in frequencies between adjacent quadrats as a measure of spatial pattern. Successive increases in quadrat size allowed scale-dependent changes in variance to be estimated (O’Neill et al. 1991; Levin 1992) (see Fig. 5.3). More efficient sampling methods have since been developed to provide more robust estimates (e.g., lacunar-

ity analysis; Plotnick et al. 1993, 1996), but the basic concept of change in variance with scale as a measure of pattern-scale dependency remains useful.

The Clark and Evans (1954) nearest-neighbor calculation produced a single value for each dataset. Ripley (1977, 1979) introduced the *K* function, which examines nearest-neighbor associations over all distances within the dataset, providing a scale-dependent measure for point patterns without the need to aggregate data into quadrats. The *K* function, as described by Crawley (2007), is calculated as:

$$K(d) = \frac{1}{N^2} A \sum_j \sum_{i \neq j} \frac{I_{(d)} \langle d_{ij} \rangle}{w_{ij}},$$

where $N$ is the number of points; $A$ is the size of the study area that contains the points; $d_{ij}$ are the subset of distances that are less than $I_{(d)}$ (values of $I_{(d)}$ will range successively from the minimum to maximum possible distance within $A$); and $w_{ij}$ is an edge correction required to avoid truncation effects when a given point is near the boundary ($w_{ij} = 1$ when a circle of radius $I_d$ around point $ij$ does not contact the boundary of $A$, and <1 when boundaries are encountered).

All these calculations can be easily done within the *spatstat* library (Baddeley and Turner 2005) of R (R Development Core Team 2010). Baddeley and Turner (2005) provide an example of analysis with Ripley's *K* for a point pattern dataset of Swedish pine trees (Strand 1972; Ripley 1988). The nearest-neighbor analysis reproduced here (Fig. 5.4) shows two lines. The solid line is the expected value at each distance class, $I_{(d)}$, if all points were randomly distributed; the segmented line shows significant deviation of a range of $I_{(d)}$ from ~6 to ~12 (see Baddeley and Turner 2005 for the statistical methods used to test for significant departures from random), indicating that trees were more uniformly distributed than the "expected" (hypothetical) random pattern. The causes of spatial regularity over small distances may include processes such processes as nonrandom seed recruitment or mortality, competition or a nonrandom substrate. The results shown in Fig. 5.4 define the

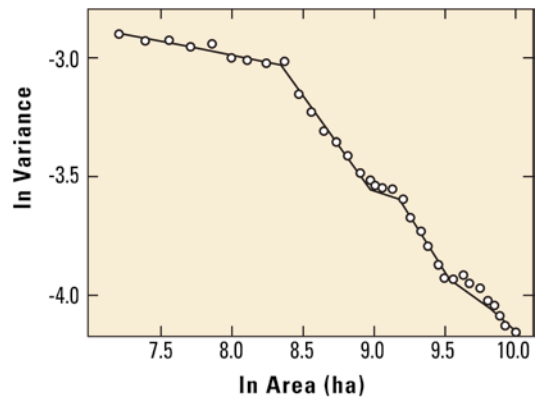


FIGURE 5.3.
Variance in percent grassland in a landscape near Goodland, Kansas changes over a range scales. These changes created a stair-step pattern hypothesized to result from human activities that vary from fields to farms, townships, and counties.

ADAPTED FROM O'NEILL ET AL. (1991).

scales over which future studies may concentrate to identify the specific processes that have resulted in these nonrandom patterns. Of course hypothesis testing requires the uncertainty around these numbers to be estimated. The *spastat* library provides an example of how to estimate uncertainties using Monte Carlo methods.

Perry et al. (2006) and Diggle (2003) provide excellent discussions of other methods for analyzing spatial point patterns. No matter the choice of method, one should be aware that when the study area, $A$, is very large multiple (new) ecosystems or cover types may be encountered. If points are restricted to a single ecosystem or cover type, then patterns will always appear to be spatially heterogeneous with areas of aggregation within preferred habitats separated by areas of overdispersion; conversely, if spatial extents are very small, then edge effects will be large and results biased despite use of edge correction methods that adjust spatial weights, $w_{ij}$. As always, there is an intimate dependency between the nature of a dataset, the scale of analysis, and the adequacy of the statistical methods used for description and hypothesis testing.

## AUTOCORRELATION AND VARIOGRAPHY

The heart of spatial statistics is the concept of correlation of spatially distributed variables. Autocorrelation and variography are two widely used methods for characterizing spatial dependence, or spatial structure, in a variable as a function of its position in a landscape. These methods are considered *global spatial statistics* because they estimate the intensity of spatial dependence for the entire study area (Fortin and Dale 2005). Both methods also provide estimates of the spatial scale(s) over which data are dependent. Factors that cause spatial dependence may be intrinsic to the system being studied, such as dispersal and competition in biotic

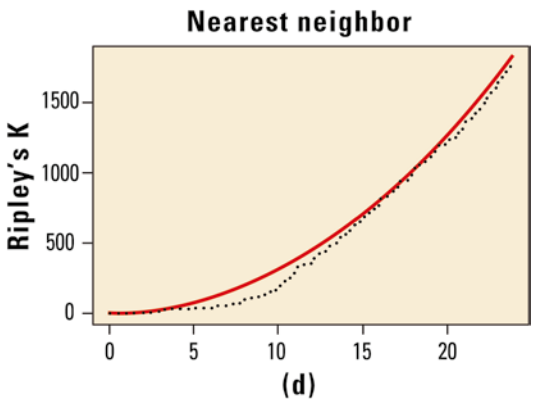


FIGURE 5.4.
Nearest neighbor analysis of the Swedish pine data using Ripley's $K$ and the *spatstat* library of Baddeley and Turner (2005). See text for details of analysis.

communities, and often manifested as fine-scale patchiness. This is certainly the case for measures of species abundances, which are always positively autocorrelated (Lichstein et al. 2002). Factors external to the system, such as topographic gradients of moisture, nutrients, and light, may also induce correlations (as noted above) but these are usually responsible for broad-scale trends (Legendre 1993; Lichstein et al. 2002). Conceptually, any observed spatial structure is a mix of induced spatial dependence (i.e., how the variable is responding to the spatial structure of an exogenous process) and inherent spatial autocorrelation (i.e., intrinsic to the variable or process of interest), but estimates of spatial autocorrelation coefficients cannot discriminate between these different components (Fortin and Dale 2005).

### *Autocorrelation*

Spatial autocorrelation is estimated by taking the average squared difference between all points separated by a given distance, $h$, the "lagged" distance between points. If we assume second-order stationarity (constant mean and variance over the entire dataset), then the autocorrelation for each distance between points, $r_{(h)}$, may be calculated as a Pearson correlation coefficient:

$$r_{(h)} = \frac{\sum_{i=1}^{N-h} (z_i - \bar{z})(z_{i+h} - \bar{z})}{\sigma^2},$$

where $z_i$ is the value at location $i$; $z_{i+h}$ is the value of a point at a distance of $h$ from $i$; $\sigma^2$ is the overall variance; and $N$ is total number of points. A graph of $r_{(h)}$ vs. $h$ is a *correlogram* that provides a visualization of the change in the dependency between points as a function of the distance between them, $h$. In general, to assure numerical adequacy of estimates of $r_{(h)}$, $h$ should never be calculated for distances greater than ½ the total distance that was sampled, and there should be at least 50 pairs of points for each lag distance (Rossi et al. 1992). Indications of ecological scale can be verified by statistically testing the peak values of the correlogram (both positive and negative) for significant differences from zero (Carlile et al. 1989). The conditions for valid tests for the significance of these peaks are restrictive, requiring (1) that only points separated by $h$ are compared for each lagged distance, (2) that gradients of change or trends in the data be removed before correlations are estimated (see Legendre 1993 for other restrictions in the analysis of gradients and autocorrelated data) and (3) that the residuals be normally distributed (Legendre and Legendre 1998).

Because $r_{(h)}$ is normalized by its variance, values will always lie between –1.0 and 1.0, which can be advantageous when comparing correlograms for different variables and/or landscapes. When $r_{(h)}$ is near 1.0, the lagged values are positively related; when $r_{(h)}$ is near –1.0, the lagged values are negatively related; and when $r_{(h)}$

is close to 0.0, the values are spatially independent. Of course, statistical tests are used to show when values of $r_{(h)}$ do not differ from 0.0 and to determine the minimal separation distance, $h_c$, beyond which the data can be considered spatially independent.

Let's consider an empirical example. In an area that was severely burned by the 1988 Yellowstone Fires, the density of postfire lodgepole pine (*Pinus contorta* var. *latifolia*) seedlings was recorded sequentially in 1-m$^2$ plots ($N = 3395$) at every meter along a 3.4-km transect that spanned a large patch of burned forest. The raw count data (seedlings m$^{-2}$, depicted in Fig. 5.5a) clearly show variability along the transect. To quantify spatial structure of the postfire tree seedlings, autocorrelation of lodgepole pine seedling density was calculated for all pair plots separated by lag distances, $h$, ranging from 1 to 250 m to generate a correlogram (Fig. 5.5b). The analysis revealed very strong spatial autocorrelation over short distances and a

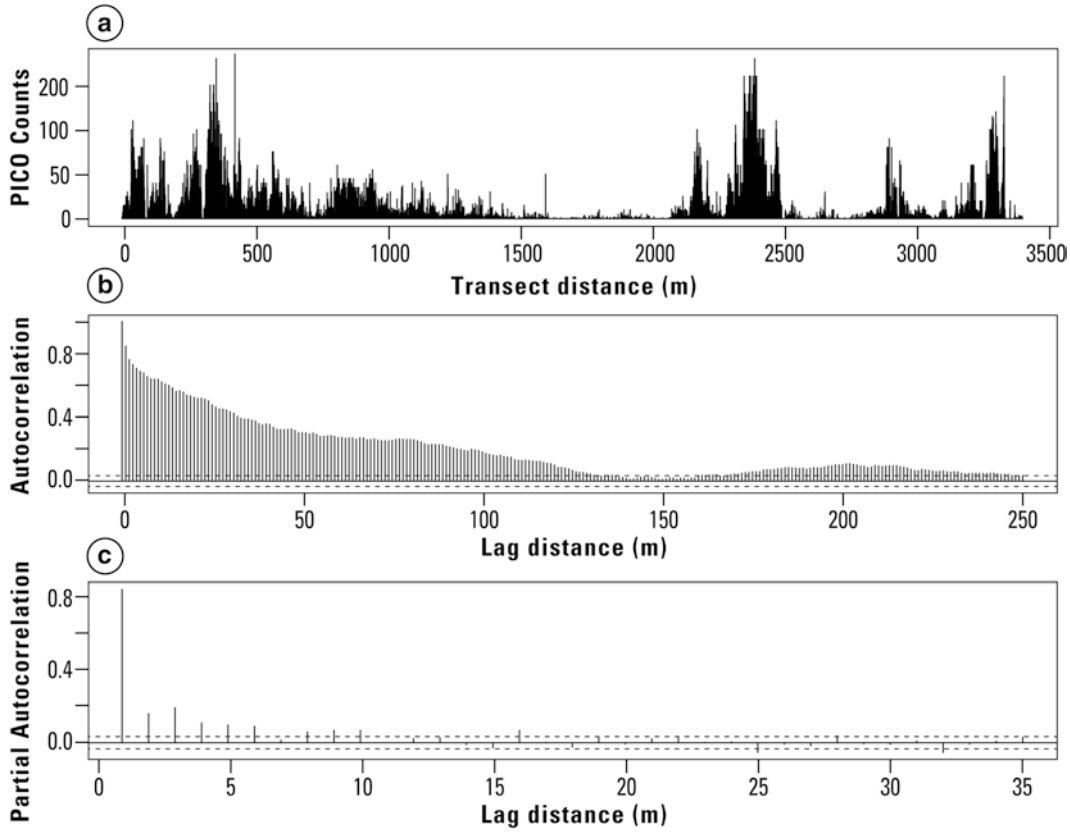


FIGURE 5.5.

(**a**) Continous counts of postfire lodgepole pine seedlings in 1-m$^2$ plots at 1-m intervals along a 3.4-km transect in Yellowstone National Park in 1991, 3 years after the 1988 fires (Turner et al. 1997b). (**b**) Autocorrelations and (**c**) partial autocorrelations of the lodgepole pine seedling counts

steady decline as lag distances increased. Correlation diminished with distance, declining to $r = 0.40$ at a lag distance of about 40 m, and to $r = 0.20$ at about 90 m. Because the sample size was so large, even very small correlations were statistically significant at $p = 0.05$, indicated by the blue dashed line in Fig. 5.5b. Thus, this example also demonstrates another important point: when sample size is very large, statistical significance doesn't always reflect ecological importance. Under such circumstances, one may set the limit of spatial correlation to a meaningful value; for instance, if $r_{(h)} = 0.2$ at a given distance $h$, then only ~4 % of the variance is accounted for by spatially correlated processes at that distance. Applying this to our example, we could infer minimal spatial autocorrelation between samples separated by at least 90 m.

Measures other than the Pearson correlation coefficient, such as Moran's $I$ and Geary's $c$, can also be used to test for spatial dependence (Fortin and Dale 2005). Interpretation of Moran's $I$ is very similar to interpretation of a correlation coefficient (Fig. 5.6). The magnitude of Moran's $I$, as well as the sign, are both important. The magnitude of the absolute value of Moran's $I$ (on the $y$ axis) in the correlogram of the random landscape (Fig. 5.6a) is about 0.12, indicating little spatial autocorrelation at any lag distance; in contrast, the maximum absolute value of Moran's $I$ in a landscape with a gradient (Fig. 5.6b) approaches 1.0, indicating very high spatial autocorrelation over some lag distances. In a landscape with repeated patterns (Fig. 5.6c), the strength of the autocorrelation is intermediate. The sign of Moran's $I$ is again informative. As with a traditional correlation analysis, a positive value indicates positive correlation, and a negative value indicates a negative correlation. When there is a repeated pattern, correlations will be positive at short distances, then become negative (peaks to valleys), then become positive again (peaks to peaks; Fig. 5.6c). Autocorrelation methods are useful for defining spatial dependencies and for discriminating between positive and negative autocorrelations, but they are unable to distinguish between certain kinds of spatial patterns, such as sharp step change vs. a gradient (Palmer and McGlinn 2016).

Pearson correlation coefficients may be used not only for their simplicity, but also because measures of the direct effect at each distance $h$ can be estimated by the partial autocorrelation coefficient (Fig. 5.5c). The partial correlation coefficient is the autocorrelation between points separated by $h$ that has first been adjusted for correlations at all other distances and therefore represents the unique effect of points $z_i$ on $z_{i+h}$. Figure 5.5c shows that the range of distances with significant direct effects (as measured by the partial correlation) is small (~10 m) in the lodgepole pine transect data, although accumulated indirect effects (as measured by the simple correlation, Fig. 5.5b) extend to ~130 m. Thus, abundance levels correlated over relatively short distances (the partial correlation) nevertheless result in patches of high abundance at scales of ~130 m (Fig. 5.5b).

When data for a spatial autocorrelation analysis come from a linear transect, directionality is not a concern. However, if the data come from a two-dimensional

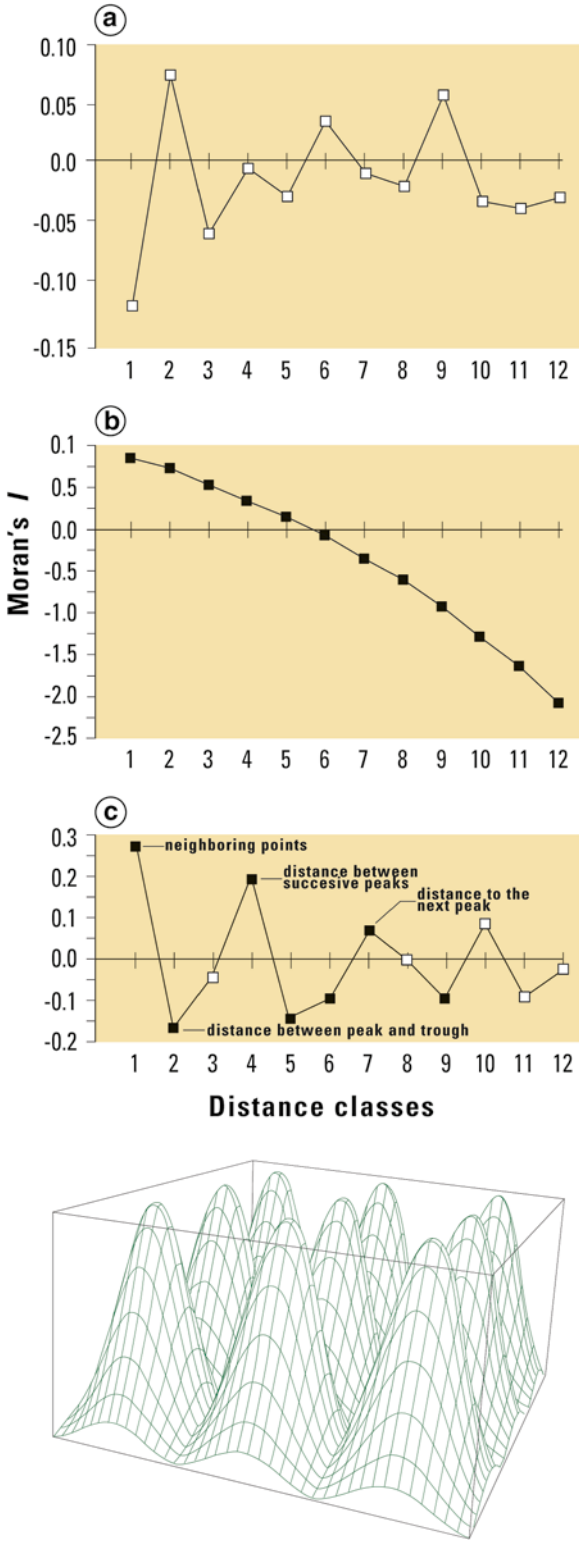


FIGURE 5.6.

All-directional spatial correlograms of artificial landscapes: (**a**) random landscape, (**b**) landscape with a gradient, and (**c**) landscape with a repeating pattern, the "nine fat bumps" shown below. Note that Moran's I behaves like a correlation coefficient.

ADAPTED FROM LEGENDRE AND FORTIN (1989).

area, directionality should be considered. For most analyses, the default is to derive the coefficients from an omni-directional (isotropic) analysis, meaning that all pairs of points at a given lag distance are used, regardless of the directionality between them. However, it is possible that the intensity of spatial dependence differs by direction, and analyses may be computed for pairs of points that lie in the same direction (anisotropy). For example, autocorrelation could be measured in both an east-west and a north-south direction and the two sets of coefficients compared.

### *Variography*

Variography is also based on spatial dependence among pairs of observations at different lag distances, but it uses a direct measure of variance in the computation and allows for spatial interpolation of point data across a landscape. Kriging techniques have been developed for this purpose with the *semivariance*, $\gamma_{(h)}$, providing the needed estimate of spatially dependent variance. The semivariance is equal to half of the squared difference of all pairs of points separated by distance $h$ (Rossi et al. 1992; Palmer 1992) and is calculated for each lag distance, $h$ as:

$$\gamma_{(h)} = \frac{N_{(h)}}{2} \sum_{i=1}^{N_{(h)}-l} \left(z_i - z_{i+h}\right)^2,$$

where $z_i$ is the value at location $i$; $z_{i+h}$ is the value of a point at a lagged distance of $h$ from $i$; and $N_{(h)}$ is number of pairs examined at lag distance $h$. Plotting $\gamma_{(h)}$ over all values of $h$ results in a *variogram*, the central tool of geostatistics (Rossi et al. 1992). In an idealized variogram (Fig. 5.7), semivariance is minimal when $h$ is small, then increases steadily until a distance is achieved where further increases in $h$ no longer cause the variance to increase. As with autocorrelation analysis, variography depends on several underlying assumptions, including absence of a trend in the $z$ values across space, that variance is constant across the dataset, that the precise location of each observation does not matter (only the distance between

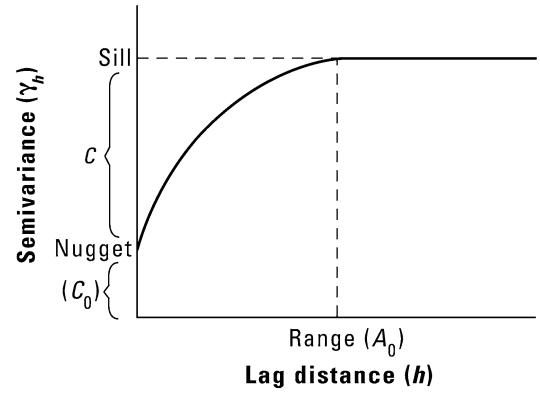


FIGURE 5.7.
Idealized semivariogram showing the nugget ($C_0$), sill ($C_0 + C$), and range ($A_0$); see text for explanation.

points is important), and (for an isotropic variogram) that the magnitude of $h$ matters, but not the direction.

Interpretation of spatial dependence using variograms is based on several key parameters. The asymptotic value of $\gamma_{(h)}$ as $h$ becomes very large is called the *sill*. The distance over which variance increases until the sill is reached is called the *range* ($A_0$, Fig. 5.7). The range is one of the most important parameters extracted from a variogram because it defines the distance (or scale) of spatial dependence in the data; at distances greater than the range, the data are considered to be spatially independent. Finally, there is often some amount of variance that is not related to spatial structure in the data, at least over the scales that were examined. This variance is estimated by the $y$ intercept of the variogram and is called the *nugget*, typically denoted as $C_0$ (Fig. 5.7). The nugget includes variance at scales smaller than the minimum separation distance between points and variance attributed to changes through time as data were sampled, measurement error, and random sampling error. The structural variance, $C$, is the difference between the sill and the nugget. The overall magnitude of spatial dependence in the data can be estimated by computing the proportion of structural variance, calculated as structural variance divided by the sill, or $[C/(C_0+C)]$. When there is no nugget or $C_0$ is small relative to $C$, the proportion of structural variance is high, and there is considerable spatial dependence in the data. When $C_0$ is nearly equal to the sill, there is no spatial dependence in the data. When the semivariance continues to increase with lag distance, $h$, and does not level off, it means that there is spatial dependence over all measurement scales. In this case, estimates of the range, $A_0$, will exceed the maximum value of $h$; in other words, the range will be greater than the largest lag distance that was analyzed. Estimates of semivariance are unreliable when there are too few pairs of points for any given lag distance, which is often the case as $h$ becomes large. Thus, the sampling restrictions for variogram estimation are the same as those given above for autocorrelation. Variograms should be calculated only for lag distances up to half the actual distance over which the data are measured, and it is best to include at least 50 pairs of points when calculating semivariance for each lag distance.

To estimate the nugget, sill, range, and proportion of structural variance from a variogram, a theoretical model must be fit to the plot of $\gamma_{(h)}$ vs. $h$ (Fig. 5.8). Spherical, exponential, and linear models are commonly used theoretical models, and analysis programs will report $C$, $C_0$, and $A_0$ values. Different models may produce different numerical estimates for these parameters, and the best choice for a theoretical model may be difficult to determine a priori. Most practitioners will inspect the shape of the empirical variogram, then fit several models to determine which one provides the best fit by comparing the proportion of structural variance among models and examining $r^2$ or AIC values. This can be straightforward when a single variogram is being evaluated. However, when a study involves multiple variograms, different theoretical models may fit best for different response variables or plots (Fig. 5.8). The analyst must then decide whether to use one theoretical model

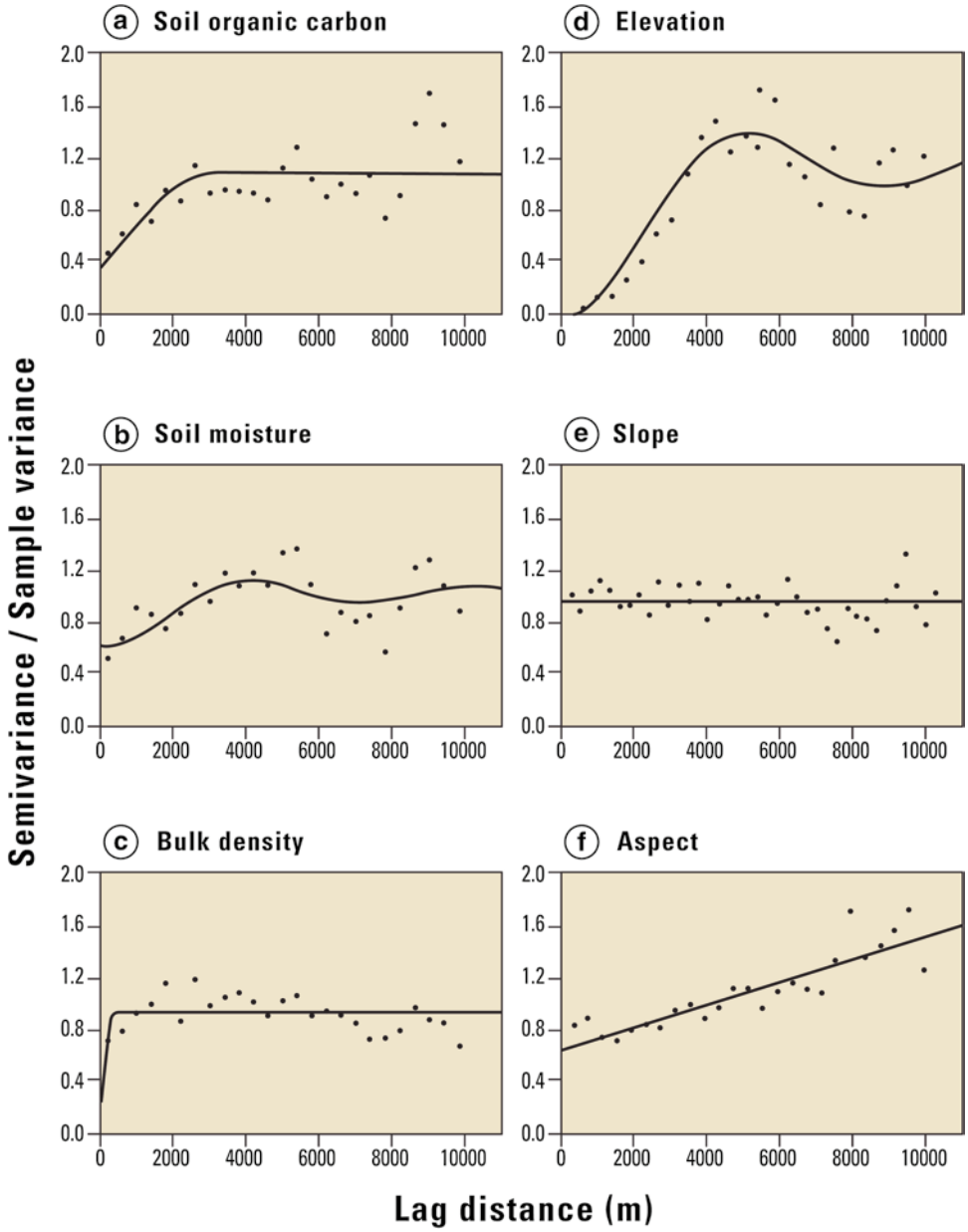


FIGURE 5.8.

Illustration of six models for fitted semivariograms using data from the Luquillo Experimental Forest of Puerto Rico: (a) soil organic carbon, fit with a spherical model; (b) soil moisture, fit with a wave/hole model; (c) bulk density, fit with a spherical model; (d) elevation, fit with a wave/hole model; (e) slope angle, fit with a random model; and (f) aspect, fit with a linear model. The equations for the illustrated semivariance models are: Random $\gamma =$ mean sample variance; Linear $\gamma(h) = C_0 + C[(h/A_0)]$; Spherical $\gamma(h) = C_0 + C(1.5(h/A_0) - 0.5(h/A_0)^3]$, $h \leq A_0$; $\gamma(h) = C_0 + C$, $h \geq A_0$. Wave/hole $\gamma(h) = C_0 + C\{1 - [\sin(A_0 * h) * h/A_0]\}$.

ADAPTED FROM WANG ET AL. (2002).

(e.g., the spherical model) for all data, thereby minimizing the effect of changing the model on parameter estimates, or to use the best model fit for each dataset. In our experience, different models for different processes are appropriate, but statisticians will differ in their opinions, and thinking about this topic may change—points the analyst should keep in mind.

The idealized curve for a variogram with a distinct scale of spatial autocorrelation is asymptotic, but empirical variograms may take a surprising number of shapes (Fig. 5.8); real-world data can be messy! If a spatial dataset is nonrandom and has been adequately sampled (i.e., the spatial extent of the data provides an adequate representation of the pattern of interest) then we expect the variogram to ascend from an initial value at $h = 0$ to an asymptotic value (Fig. 5.8a, d). Variograms that do not asymptote, but rather continue to increase as $h$ increases (Fig. 5.8e), indicate an underlying trend or nonstationary stochastic process (Crawley 2007) which must be accounted for before finer scale dependencies can be explored. A relatively flat horizontal variogram indicates a pattern that lacks spatial dependencies (i.e., a random pattern, Fig. 5.8c).

The form of the variogram does not lend itself to statistical testing, but confidence intervals for semivariance estimates can be calculated (Shafer and Varljen 1990; Zheng and Silliman 2000; Lin and Chen 2005; Xiao et al. 2005). Because the estimate of semivariance for each lag distance is obtained from multiple pairs of points, the variance around each estimate can be determined. It is increasingly common for a 95 % confidence interval to be plotted in empirical variograms. Although confidence intervals will often bounce around the sill, they are helpful in determining the range over which significant changes in variance may be expected (Xiao et al. 2005). Significance testing for each lag distance $h$ is not commonly done in variography (in contrast to correlograms), but the occurrence of a semivariance estimates below the 95 % confidence interval usually corresponds to a significant autocorrelation at that lag distance.

Spatial interpolation, or kriging (Rossi et al. 1992), uses the semivariogram to predict an expected value at unmeasured locations. Kriging methods predict $z$ by using a weighted average of the expected values based on the distance from other points, with the weights accounting for autocorrelation in the observed data. Results yield a best linear unbiased estimate of a variable at a given point. Kriging has found many ecological adaptations, including mapping vegetation community distributions (Arieira et al. 2011), determining patterns of exotic species invasion (Cilliers et al. 2008; Schreiber and Lloyd-Smith 2009) and locations of exotic weeds (Kalivas et al. 2012), and designing optimal spatial sampling methods (Xiao et al. 2005).

Although semivariance has many similarities to the autocorrelation function (Box 5.1, and see Fig. 1 in Palmer and McGlinn 2016 for helpful comparisons), there are several notable differences between these methods that affect their use and interpretation. Because correlograms emphasize the strength of the correlation

## BOX 5.1
## GENERAL RELATIONSHIPS BETWEEN THE AUTOCORRELATION AND SEMIVARIANCE OF VALUES LAGGED BY DISTANCE *H* (ADAPTED FROM ROSSI ET AL. 1992)

If the population mean and variance are constant (i.e., there is no trend) and $C(h) = \text{thecovariance} = 1/N(h)\sum(z_{-h} - \bar{z}_{-h})(z_{+h} - \bar{z}_{+h})$, where $-h$ and $+h$ represent the head and tail of data points separated by distance $h$, then:

$$\gamma(h) = \sigma^2 - C(h)$$

$$\rho(h) = C(h)/\sigma^2$$

$$1 - \rho(h) = \gamma(h)/\sigma^2$$

where $\gamma(h)$, $\rho(h)$ are the semivariance and autocorrelation, respectively, for points separated by distance $h$.

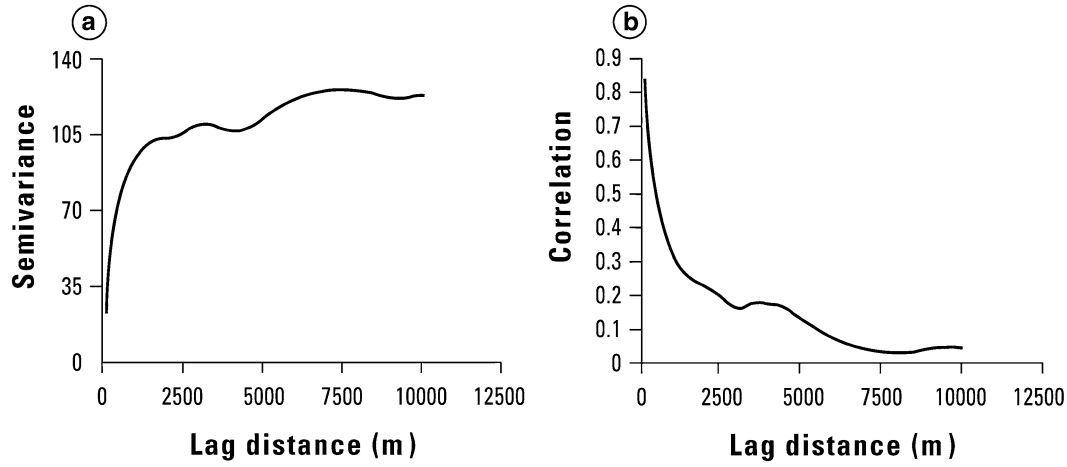


FIGURE 5.9.

Example semivariogram (**a**) and correlogram (**b**) computed for a landscape in northern Yellowstone National Park. Note that the shape of the correlogram is nearly identical, although inverted, to the shape of the semivariogram of the same data. These two approaches are complementary.

ADAPTED FROM MEISEL AND TURNER (1998).

at each lag distance and variograms report the magnitude of the variation (Box 5.1), the empirical plots are of opposite shape (Fig. 5.9). The semivariance estimates, which are not normalized by either the mean or the variance, may take on any positive value as all forms of variance also do. However, the magnitude of the semivari-

ance is informative when comparing variables measured in the same units—for example, one could ask whether the total amount variance in plant biomass is increasing across a chronosequence of plots, or is greater in certain landscapes than in others. When comparing variograms for variables that are reported in different units (e.g., plant biomass in g and soil nutrient properties in ppm), the magnitude of variance cannot be readily compared. In such cases, variograms may be standardized by dividing by the maximum semivariance ($C + C_0$) so that the $y$ axis scales from zero to one. This normalization allows the shape of the curve and the range estimates ($A_0$) to be easily read and interpreted, regardless of the units in which semivariance was reported. In contrast to correlograms, semivariograms provide no information about whether spatial dependencies are positive or negative because variances are always positive. Lastly, one might expect that *range* estimates would be the same for correlograms and variograms. This is not always the case because the semivariance measures an asymptote where variance no longer increases with distance, whereas autocorrelation measures multiple distances where autocorrelation is present rather than extracting a single dominant scale.

### *Cross-Correlograms and Co-variograms*

When multiple variables are recorded in either space or time then the correlations among these variables can be used to more accurately describe and predict the spatial patterns of one or more predicted variables (Kalkhan and Stohlgren 2000; Kalkhan et al. 2007). The additional information provided by these cross-correlations allows a reduction in the total number of samples needed to detect significant effects (Fortin and Dale 2005). Covariance analysis has long been an essential element of regression methods (Cressie 1991) and may also be used to advantage with kriging (referred to as *co-kriging* when cross-correlations are used) to improve estimates of spatial interpolations. Although co-kriging is a more complicated process and long-regarded as a difficult computation (Deutsch and Journel 1998), advanced methods are now available making the use of spatial cross-correlations among multiple predictor variables an attractive option (see gstat package in R Development Core Team 2010; Rossiter 2012). For instance, precise knowledge of the spatial distribution of perennial weeds, which flourish in cotton agriculture, would allow better-targeted management options to be developed and employed. If the distribution of weeds could be estimated as a function of environmental variables, measured at lower expense, then significant savings would be realized by the use of co-kriging methods (Rossiter 2012). Kalivas et al. (2012) used this approach to reduce prediction errors over ordinary kriging (Fig. 5.10) with co-kriging providing better estimates of the distribution of bindweed and purple nutsedge, two difficult weeds to control (Kalivas et al. 2012).

More advanced statistical methods hold great promise for landscape studies but, unfortunately, are beyond the scope of this chapter. The reader may wish to refer

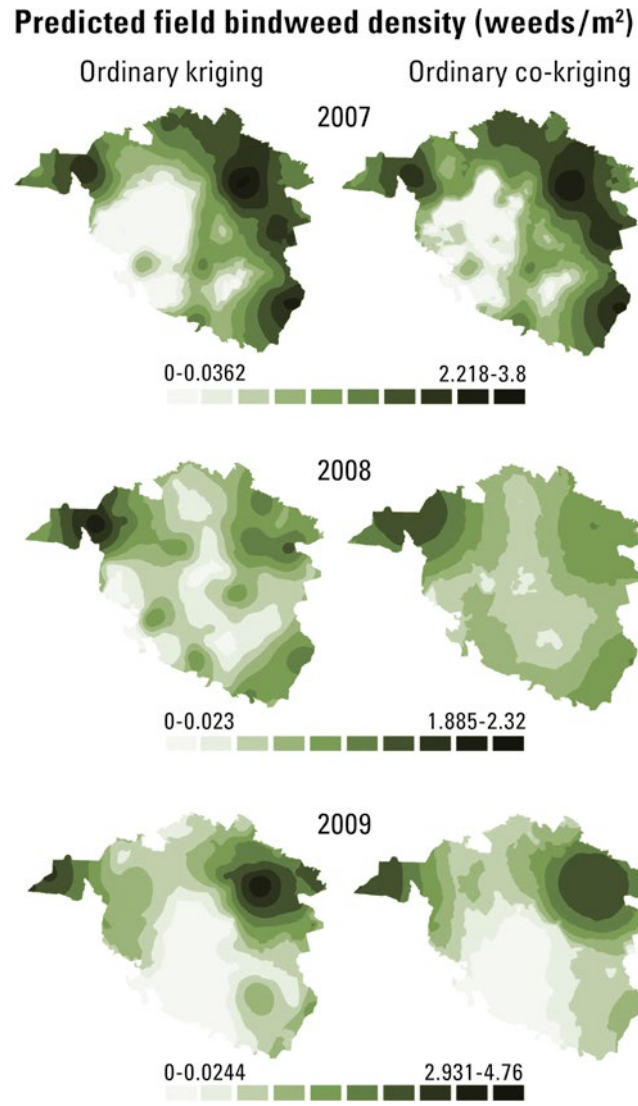


FIGURE 5.10.
Interpolated maps of field bindweed density with the use of ordinary kriging (**a**) and co-kriging (**b**) for three different years (Reproduced with permission from Kalivas et al. 2012).

to appropriate texts (e.g., Cressie 1991; Fortin and Dale 2005; Ripley 2005) and articles (e.g., Wagner and Fortin 2005; Perry et al. 2006), which provide guidance and caveats for use of additional multivariate methods to describe and predict spatial patterns.

### *Optimized Sampling Designs for Spatial Statistics*

When landscape ecologists think about evaluating spatial dependence using a field study, the default plan is often to sample continuously along transect (as was done for the lodgepole pine seedlings in Fig. 5.5) or within a full uniform grid of points. However, these are inefficient, not only because they take a lot of work, but also

TABLE 5.1.

EXAMPLES OF CYCLIC SAMPLING DESIGNS (ADAPTED FROM BURROWS ET AL. 2002), WHERE $x$ IS THE LENGTH OF THE CYCLE AND SAMPLE LOCATIONS ARE GIVEN IN THE SAME UNITS (E.G., METERS).

| Cycle definition | Length of cycle ($x$) | Number of plots sampled per cycle | Sample locations (0 to $x-1$) | Illustration of sampling cycle (+indicates sample location, o is not sampled) |
|---|---|---|---|---|
| 2/3 | 3 | 2 | 0, 1 | + + o |
| 3/7 | 7 | 3 | 0, 1, 3 | + + o + o o o |
| 4/13 | 13 | 4 | 0, 1, 3, 9 | + + o + o o o o + |
| 5/21 | 21 | 5 | 0, 1, 4, 14, 16 | + + o o + o o o o o o o o o + o + o o o o |

A 3/7 CYCLE INDICATES THAT THREE PLOTS IN EVERY SEVEN ARE MEASURED. WHEN REPEATED, THE SPACING OF THE THREE PLOTS WILL GIVE PAIRS OF PLOTS SEPARATED BY ONE, TWO, THREE, FOUR, FIVE, SIX, AND SEVEN LAG DISTANCES

because they produce data in which the point pairs for short lag distances are overrepresented, and the point pairs for long lag distances are underrepresented. Sampling designs that systematically vary the distances between sample points then repeat that sequence can produce a dataset with similar numbers of point pairs across all lag distances (Table 5.1). These clever designs will give comparable statistical power over a range of lag distances and are thus a valuable alternative to systematic sampling schemes. Because the repeated distances are considered to be a cycle, these sampling designs are often referred to as *cyclic*. As an example, consider sampling a 1000-m transect at 1-m intervals with traditional protocols that would produce 1000 data points. If 4/13 cyclic sampling was used instead in place of complete sampling (Table 5.1), then approximately 300 sampling points could efficiently measure a wide range of lag distances. There is detailed treatment of these methods in Cressie (1991), but we also recommend Burrows et al. (2002) for an accessible explanation and discussion. Burrows et al. (2002) compared different approaches (random, uniform, and cyclic sampling) for quantifying the spatial pattern of leaf area index (LAI) measures in terrestrial ecosystems that surrounded an eddy flux tower. They found a 60 % reduction in effort required for cyclic sampling designs vs. random and uniform sampling. As with all study designs, pilot data can be extremely valuable for figuring out the best cycle of distances to repeat.

In another example, Turner et al. (2011) used a cyclic design to determine how aboveground vegetation and soil nitrogen availability changed within forested plots following stand-replacing fires. Because these measurements are very labor intensive and the laboratory analyses costly, an alternative to sampling a full grid was desired. In each of four plots, they used a cyclic design ($n = 81$ points) with a minimum separation distance of 2 m between sampling points and a cyclic sequence

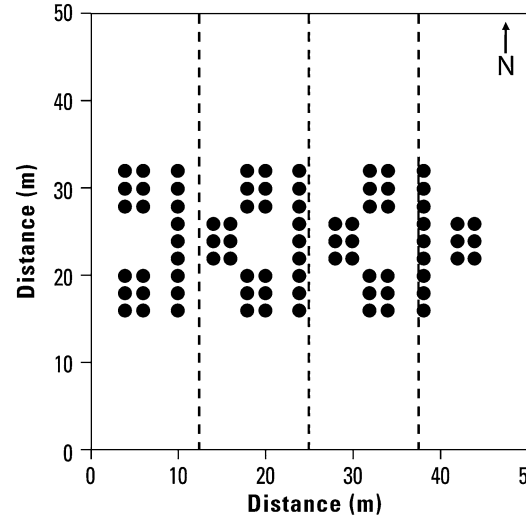


FIGURE 5.11.
Illustration of how a 3/7 cyclic sampling design (see Table 5.1) was implemented to sample soils and aboveground vegetation in Greater Yellowstone (adapted from Turner et al. 2011). The 3/7 scheme is applied in the horizontal direction and reverses direction for the middle three transects.

of points that was repeated (Fig. 5.11). The grid covered an 18 m × 40 m area positioned in the center of each 0.25-ha plot and included nine parallel rows, each separated by 2 m. The 2-m spatial resolution was based on the observed patchiness of aboveground vegetation given our focus on the relationship between vegetation and soil nutrients. Each row included three 3/7 cycles that were 14 m long in which 3 of 7 grid points (the 0, 1, 3 design in Table 5.1, which translated to samples at 0, 2 and 6 m in each cycle) were sampled; the middle three rows were offset by 6 m to account for potential anisotropy. Variography was then used to assess spatial dependence in vegetation and soil variables over 4 years. Using spatial statistics to improve sampling design is extremely valuable in any studies that include the goal of detecting the magnitude and scale of spatial variation in the data.

## EXAMPLES OF SPATIAL STATISTICS IN LANDSCAPE ECOLOGY

Because spatial statistics are less familiar than landscape metrics for many landscape ecologists, and therefore not used as frequently, we conclude this chapter with a few additional examples to illustrate the diversity of applications in landscape ecology. As is the case throughout this book, the selected examples are not intended to exhaustive but rather to illustrate how these methods can be used to gain insights into landscape-level questions. We also suggest software resources that readers may find useful (Box 5.2).

Spatial statistics have been used in several studies assessing organism responses to spatial variation in their habitat. The persistence of elephant populations in

## BOX 5.2
## THE PRACTICAL SIDE: SELECTED SOFTWARE FOR SPATIAL STATISTICS

There are many options for computing spatial statistics, including utilities within widely used GIS software programs such as ArcGIS. We highlight a few here. There are many online resources that provide current developments in spatial ecology including Wikipedia http://en.wikipedia.org/wiki/Spatial_ecology with lists and links to open-source GIS software (http://en.wikipedia.org/wiki/Category:Free_GIS_software). For hands-on learning, we recommend two chapters in the second edition of Learning Landscape Ecology. Exercises developed by Palmer and McGlinn (2015) lead students through the calculation and interpretation of correlograms and semivariograms using Excel and R; this is a great starting point. Exercises developed by Turner and Simard (2016) are more advanced, use GS+ (see below), and are designed to compare and contrast insights gained from traditional landscape metrics and spatial statistics.

**R Software**. There are a wide variety of utilities available for reading, writing, display, and analysis of spatial data in R. The R software and associated libraries may be explored and downloaded at http://www.r-project.org/foundation/. An overview of the many resources within R may be found at http://cran.r--project.org/web/views/Spatial.html. Libraries used in this chapter include *spatial* and *spastat* for point pattern analysis and *gstat* for geostatistics.

**GS+** is commercially available software published by Gamma Design Software, LLC; see http://www.gammadesign.com/. GS+ is a comprehensive geostatistics program that is menu driven and easy to use, and it readily produces kriged maps from empirical variograms. Users should take care to specify the minimum separation distance, however, as this program defaults to bins that may be larger than desired. Demo versions of GS+ can be downloaded for free and used for a limited time.

**PASSaGE**, which is an acronym for Pattern Analysis, Spatial Statistics, and Geographic Exegesis, is a free and integrated software package for spatial analysis and statistics (Rosenberg and Anderson 2011). It incorporates a wide range of analyses, including point pattern analyses, correlograms, semivariograms, and many more. The developers work in computational biology, bioinformatics, and landscape genetics. See http://www.passagesoftware.net/index.php

Zimbabwe is being threatened by continued conversion of natural habitat into agricultural lands. Murwira and Skidmore (2005) wanted to determine whether and how the spatial distribution of the African elephant (*Loxodonta africana*) responded to the spatial heterogeneity of vegetation cover based on data obtained in the early 1980s and 1990s. The distribution of elephants was measured from direct counts, and vegetation cover was derived from remotely sensed data. Variography applied to these data produced measures of the dominant scale (i.e., the range) and intensity (the sill) of the normalized difference vegetation index (NDVI) in 3.84-km × 3.84-km windows distributed in a 61-km × 61-km landscape. The range and sill were then used as predictor variables of elephant occupancy within mixed-use landscapes and to determine mixtures that optimized probability of elephant occurrence (Fig. 5.12). Examination of trends in agricultural intensity

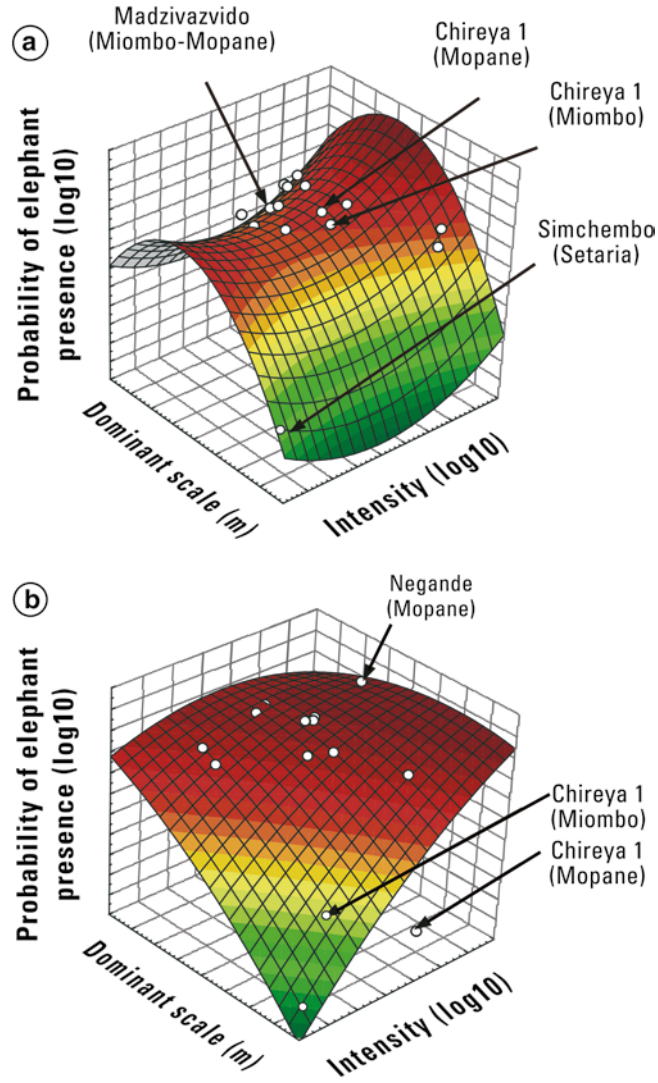


FIGURE 5.12.
The dominant scale and intensity of spatial heterogeneity in vegetation cover jointly influence the probability of elephant presence in (**a**) the 1980s and (**b**) the 1990s in an agricultural landscape in Zimbabwe.

ADAPTED FROM MURWIRA AND SKIDMORE (2005).

were found to be moving vegetation patterns away from an optimum mixture towards uniform agricultural areas which would not allow continued persistence of elephants within these landscapes.

Another example comes from a study of winter habitat use by woodland caribou (*Rangifer tarandus caribou*) in Newfoundland, Canada that blended geostatistical analyses with habitat selection (Mayor et al. 2007). Through careful analyses of multiple levels of habitat use (seasonal range, travel routes, feeding areas, and microsites) in response to snow depth and the abundance of lichens, Mayor et al. (2007) found that caribou reduced the variance in these key habitat features by selecting favorable habitat. By comparing variability of habitat components measured at four levels of habitat use (from feeding microsites to population winter

range) across a spectrum of spatial scales (from 1 to 28,000 m), Mayor et al. (2007) showed that caribou first selected for lichens at a broad scale (13 km) then selected areas with shallower snow at all scales.

In a landscape study of plant–animal interactions, De Jager and Pastor (2009) used spatial statistics to evaluate how available and consumed browse has changed with time in Isle Royale National Park, Michigan (USA) as the size of the moose (*Alces alces*) population declined. Four different variogram models were fit to measured levels of moose consumption, browse biomass, plant basal area, and soil fertility; each model representing a different type of landscape pattern from random to regularly arranged patches. The best models were then fit to patterns of browse availability, nitrogen availability, and conifer basal area. The low levels of moose populations have resulted in declining levels of annual consumption which, in turn, have resulted in random distributions of browse consumption: there were simply too few moose to impose detectable spatial patterns as occurred in the past. It appears that lower grazing levels have resulted in new process-pattern relationships with current vegetation patterns more related to variation in fertility, light availability through the canopy and canopy gaps, or seed dispersal patterns than to moose consumption (De Jager and Pastor 2009).

Spatial statistics are also useful in studying ecosystem properties such as nutrient pools and flux rates (e.g., Fig. 5.8, Wang et al. 2002). For example, Vasquez et al. (2012) explored the relationship between total soil carbon and hydrologic and biotic processes in a subtropical landscape in Florida, USA. They studied three nested regions and found that total soil carbon varied at two key scales, one over a relatively short range (5.6 km) in association with local soil and landscape variation, and another at a longer range (119 km) in association with regional variation. Spatial statistics have also detected pattern at multiple scales in a seascape and been used to map seagrass cover. In an estuary in eastern Canada, Barrell and Grant (2013) used acoustic data to detect aquatic vegetation, then used these data to map seagrass beds. Semivariograms were computed, and kriging was used to map seagrass cover and identify "hot spots" where seagrass cover was high relative to the mean.

The above examples employed methods related to variography. Many other examples may be found using autocorrelation (Ishihama et al. 2010; Liang 2012), spectral analysis (Jollineau et al. 2008; Huang et al. 2009), Bayesian statistics (Romero-Calcerrada et al. 2008; Fitzpatrick et al. 2010), and spatial modeling (Kellogg et al. 2008).

As the science of landscape ecology matures, studies will continue to rely on the considerable power of analysis available from methods based on spatial statistics. Because the data and methods of spatial statistics differ from those using landscape metrics, each approach will provide unique and important insights into the broad-scale patterns of ecological processes. The use and interpretation of categori-

cal landscape metrics are understood to a much greater degree than for spatial statistics. The latter seem more difficult for many students to grasp, and the parameter estimates more abstract. Studies that systematically compare results from these two approaches will be most effective in advancing our understanding of landscape dynamics. There is tremendous opportunity for developing richer interpretations of spatial statistics and exploiting the potential for hypothesis testing in landscape ecology. As georeferenced data become more readily available (e.g., through sensor networks and other remotely sensed measurements), there is much to be gained from greater exploration of how to enhance the use of spatial statistics in hypothesis testing.

## SUMMARY

Spatial statistics are widely applied in landscape studies for quantifying the spatial structure of spatially distributed data represented by real numbers. The diversity of methods available and the proliferation of jargon used in spatial statistics makes their appropriate use a demanding endeavor. Spatial statistics do not require the a priori categorization of landscape data, nor do they assume a patchy structure with delineated boundaries. Spatial statistics quantify the magnitude of variance in the data, the proportion of that variance that is spatially dependent (i.e., spatially correlated), and the distances over which variables are spatially dependent. These methods differ from, but are complementary to, those based on landscape metrics (Chap. 4).

Landscape ecologists use spatial statistics for wide variety of purposes, but three are particularly important. First, spatial statistics are used to test for independence in spatially distributed data prior to use of parametric statistics for hypothesis testing. When data are spatially dependent (i.e., the degree of correlation between observations changes as a function of distance), statistical tests for hypothesis testing may lead to Type I error (false rejection of the null hypothesis) unless corrective measures are taken. When data are spatially correlated, an adequate separation distance between samples may allow the assumption of spatial independence to be met. Alternatively, methods that adjust and remove spatial dependence prior to statistical testing may be used. Second, spatial statistics are used to quantify the nature of the spatial structure in continuous variables. The magnitude and scale of spatial dependence can be informative and used explicitly to test hypotheses about spatial structure. Third, spatial statistics are used for spatial interpolation, using values at sampled locations to predict values at locations that were not sampled or to produce a continuous surface or map of expected values.

There are several important points to keep in mind when analyzing spatial patterns. Reviewers of grant proposals and manuscripts now expect that the spatial dependence in a dataset has been considered and characterized. Although long considered a nuisance, spatial autocorrelation is not always a problem, but rather an informative attribute of the data leading to new insights and hypotheses. When multiple variables are considered, and similar scales of dependency are found, we now know that this, alone, does not prove causality. However, the coincidence of scales does suggest linkages that are worth exploring. As with all spatial data, the grain and extent of the data strongly affect results of any analysis. Of course, the analyst also must attend to the assumptions required for spatial statistics, including that of stationarity.

One set of methods used in spatial statistics is point pattern analysis, which is used for data formed from event-based records (e.g., location of nest sites). Geographically referenced events have had a long history of study in ecology, and approaches include quadrat-based measures that explore change in variance with scale. Ripley's $K$ function provides a scale-dependent measure that does not require aggregating data into quadrats.

Another set of methods is based on correlation among spatially distributed variables. Autocorrelation and variography measure changes in the relatedness or variance of continuous measurements (such as a rate) as a function of the distance, $h$, between measured points. Correlograms and variograms provide similar but not identical insights; correlograms indicate the direction and magnitude of autocorrelation, whereas variograms change in variance with distance. Variograms provide the means to interpolate data across space using parameters such as the nugget, sill, and range.

The downside of spatial statistics is that very large datasets may be required to assess the change in pattern across space. However, efficient sampling designs can reduce sampling redundancies (i.e., numerous samples taken at short distances and few samples at long distances) by systematically changing the density of sample points as $h$ increases. An illustration of the efficiencies gained using a cyclic sampling design is presented in this chapter and shows how samples may be arranged to greatly reduce sampling efforts.

As the science of landscape ecology continues to mature, studies will increasingly rely on the considerable power of analysis available from methods based on spatial statistics. Hypothesis testing based on spatial models and spatial statistical tests from replicate study areas is becoming the norm. There is much opportunity and considerable excitement for developing richer interpretations using spatial statistics and exploiting the potential for greater rigor in hypothesis testing in landscape ecology.

## DISCUSSION QUESTIONS

1. Gradients due to elevation change often produce spatial correlations in ecological data. How would one design a sampling scheme to separate these correlations from other factors due to biological process of interest (e.g., effects of dispersal patterns on species abundance).
2. Disturbances that change habitat will alter patterns of spatial association for species residing in those habitats. What methods of analysis would you use to characterize the change in scale following disturbance and then monitor the recovery of spatial dynamics with time? What data would be required for your selected method(s)?
3. What will a correlogram of species abundances look like in: (a) continuous, optimal habitat; (b) a landscape with a steep elevation gradient; (c) in a patchy, disturbed landscape?
4. Why does normalization of a statistic (e.g., correlogram) allow statistical tests to be efficiently (and rigorously) applied? Why are such tests more difficult for nonnormalized statistics (e.g., semivariogram)?
5. What is the difference between interpolation and extrapolation? How will the errors of these two different estimates differ? (hint: consider the error bounds for a linear regression).

## FURTHER READING

Burrows SN, Gower ST, Clayton MK, Mackay DS, Ahl DE, Norman JM, Diak G (2002) Application of geostatistics to characterize leaf area index (LAI) from flux tower to landscape scales using a cyclic sampling design. Ecosystems 5:667–679

Crawley MJ (2007) The R book. Wiley, Chichester

Dungan JL, Perry JN, Dale MRT, Legendre P, Citron-Pousty S, Fortin M-J, Jakomulska A, Miriti M, Rosenberg MS (2002) A balanced view of scale in spatial statistical analysis. Ecography 25:626–640

Fortin MJ, Dale MRT (2005) Spatial analysis: a guide for ecologists. Cambridge University Press, Cambridge

Legendre P, Dale MRT, Fortin MJ, Gurevitch J, Hohn M, Myers D (2002) The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography 25:601–615

Perry GLW, Miller BP, Enright NJ (2006) A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. Plant Ecol 187:59–82

Schwarz PA, Fahey TJ, McCullouch CE (2003) Factors controlling spatial variation of tree species abundance in a forested landscape. Ecology 84:1862–1878

Rossi RE, Mulla DJ, Journel AG, Franz EH (1992) Geostatistical tools for modeling and interpreting ecological spatial dependence. Ecol Monogr 62:277–314