Chapter 5
Scale Detection Using Semivariograms and Autocorrelograms

Michael W. Palmer and Daniel J. McGlinn

OBJECTIVES

The evolution and ecology of all organisms are contingent on the complex variation seen in nature. Landscape ecology differs from most other branches of ecology in that it explicitly involves spatial variation. Therefore, one of the goals of landscape ecology is to describe spatial variation. The purpose of this exercise is to:

1. Introduce two tools for describing this variation: semivariance and autocorrelation; and
2. Give students experience creating and interpreting semivariograms and autocorrelograms.

In this lab, you will collect field data from quadrats arranged along a transect (or alternatively, you will use supplied data). You will then calculate and graph semivariograms and autocorrelograms using a spreadsheet, and you will use these graphs to determine how spatial patterns vary as a function of scale in your system. For this lab, you will need access to a spreadsheet program (such as Excel) and the file **vario.xlsx** provided on the book's website. If you choose the fieldwork option, you will also need two 100-m measuring tapes and one 1×1-m sampling quadrat. We've also provided code (see book's website) if you'd like to try the lab using R software.

M.W. Palmer (✉)
Department of Plant Biology, Ecology and Evolution,
Oklahoma State University, Stillwater, OK, USA
e-mail: mike.palmer@okstate.edu

D.J. McGlinn
College of Charleston, Charleston, SC, USA

© Springer-Verlag New York 2017

S.E. Gergel, M.G. Turner (eds.), *Learning Landscape Ecology*,
DOI 10.1007/978-1-4939-6374-4_5

INTRODUCTION

Nature is intrinsically variable, and the evolution and ecology of all organisms are contingent on such variation. Landscape ecology is concerned not only with the magnitude of this variation, but also with its geometry. Most patterns in nature are far more complex than the simple polygons and curves of Euclidean geometry. For example, forest edges are rarely straight lines, animal home ranges are not rectangles, and trees are not cones. Therefore, we need special methods to describe the shape of nature.

The discipline of spatial statistics has diversified and matured (see Cressie 1991; Bailey and Gatrell 1995), and it is not possible here to give a full summary of the wealth of methods available. Instead, the purpose of this exercise is to describe two different methods for characterizing variation in a variable as a function of position in the landscape. This variable could be a soil nutrient, a measure of vegetation height, an index of species composition, or anything else of interest. In spatial statistics, we term variables with known locations **regionalized variables**, and we label them z (so as not to confuse them with x and y, typically reserved for the spatial coordinates, or for independent and dependent variables, respectively).

The two methods covered in this exercise are variography, which is part of the discipline of geostatistics (see Isaaks and Srivastava 1989), and autocorrelation, which is derived from the familiar correlation coefficient (Sokal and Rohlf 1981). Recall that the **correlation coefficient**, r, is a number that varies between −1 and +1 and reveals the nature of the relationship between two variables. It is close to −1 for two variables that are strongly negatively related, close to 0 for unrelated variables, and close to +1 for positively related variables. In contrast, variography is derived from the *variance*, which must be a positive number but can otherwise take any value.

One of the most important properties of almost all regionalized variables is **spatial dependence**. Spatial dependence (as assessed by **spatial autocorrelation**, or the tendency of a random variable to be correlated with itself at finite distances) means that a variable measured at one location *depends*, in one way or another, on the same variable measured at a different location. Spatial dependence arises for a number of different reasons, but let us consider two examples.

If you examine mean annual temperature as a function of position on the globe, you will note that (with many important and interesting exceptions) there is a gradient from warm temperatures at the equator to cool temperatures at the poles. If you have two sites that are almost at the same latitude, they will have similar temperatures. On the other hand, two sites that are on different latitudes will have different temperatures, and the amount of the difference in temperature will be positively (and gradually) related to the difference in latitude. **Spatial dependence** occurs when information available at one location allows you to infer information about the other location.

Another example is in a savanna landscape where widely spaced trees provide islands of shade in an otherwise sunny landscape. Two sites that are centimeters apart are likely to have a similar amount of sunshine. However, two sites that are several meters apart may, or may not, have similar amounts of sunshine—a lot depends on the size and spacing of trees. If the sites are hundreds of meters apart, you may not be able to predict the sunlight regime very well. So in this case, we have spatial dependence at fine scales, but not necessarily at coarse scales. Also, unlike the example of global temperature, our regionalized variable consists of fairly discrete *patches* of sun and shade.

The first column of graphs in Figure 5.1 displays a variety of made-up regionalized variables with identical means and variances. These hypothetical variables have been constructed to illustrate the diversity of patterns that could potentially be found in nature. Note that regionalized variables can consist of a variety of features such as patches (i.e., homogeneous regions), noise (random, independent variation), random walks (a random walk is when a value at a given location equals the value at an adjacent location, plus or minus a small random number), or some combination of these. Also, note that the different variables behave differently as a function of scale. For example, patches can be large, small, or intermediate. Stretches of linear behavior can also be large, small, or intermediate. Also, noise can operate at any scale. If the graphs in Figure 5.1 were based on real data, we would seek biological explanations for the different scales. Such explanations might involve the size and shape of underlying geomorphology, the average size of plant clones, the average size of a natural disturbance, the home range size of the dominant mammal species, or the average farm size.

Except for variable A in Figure 5.1, there is some spatial dependence. That is, nearby locations are, *on average*, more similar than distant locations. Since similarity typically decreases as a function of distance of separation, we also call this phenomenon **distance decay**. Distance decay has important consequences for living things. For example, if soil conditions are very similar at nearby locations (as for variables C and F in Figure 5.1), then natural selection *might* favor plants with short dispersal distances. If, on the other hand, soil conditions were spatially unpredictable (as for variable A), a long dispersal distance *might* be advantageous. Similarly, the foraging behavior of animals, the growth of plant roots, the spread of fire, the flow of water, and the behavior of many other ecological phenomena all depend on the nature of distance decay in environmental factors.

In statistics, spatial dependence has both desirable and undesirable attributes (Legendre 1993). It means that one can predict variables (to some degree) based on geographic location, which can aid in mapping the environment. However, spatial dependence also violates the standard statistical assumption of independent observations (even if samples are randomly located). Thus, unless specifically corrected for, many statistical methods are invalid if your data exhibit distance decay. Fortunately, there are tools to evaluate the *degree* and the *scales* of spatial dependence. The two tools we introduce in this laboratory exercise are the semivariogram and the autocorrelogram.

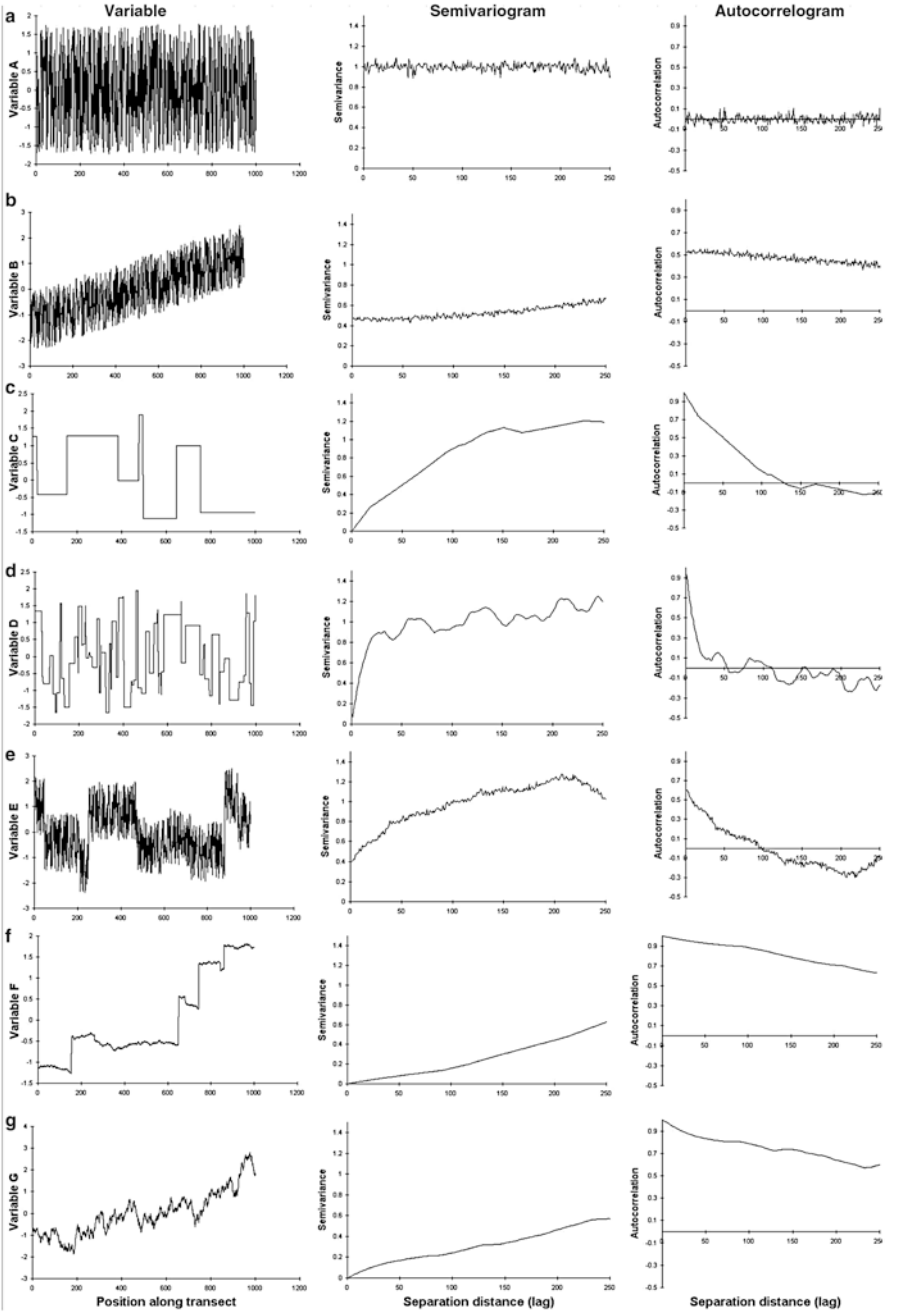

Figure 5.1 Seven artificial regionalized variables (column 1) as a function of position along a transect, along with their corresponding semivariograms (column 2) and autocorrelograms (column 3). All variables have identical means and variances. The variables can be described as follows: (**a**) pure noise; (**b**) fine-scale noise superimposed on a linear trend; (**c**) large patches; (**d**) small patches; (**e**) noise superimposed on large patches; (**f**) patches with "drift" in their mean values, plus fine-scale noise; (**g**) random walk

Variography

Variography is the discipline of using semivariograms (and related graphs such as covariograms) to uncover the degree to which the variance in a regionalized variable depends on distance (Rossi et al. 1992). The geographic distance between two samples is termed the **spatial lag**. Recall that the variance is the square of the standard deviation and is a measure of the spread or variation of data. The word **semivariance** is derived from "half of the variance," and indeed it is a measure of the variance of the regionalized variable, z. But what is special about the semivariance is that it changes as a function of distance. The semivariance is computed as follows:

$$\gamma(h) = \left\{\Sigma\left[z(i) - z(i+h)\right]^2\right\} / 2N(h)$$

where $\gamma(h)$ is the semivariance of a **lag** of distance h, $z(i)$ is the value of a regionalized variable z at location i, $z(i+h)$ is the value of z at a location separated from i by lag h, and $N(h)$ is the number of pairs of points separated by lag h. The summation is over all pairs of points separated by distance h. In plain English, the semivariance is half of the average squared difference of all pairs of points separated by a given distance. A semivariogram is a plot of semivariance versus the lag distance. As with the variance, the semivariance cannot be less than zero, but it is not bounded on the top.

An idealized, hypothetical semivariogram is given in Figure 5.2. Since the semivariance is directly related to variance, a high value indicates high variation, and a low value indicates low variation. Almost always, variance increases as a function of lag

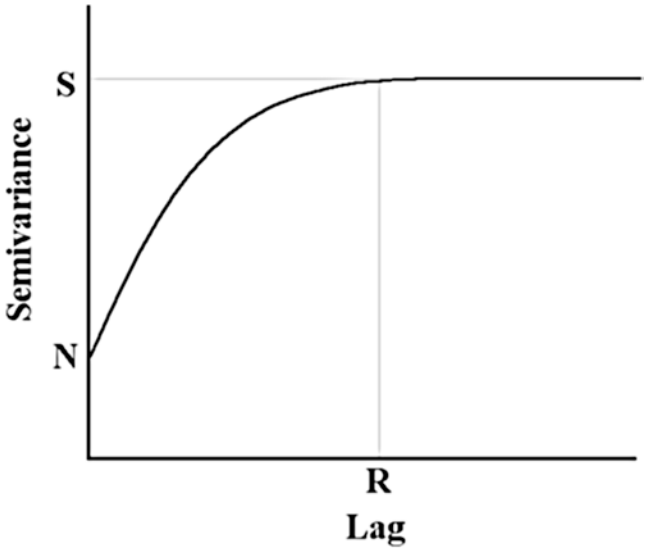

Figure 5.2 An idealized semivariogram. N = nugget, R = range, S = sill

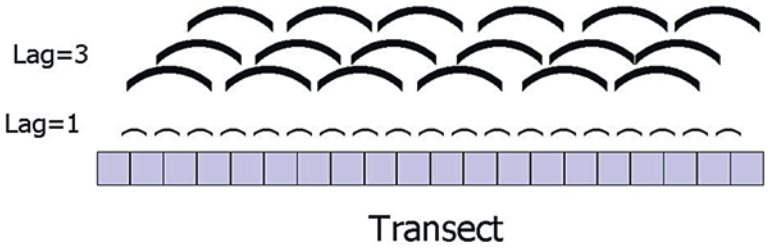

Figure 5.3 A 20-quadrat-long transect illustrating all of the pairs of points separated by two selected lags: 1 and 3

distance. In other words, the larger the area you study, the more variable your conditions are. It is important to reiterate that the lag distance is not the same as the distance from the origin or starting point. Rather, it is calculated for all pairs of points (Figure 5.3).

At distances less than R (the **range**), we have spatial dependence (Figure 5.2). That is, closer samples are more similar than distant samples. At distances of at least h, we have spatial independence; therefore, samples separated by longer distances would be valid for conventional statistics. Any area with linear dimensions of at least R would have as much variance as the landscape as a whole. A horizontal asymptote at distances greater than R is known as the **sill**. The sill indicates the amount of "background" variation.

For a very smooth, regionalized variable, two samples that are infinitesimally close to each other may have almost identical values. Elevation of the ground surface almost always behaves this way. However, most variables are not so smooth, if for no other reason than measurement error. Such unresolved variation at very fine scales is termed the **nugget effect**, and is indicated by N (Figure 5.2). The term derives from the original use of variography in gold mining: at fine spatial scales, you either find the gold nugget or miss it. Soil variables such as pH or nutrient concentrations typically have very high nugget effects.

The semivariogram specifically address how variance increases as a function of scale. Although it can only describe *patterns*, we often hope to infer the *processes* that generate such patterns. If we find a distinct range, or even a pronounced inflection in the semivariogram, we suspect that there are different processes operating at different scales. For example, if we discover that the range equals approximately 10 m, we need to seek an underlying process that operates on a scale of 10 m. In a forest, this scale could represent the average canopy gap size or the average size of a canopy tree crown. In the arctic tundra, the range could represent the average size of a permafrost polygon. Of course, you are never guaranteed to actually find a range. It is possible (and indeed, likely) that variation in nature increases continuously as a function of scale.

The second column of Figure 5.1 shows the semivariograms for the hypothetical regionalized variables. Since very few pairs of points represent very far distances, it is usually not advisable to plot $\gamma(h)$ for large h. A general rule of thumb (adopted here) is to plot only up to half of the maximum distance between samples. The

bumps and wiggles at such far distances in the semivariograms are due to chance variation in the data, and not to the underlying process generating the patterns. Note that only three of the variables (C, D, and E) have semivariograms remotely resembling those in Figure 5.2. The range of variable C (≈150 units) is much larger than the range of variable D (≈25 units), which reflects the differences in patch size. Variable E seems to have an inflection at 50 m, which marks the difference between noise within patches, and the differences between patches.

The semivariograms of variables B, F, and G are continuously increasing functions. Therefore, there is spatial dependence at broad spatial scales. Three variables (A, B, and E) have much fine-scale noise, and hence a substantial nugget effect. Note that variable F has very little fine-scale noise, and hence has a negligible nugget. Variable A represents pure noise and hence pure spatial independence. For such variables, the nugget equals the sill.

Autocorrelation

Autocorrelograms are plots of the correlation coefficient, *r*, as a function of lag:

$$r(h) = \mathrm{corr}\left[z(i), z(i+h)\right]$$

It is called "auto"-correlation because the variable is correlated with itself. Autocorrelation can take values from −1 to +1 although for most applications positive values are most common. In situations with distance decay, autocorrelograms are declining functions and often look like upside-down semivariograms (third column of Figure 5.1). If there is little fine-scale noise, the *y*-intercept will be close to 1.0. In situations in which the semivariogram displays a nugget effect, the *y*-intercept of the autocorrelogram will be less than 1. An autocorrelation of 0 means there is no spatial predictability; this is related to the concept of the sill.

This describes only the simplest kind of autocorrelogram. More complex (and usually more appropriate) ways to calculate autocorrelograms, as well as testing their statistical significance, are described by Legendre and Fortin (1989), Bailey and Gatrell (1995), and Legendre and Legendre (1998). Autocorrelograms are also used in the analysis of change through time (also known as "time series").

Comparing Autocorrelation to Semivariance

The interpretation of autocorrelograms is very similar to that of semivariograms, so the choice between them is largely a matter of taste. Since the correlation coefficient is a dimensionless number (i.e., it is standardized), autocorrelograms are useful in comparing variables with different units (e.g., plant density and soil calcium). Semivariance has a dimension of units squared (so if the regionalized variable is in

parts per million or ppm, semivariance is in ppm^2). Thus, it is useful in comparing different commensurate variables or (more commonly) the same variable in different locations. However, semivariance can be standardized for comparing variables measured in different units (see Rossi et al. 1992).

Since it is derived from the correlation coefficient, autocorrelation is closely related to classical statistical theory. Variography, on the other hand, is a branch of geostatistics. This discipline was largely developed for the mining industry to help predict the locations of mineral deposits. Variography is a precursor to geostatistical interpolation (for mapping) or "kriging" (see Isaaks and Srivastava 1989) and to fractal geometry (Burrough 1983; Palmer 1988).

EXERCISES

EXERCISE 1: Data Collection

Option 1: Field Exercise Using Vegetation Height

1. Choose a field site in which the maximum height of the vegetation is about 2.5 m or less, and in which it is possible to fit a 200-m transect. If the site is large enough, randomly choose a starting location and compass direction. If it is too small for this, choose an appropriate direction but randomize the starting point, so you are not biased by particular plants.
2. Extend two (or more) 100-m tapes end to end along the chosen compass direction. Ideally, you would do this with a surveyor's compass or level. It is crucial that the transect be as straight as possible and not influenced by the vegetation!
3. Beginning at 0 m, establish a 1×1-m quadrat (this can consist of three meter sticks plus the meter tape as the fourth boundary).
4. Within this plot, measure and record the height of the tallest plant.
5. Now repeat the process with an adjacent plot at 1 m, then at 2 m, and continue to the end of the transect.

If you do not have the luxury of a large enough field site, it is possible to perform this exercise with smaller contiguous quadrats. Also, the regionalized variable need not be height: you can perform this same exercise using stem density, biomass, species richness, ordination scores, percent cover of bare soil, elevation, percent sunlight, soil parameters, and more. It may be possible to derive a regionalized variable from a map or a remotely sensed image, but be aware that data on such images *may* have already been "smoothed" or interpreted for ease of display, and hence your analyses would be inappropriate. Regardless of the overall length and quadrat size, try to have at least 200 quadrats in your transect (spatial analyses typically require large sample sizes). Another option is to split the class up into two or more groups, with each group studying either a different regionalized variable along the same transect, or the same variable in a different vegetation type.

Option 2: Using Provided Data Sets

Some example data sets are provided on the book's website, in case there are no opportunities for collecting new data. The file is entitled **vario.xlsx** and contains a worksheet with three different **example data sets**.

EXERCISE 2: Data Analysis

You will analyze the data using a spreadsheet. The example given here is for Microsoft Excel, but similar commands exist in other spreadsheets. Before beginning this exercise, review absolute and relative cell references, how to graph data, as well as the following Excel functions: **OFFSET**, **SUMXMY2**, **CORREL**. Make sure that automatic calculation of formulas is in effect (this is the most usual default; it means that the results will be continuously updated. Check under **File - Options - Calculation - Automatic**).

1. Enter your data in the blank worksheet labeled **Vegetation Heights**. The following description assumes that the transect is 200 quadrats long; if not, substitute "200" with the correct number.
2. In row 1, label columns **A**–**F** as follows:

A	B	C	D	E	F
POSITION	VALUE	LAG	SEMIVARIANCE	LAG	AUTOCORRELATION

3. In column **A**, fill rows **2**–**201** with the numbers 1–200.
   (*HINT:* One quick way to do this is to put "`1`" in cell **A2**, and then put the formula: "`=A2+1`" in cell **A3**. Then copy the contents of **A3** and paste them into cells **A4**–**A201**. Since there were no dollar signs ($) in the original formula, the cell reference of **A2** is copied as a relative location. Therefore, each one of the cells will equal the cell above it plus 1).
4. In column **B**, fill rows **2**–**201** with the data you collected (or copied from the provided data sets) in the correct spatial sequence.
5. In column **C**, fill rows **2**–**101** with the numbers 1–100. (Recall the rule of thumb that it is best not to plot semivariograms for more than half of the maximum lag distance). Repeat this for column **E**.

Two different ways to calculate the semivariance follow. Method 1 is conceptually easier, but method 2 is less labor intensive. Therefore, read and understand method 1, but use method 2. It is important to keep in mind that in a transect Q units long, the number of pairs of quadrats separated by a given lag distance h will equal $Q - h$. In our example, there are 199 pairs of quadrats separated by 1 m, 198 separated by 2 m, and 1 pair separated by 199 m (i.e., the first quadrat and the last quadrat). Before continuing, review the equation for semivariance.

Semivariance Method 1: Read and understand this method

1. In cell **D2**, put the formula:
   "=SUMXMY2(B2:B200,B3:B201)/(2*(200−1))"
   The formula SUMXMY2 means "sum of (*x* minus *y*) squared". The two selected blocks (**B2:B200**) and (**B3:B201**) are actually the same data, but shifted by a lag of 1 unit. The denominator is two times the number of pairs of points separated by distance *h*. It is, of course, possible to put the number 198 in the denominator, but writing the formula out often helps with troubleshooting.
2. In cell **D3**, write the formula:
   "=SUMXMY2(B2:B199,B4:B201)/(2*(200−2))"
   This is the semivariance for a lag of 2. The formula for **D4** should be:
   "=SUMXMY2(B2:B198,B5:B201)/(2*(200−3))"
3. Continue filling in column **D** until you reach a lag of 100.

Semivariance Method 2: Use this method

1. Instead of typing in a unique formula for each cell of column **D**, it is more time-efficient to type in a generic formula in cell **D2:**
   "=SUMXMY2(B$2:OFFSET(B$2,200−C2−1,0),OFFSET(B$2,C2,0):B$201)/(2*(200−C2))"

NOTE: This is precisely the same formula as in method 1, except for how we specify addresses. The dollar signs ($) before the row means a reference to that exact row, no matter where you copy and paste the formula. **OFFSET** returns a new cell address and has three arguments: a cell address, the number of rows of separation, and the number of columns of separation. Therefore, the block B$2:OFFSET(B$2,200−C2−1,0) refers to a column of data beginning at cell B2 and ending (199−C2) cells below B2. Since cell C2 indicates a lag of 1, the column of data will be the same as B2:B200, as desired. The second block, OFFSET(B$2,C2,0):B$201, means a block beginning at C2 below B2, and ending at B201. This will be the same as B3:B201. The denominator of the equation will equal 2*(200−1).

2. Copy cell **D2** and paste it into cells **D3–D101.** Note that when you do so, the formula remains identical in all cells *except* that the reference to the lag, column **C,** changes. Thus, the formula will return the semivariance for whatever lag is indicated in the same row of column **C**.

You can calculate autocorrelation by similar methods to those described earlier for semivariance.

Autocorrelation Method 1: Read and understand this method

1. In cell **F2,** type: "=CORREL(B2:B200,B3:B201)". This will return the correlation between the variable and itself, with a lag of 1.
2. In cell **F3**, type "=CORREL(B2:B199,B4:B201)" for a lag of 2, and continue filling column **F** until lag 100.

Autocorrelation Method 2: Use this method

1. Following the same reasoning as method 2 for the semivariance, type the following in cell **F2:**
   "=CORREL(B$2:OFFSET(B$2,200-E2-1,0),OFFSET(B$2,E2,0):B$201)"
2. Copy this formula and paste it into cells **F3–F101.**

Before proceeding further, make sure to save your results.

EXERCISE 3: Results

Using your spreadsheet program, create the following plots:

1. Vegetation height as a function of transect position
2. Create a semivariogram as follows:
   a. Plot the semivariance as a function of lag. The data will be in the block **C2:D101**.
   b. Label the *X*-axis "Lag (meters)", and the *Y*-axis "Semivariance".
   c. Drag the graph immediately under the graph of the raw data.

HINT: First, make an X,Y (scatter) plot under **Insert - Charts - Scatter**. Then, double-click on any point in the graph and select **Format Data Series**. In Excel 2013 onwards, select the icon resembling a paint can (aka Fill & Line) and choose **Solid Line**. In older versions of Excel, select **Patterns - Line - Automatic**.

3. Create an autocorrelogram as follows:
   a. Graph the data in **E2:F101** and drag the graph under the semivariogram.
   b. Label the axes appropriately.

Interpretations and Rules of Thumb

As with any bivariate (two-variable) graph, the scaling of the *Y*-axis relative to the *X*-axis should not affect our interpretation, but it often does. A short, long graph often appears less "noisy" than a tall, narrow one. It is generally best to choose a scaling relatively close to 1:1 (that is, square), or at most 1.5:1 or 1:1.5. Of course, there may be exceptions (e.g., if one wants to display the results of numerous transects, one graph on top of the other, it might be useful to have them short and long).

For semivariograms, it is conventional and advisable for both the *x*-minimum and the *y*-minimum to be zero. The *x*-minimum should be zero for the autocorrelogram, but a case can be made that the *y*-minimum and the *y*-maximum should be −1.0 and +1.0, respectively. If part of the goal of the research is to compare the results from different transects, *x*-axes and *y*-axes of the same kind of plot should be scaled identically.

The plot of vegetation height as a function of transect position will typically have some sort of broad-scale pattern, immediately detectable upon inspection, in addition to fine-scale variation. The details will vary markedly depending on the nature of your plant community. The fine-scale variation may be partially measurement error, but in

most cases it is predominantly caused by natural variation. Increasing the number of samples (i.e., transect length) will not reduce the magnitude of this fine-scale variation.

The semivariograms and autocorrelograms will also have an overall shape, summarizing the spatial patterns of the community, as well as fine-scale variation. However, in contrast to the graph of height, increasing the sample size (transect length) will tend to decrease the finer-scale patterns. This means that we are increasingly confident that we have described how spatial pattern (variance or correlation) is related to scale (spatial lag).

Q1 How does height behave as a function of distance along the transect? Is this generally consistent with your impression of the field site?

Q2 Examine the semivariogram. Is there an identifiable nugget? Range? Sill?

Q3 Does the regionalized variable (height) exhibit spatial dependence?

Q4 Examine the autocorrelogram. Is there spatial autocorrelation?

Q5 How would you describe the nature of your spatial variation? Does your pattern consist of patches? Noise? A dominant trend? Nested patterns of variation? Random walk? A random walk (also known as "drift") is when there is spatial dependence, but the difference between each number and the previous number is random. The term *random walk* derives from a plot of distance from the starting point as a function of time for an animal whose direction of movement is purely random.

Q6 Is there periodicity in your data (i.e., did the response change regularly at several spatial intervals)? How would you know this from the shape of the semivariogram or autocorrelogram? Note that both the semivariogram and the autocorrelogram can describe variance as a function of scale, but neither can completely summarize the *nature* of spatial variation (e.g., patches, gradients, or a combination). This is akin to the observation that variance does not fully describe the statistical distribution of data (e.g., whether it is normally distributed), and that the correlation coefficient does not fully describe the nature of the relationship between two variables (e.g., whether they might have a nonlinear relationship).

Q7 Suppose a rodent species requires tall vegetation for cover. Does the nature of the spatial pattern you observe have implications for this species?

Q8 Suppose a predator only hunts in relatively short vegetation. Does the nature of spatial variation have implications for foraging behavior?

Q9 Can you think of any other biological ramifications of your results?

Q10 If you collected data from more than one site or variable, how do their spatial patterns compare?

SYNTHESIS

Q11 How does your variable behave in comparison to the supplied data sets? The supplied data sets (on the page **example data sets** in the spreadsheet accompanying this laboratory) can be pasted into the data column (column **B**) in the worksheet **ready-to-go blank**, and the semivariograms and autocorrelograms will be recalculated automatically (however, note that you may need to change the *Y*-axis scaling on the graphs).

Q12 Refer to Figure 5.1. Choose two or three of the variables and describe what natural phenomena might lead to those patterns.

Q13 If you find spatial dependence, what does this imply for the use of conventional statistics?

Q14 Are some spatial scales better than others for studying your system? Why or why not?

Q15 In theory, regionalized variables are measured at points. However, you have measured them in a quadrat. What do you expect would happen to the semivariogram if you reduced the size of the quadrat?

Q16 What is noise? Is it a useful concept?

OPTIONAL EXERCISES

EXERCISE 4: Correlation and Variation

As their names imply, the auto**correl**ograms and semi**vari**ograms stress correlation and variance, respectively. Therefore, they are likely to behave differently in data sets with different variance. In the supplied spreadsheet accompanying this exercise, locate a worksheet entitled **2 hypothetical variables**. Examine the two variables carefully.

Q17 How do these variables differ?

Q18 How do you expect their semivariograms and autocorrelograms to differ?

Q19 Now copy one of the variables and paste it in the data column (column **B**) in the sheet labeled **ready-to-go blank**. Examine both the semivariogram and autocorrelogram. Now repeat with the second variable. Were you right in your answer to the previous question?

EXERCISE 5: Variography and Fractals

Plot your semivariogram on a double logarithmic scale. Do this by left-clicking on the *X*-axis of your semivariogram. Then right-click and choose **Format axis**. Select **Scale** and click **Logarithmic**. Repeat the same procedure for the *Y*-axis.

Q20 Is the semivariogram a straight line? If so, we can say that the variable is *statistically self-similar*. This means that fine-scale patterns are indistinguishable from scaled-down versions of broader-scale patterns. The concept of "self-similarity" is intrinsic to the study of *fractal geometry*. The fractal dimension D can be determined from the slope m of the log-log semivariogram with the formula $D=(4-m)/2$. The interpretation of the fractal dimension is beyond the purpose of this chapter; see Burrough (1983) and Palmer (1988) for more details.

Q21 Are there multiple plateaus? If so, we have a hierarchy of spatial patterns. This would imply that we have distinctly different processes operating at distinctly different *scale domains*.

Q22 Would you predict that most spatial patterns in nature are self-similar, hierarchical, or neither?

EXERCISE 6: Variography using R

R code for semivariograms and autocorrelograms is provided on the website for this book. Repeat the same analyses as presented in the main lab, but using the supplied code instead.

FURTHER STUDY

This exercise only considered one-dimensional patterns. However, ecologists typically study spatial patterns in two dimensions. The same formulas for semivariance and autocorrelation hold, but the calculations are a bit more complicated because distances no longer fall in discrete lag intervals. Therefore, we typically average semivariance over a certain range of lags. A further complication arises if the patterns are not **isotropic** (statistically the same in all compass directions). In such cases, we usually calculate different semivariograms and autocorrelograms for different directions.

Furthermore, sampling need not be in a perfectly sampled transect as in this lab. It is perfectly legitimate for samples to have locations that are random, on interrupted transects and grids, or any other objective method. When samples are located

at irregular intervals (and/or in two dimensions), many of the spatial lags are not a simple multiple of the minimum spacing. We deal with this by creating "lag classes" (e.g., 0–1 m, 1–2 m, 2–3 m) much in the same way as we would generate a histogram. Although there are no firm rules about how many pairs of points should fall within a lag class for an accurate semivariogram or autocorrelogram, a general rule of thumb is that it should be at least 80.

In this lab, we interpreted semivariograms and attempted to find the range, sill, and nugget by eye-balling. However, it is common to use a curve-fitting procedure such as nonlinear regression to actually obtain estimates of these parameters (see Legendre and Legendre 1998), as in Chapter 11. Such curve-fitting is an essential step for procedures such as kriging, discussed next.

Variography is often a precursor to a geostatistical interpolation procedure known as **kriging** (Hohn 1988; Isaaks and Srivastava 1989; Cressie 1991). By interpolation, we mean that we estimate the value of a regionalized variable at an unsampled location, based on knowledge from sampled locations. The most common product of kriging is a map (usually a contour map) of the variable of interest. Kriging performs best when the nugget is small relative to the sill and when the average distance between nearby samples is less than the range. See Legendre and Fortin (1989), Halvorson et al. (1994), Marinussen and Van Der Zee (1996), and Carroll and Pearson (1998) for examples of kriging in ecology.

One may have noticed a resemblance between the semivariogram and the well-known species–area relationship (Scheiner 2004). This resemblance is more than casual. As Wagner (2003) elegantly illustrates, distance decay in species distributions scales up to distance decay in species richness, one of the root causes of the species–area relationship (Palmer and White 1994). Indeed, Wagner (2003) develops numerical techniques by which one can separate how much of the semivariogram for species richness is due to intraspecific autocorrelation, and how much is due strictly to interspecific co-occurrence. Such techniques elevate variography beyond mere description of pattern, and into the realm of uncovering fundamental properties of biodiversity such as those explored in Chapter 15. While we have only discussed univariate patterns in this lab, bivariate or multivariat patterns are often of interest. If so, we can use covariograms or cross-correlograms to determine whether the relationships between variables change as a function of spatial scale. Additional multivariate approaches are also explored further in Chapter 15.

Lastly, for a simple analysis of spatial pattern along a transect (as in this lab), it is possible to perform basic calculations on a spreadsheet. However, a spreadsheet becomes cumbersome for more complex sampling designs such as interrupted or two-dimensional sampling and for complex analyses such as detection of anisotropy, significance testing, nonlinear curve-fitting, multivariate patterns, or kriging. Fortunately, a wide range of software exists for such analyses, including packages in R and GIS software (Chapters 11 and 15).

Acknowledgements We thank Steven Thompson, Sam Fuhlendorf, Anne Cross, Sophonia Roe, Marie-Josée Fortin, and four anonymous reviewers for comments on earlier versions. We especially thank Sam Fuhlendorf for allowing the use of his data, and José Ramón Arévalo for collecting new data for the lab.

REFERENCES AND RECOMMENDED READINGS[1]

Bailey TC, Gatrell AC (1995) *Interactive spatial data analysis*. Longman Scientific & Technical, Essex *An excellent text on spatial statistics; includes some demonstration software*

Burrough PA (1983) Multiscale sources of spatial variation in soil. Application of fractal concepts to nested levels of soil variations. J Soil Sci 34:577–597

*Carroll SS, Pearson DL (1998) Spatial modeling of butterfly species richness using tiger beetles (*Cicindelidae*) as a bioindicator taxon. Ecol Appl 8:531–543. *A good example of the application of spatial statistics in biodiversity studies.*

*Cohen WB, Spies TA, Bradshaw GA (1990) Semivariograms of digital imagery for analysis of conifer canopy structure. Remote Sens Environ 34:167–178. *A good example of the use of variography in digital imagery.*

*Cressie NAC (1991) *Statistics for spatial data*. Wiley Interscience, New York. *Comprehensive and authoritative work on spatial statistics.*

*Fortin M-J, Dale MRT (2014) *Spatial analysis: a guide for ecologists.* Cambridge University Press, Cambridge. *Excellent survey of different spatial analysis tools for ecologists including variograms and autocorrelation with many helpful visual aids.*

*Halvorson JJ, Bolton Jr H, Smith JL et al (1994) Geostatistical analysis of resource islands under Artemisia tridentata in the shrub-steppe. Great Basin Naturalist 54:313–328. *An example of ecological applications of geostatistics on a fine scale.*

*Hohn ME (1988) Geostatistics and petroleum geology. Van Nostrand Reinhold, New York. *A good, clear text for the use of geostatistics for interpolation and mapping.*

*Houle G (1998) Seed dispersal and seedling recruitment of Betula alleghaniensis: spatial inconsistency in time. Ecology 79:807–818. *Another good example of geostatistics in fine-scale plant ecology.*

*Isaaks EH, Srivastava RM (1989) An introduction to applied geostatistics. Oxford University Press, New York. *A popular text on the use of geostatistics for interpolation.*

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? Ecology 74:1659–1673

*Legendre P, Fortin M-J (1989) Spatial pattern and ecological analysis. Vegetatio 80:107–138. *A good, brief introduction to spatial methods available to ecologists.*

Legendre P, Legendre L (1998) Numerical ecology, 2nd English edn. Elsevier, Amsterdam

*Marinussen MPJC, Van Der Zee SEATM (1996) Conceptual approach to estimating the effect of home-range size on the exposure of organisms to spatially variable soil contamination. Ecol Model 87:83–89. *A good applied example of the use of spatial statistics.*

*Palmer MW (1988) Fractal geometry: a tool for describing spatial patterns of plant communities. Vegetatio 75:91–102. *Follows Burrough (1983) to describe how variography can be used as an introduction to the world of fractals.*

Palmer MW, White PS (1994) Scale dependence and the species-area relationship. Am Nat 144: 717–740

*Rossi RE, Mulla DJ, Journel AJ et al (1992) Geostatistical tools for modeling and interpreting ecological spatial dependence. Ecol Monogr 62:277–314. *A good general introduction to variography, geared toward the ecologist.*

[1] *NOTE:* An asterisk preceding the entry indicates that it is a suggested reading.

Scheiner SM (2004) A mélange of curves—further dialogue about species-area relationships. Global Ecol Biogeogr 13:479–484

Schlesinger WH, Raikes JA, Hartley AE et al (1996) On the spatial pattern of soil nutrients in desert ecosystems. Ecology 77:364–374 *A good geostatistical study of soil nutrients*

Sokal RR, Rohlf FJ (1981) *Biometry*. Freeman, New York

*Wagner HH (2003) Spatial covariance in plant communities: integrating ordination, geostatistics, and variance testing. Ecology 84:1045–1057. *Seminal paper uniting variography with biodiversity statistics.*