

LAB EXERCISE #1 – Scaling Techniques

Instructor: K. McGarigal

Overview: In this exercise, you will gain familiarity with a few common procedures for elucidating the scale of pattern in point, continuous and categorical spatial patterns. Specifically, you will: 1) calculate the Clark and Evans Nearest Neighbor Index and Ripley's K distribution for the pattern of vernal pools in a portion of western Massachusetts and interpret the results, 2) derive a Correlogram and Semivariogram for a snag transect in western Oregon and interpret the results, and 3) derive a lacunarity plot for several categorical map patterns and interpret the results.

Objectives

- To learn some basic techniques for scaling patterns.
- To gain a practical understanding of how to compute the Clark and Evans (1954) Nearest Neighbor Index and Ripley's K distribution for spatial point patterns.
- To gain a practical understanding of how to conduct a semivariance and autocorrelation analysis for continuous data.
- To gain a practical understanding of how to conduct a Lacunarity analysis of categorical map patterns.
- To gain an appreciation for some of the limitations in scaling spatial patterns.

Background Information

Scaling Point Patterns - Ripley's K :

First-order analysis of point patterns allows one to assess the overall dispersion of points; that is, whether the points are distributed randomly or in a more uniform or clumped manner. While first-order pattern analysis does allow one to assess dispersion over a range of scales, and is thus not a scaling technique per se, it is a useful first step to evaluating point patterns. A host of first-order techniques have been developed based on nearest-neighbor distances. A typical approach is to use the mean point-to-point distance to derive a mean area per point, and then to invert this to get a mean point density (points per unit area), λ , from which the test statistics about expected point density are derived. A classic index is by Clark and Evans (1954):

$$ID_2 = \frac{\bar{x}_d}{\left(\frac{1}{2\sqrt{\lambda}} \right)}$$

where \bar{x}_d is the mean nearest-neighbor distance over all points $i=1, 2, \dots, n$, and the denominator is the expected mean nearest-neighbor distance under the assumption of a Poisson process.

Second-order analysis of point patterns allows one to assess point dispersion over range of scales. Perhaps the most well-known is the Ripley's K -distribution or K -function (Ripley 1976). The K -distribution is the cumulative frequency distribution of observations at a given point-to-point distance (or within a distance class). Because it preserves distances at multiple scales, Ripley's K can quantify the intensity of pattern at multiple scales.

Consider a *spatially random* distribution of N points. If circles of radius d_s , are drawn around each point, where s is the order of radii from the smallest to the largest, and the number of other points that are found within the circle are counted, and then summed over all points (allowing for duplication), then the expected number points within that radius $E(d_s)$ from an arbitrary point are:

$$E(d_s) = \frac{N}{A} K(d_s)$$

where N is the sample size, A is the total study area, and $K(d_s)$ is the area of circle defined by radius d_s . For example, if the area defined by a particular radius is one-fourth the total study area and if there is a spatially random distribution, on average approximately one-fourth of the cases will fall within any one circle (plus or minus a sampling error). More formally, with *complete spatial randomness* (csr), the expected number of points within distance, d_s , is:

$$E(d_s [csr]) = \frac{N}{A} \pi \cdot d_s^2$$

On the other hand, if the average number of points found within a circle for a particular radius placed over each point, in turn, is greater than that found by the equation above, this points to clustering for that radius. Conversely, if the average number of points found within a circle for a particular radius placed over each point, in turn, is less than that found by the equation above, this points to dispersion; that is, points are, on average, farther apart than would be expected on the basis of chance for that radius. By counting the number of total number of events within a particular radius and comparing it to the number expected on the basis of complete spatial randomness, the K -statistic is an indicator of non-randomness. In this sense, the K -statistic is similar to the first-order nearest neighbor indices, such as the Clark and Evans index; however, it is more comprehensive than these first-order metrics for two reasons. First, it applies to all orders cumulatively, not just a single order. Second, it applies to all distances up to the limit of the study area because the count is conducted over successively increasing radii. Indeed, this is the great utility of Ripley's K for investigating point intensity over a range of spatial scales.

Given the above, Ripley's K is defined as:

$$K(d_s) = \frac{E(d_s)}{\lambda}$$

where $E(d_s)$ is the expected number of points within a distance d_s from an arbitrary point. Again, the mean intensity λ is estimated simply as N/A , where N is total number of points and A is total area sampled. The cumulative distribution $E(d)$ is estimated empirically, and so in practice this requires completely surveyed data (not sparse samples of points). The typical estimate of $K(d)$ is tallied as:

$$K(d_s) = \frac{1}{\lambda} \sum_{i=1}^n \sum_{j=1}^n \left[\frac{\delta_{ij}(\|x_i - x_j\| \leq d_s)}{N} \right]$$

for i not equal to j , where δ is an indicator function that takes on a value of 1 if the condition is true, else 0. Here, the condition is that the Euclidean distance between points is less than or equal to d . That is, $K(d)$ is a tally of the average number of points that fall within each distance class from an arbitrary point. The inverse of λ in the formula converts K to units of area. Other variants of the equation can be used to correct for edge effects, in which cases δ takes on values less than 1.0 in the summation.

Edge effects can seriously degrade distance-based statistics, and there are at least two ways to deal with these. One way is to invoke a buffer area around the study area, and to analyze only a smaller area nested within the buffer. By common convention, the analysis is restricted to distances of half the smallest dimension of the study area. This, of course, is expensive in terms of the data not used in the analysis. A second approach is to apply an edge correction to the indicator function for those points that fall near the edges of the study area; Ripley and others have suggested a variety of geometric corrections.

As noted above, for completely random data, the expected value of $K(d_s)$ is $\pi*d_s^2$. For clarity of presentation, the K distribution is often transformed in one of two ways. The transformed distribution is defined:

$$L(d_s) = \sqrt{\frac{K(d_s)}{\pi}}$$

which produces a plot of $L(d)$ against d where the expectation under randomness is a line with slope 1.0 (i.e., $L(d)=d$). Subtracting d from $L(d)$ transforms this expectation to a horizontal line under randomness. Which transformation to use is purely cosmetic.

Scaling Continuous Data - Spatial Autocorrelation and Semi-variance:

Two methods, autocorrelation and semivariance analysis, are popular tools to discover pattern in geostatistical data (Legendre and Fortin 1989).

(1) Autocorrelation

First, let's consider a way to index similarity in the variable as measured for two samples. A simple correlation coefficient will suffice:

$$r_{ij} = \frac{n \sum_{i=1}^n \sum_{j=1}^n (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

where the values y for two samples denoted i and j are deviates $y_i - y_{mean}$ and there are a total of n of these. Now we need a way to qualify this index so that it is scale-specific, that is, so it reflects the correlation between points separated by some particular distance. An easy way to do this is to introduce an indicator or switching function (a weight) such that the weight takes on a value of 1 only if the two points are separated by a given distance d ; otherwise the weight is set to 0. The new index, computed as Moran's I , is:

$$I(d) = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{W \sum_{i=1}^n (y_i - \bar{y})^2}$$

where w_{ij} is the indicator (weight) for samples i and j , and W is the sum of all the weights. Like a familiar Pearson's correlation coefficient, $I_{(d)}$ is bounded on $[-1,1]$, with 1 being perfect positive autocorrelation. The expected value under randomness is a value near 0, signifying a lack of autocorrelation.

A **correlogram** is a graph of autocorrelation versus distance. For a repeating pattern of clumps separated by some distance, the clump size appears as a peak in autocorrelation for distances up to average clump size; the distance between clumps appears as a second peak. For very regular patterns such as a sine wave, the size of the pattern is indicated by the distance between peak and trough (maximum dissimilarity of measurements).

As a scaling technique, we are interested in distances at which a variable is strongly autocorrelated. With ecological data, this is usually positive autocorrelation (negative cases are not common). In the case of topography, we'd expect positive autocorrelation within the scale of slope facets, with $I_{(d)}$ approaching 0 at scales beyond the grain of topography.

(2) Semivariance

An alternative way to look at spatial structure in geostatistical data is to consider the dissimilarity in values measured at two points some distance apart. Following the notation

used for autocorrelation, where y_i is an attribute value measured at location i and n is the total number of points, semivariance g (gamma) at distance d is defined:

$$g_{(d)} = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - y_j)^2}{2n_d}$$

where j is a point at distance d from i , and n_d is the number of points at that distance (or in that distance class; $n_d = N-d$). The summation term is a distance measure or variance term; by convention this is divided by 2 to rescale it. Note that if two points close together are similar, the difference term in the numerator is small, and they have low semivariance at that distance. At some distance farther removed, the association between points is no different from that of a random pair of points sampled from the grid (as rescaled, semivariance asymptotes to simple variance as estimated over all points).

A **semivariogram** is a plot of semivariance against distance (or distance class). A semivariogram is defined by its sill (asymptotic value), its range (the distance at which the sill is reached), and its nugget semivariance (the intercept). Semivariance may be divided by the simple variance to rescale it so that it should approach a value of 1 beyond the range of the variogram.

As a scaling technique, the range is of primary interest as it indicates the characteristic scale of the variable. The nugget is often of interpretative value, in that a large (nonzero) nugget might indicate spatial structure at scales smaller than the minimum resolution of the sample points (i.e., at scales finer than the minimum between-sample distance); this would indicate that some finer-scale pattern was missed in sampling.

Note that even if spatial scaling is not the primary goal of the analysis, autocorrelation and semivariance analysis can be an important part of screening geostatistical data. Samples that are autocorrelated are not independent; indeed one definition of autocorrelation is that it implies that the value of a variable at one location can be predicted from values measured at other (nearby) locations. This lack of independence violates a fundamental assumption in parametric statistical tests, and so autocorrelation analysis should be conducted prior to parametric statistical analysis.

Scaling Categorical Map Patterns - Lacunarity Analysis:

Lacunarity is a measure of scale-dependent contagion, or aggregation of patch types, suitable for categorical maps (Plotnick et al. 1993). Lacunarity essentially assesses contagion across a range of spatial scales. Briefly, a sliding window is moved throughout the grid, a row and column at a time, so that each cell is sampled several times at each window size. Consider a binary ("on/off") raster map of dimension m . The analysis is concerned with the frequency with which one encounters "on" cells in a window of size r . To do this, a gliding box is constructed of dimension

r , and a frequency distribution of tallies of $S=1,2, \dots, r^2$ "on" cells is constructed. In this gliding, the box is moved over one cell at a time, so that the boxes overlap during the sliding. Define $n(S,r)$ as the number of boxes that tally S "on" cells for a box of size r . There are $N(r)=(m-r+1)^2$ boxes of size r , and the probability associated with $n(S,r)$ is $Q(S,r)=n(S,r)/N(r)$. The first and second moments of Q are used to define lacunarity L (lambda) for this box size as a variance to mean-square ratio. This is repeated for a range of box sizes r . Lacunarity is summarized as a log-log plot of $L(r)$ versus r . Note that as r increases, the tally $n(S,r)$ converges on the mean and its variance converges on 0. Thus, for very large boxes $L=1$ and $\ln(L)=0$. At the smallest box size ($r=1$), $L(1) = 1/p$, where p is the proportion of the map occupied ("on"). Thus, for the finest resolution of a map, L depends solely on p . At intermediate scales, lacunarity expresses the contagion of the map, or its tendency to clump into discrete patches. So the lacunarity plot summarizes the texture of the map across all scales.

Detailed Instructions

There is no written assignment (i.e., report) to turn in for this lab. Instead, you will complete this project in class. The assignment is divided into three parts corresponding to the three classes of spatial patterns described above. You may work on this assignment individually or in groups.

Part 1: Point Pattern Analysis - Nearest Neighbor Index and Ripley's K

This data set consists of the complete set of 225 potential vernal pools (points) in the towns of South Hadley and Granby, Massachusetts (Appendix A; total project area = 120,592,932 m²).

Open up in Excel the spreadsheet "...scale\scalelab.xls"

1.1. Clark and Evans Dispersion Index: First-order nearest neighbor analysis

First, review the worksheet titled "points" which contains the x-y coordinates of the 225 vernal pools.

Second, review the worksheet titled "matrix" which contains the point-to-point distance matrix. This is a square symmetrical distance matrix containing the Euclidean distance (in meters) between each pair of vernal pools.

Third, review the worksheet titled "nearest neighbor" which contains the calculations for the Clark and Evans Dispersion Index. Specifically, review the calculations in the table titled "first order nearest neighbor analysis". The row labeled "nearest neighbor index" is the Clark and Evans Dispersion Index. The standard error of the index and a test statistic (Z) are given as well. The significance of the test statistic can be determined by comparing the observed value to a z statistic table for $N-1$ degrees of freedom, where N is the number of points.

Lastly, for comparison, the results of a K^{th} order nearest neighbor analysis are provided as well in the table labeled as such, and the results are graphed in the plot provided.

Questions

1. What can you say about the point pattern from this ‘first-order’ analysis?
2. What are some of the limitations of first-order analyses of point patterns?

1.2. Ripley’s K distribution: second-order point pattern analysis

Next, review the worksheet titled “Ripley’s K” which contains the calculations for the Ripley’s K distribution.

First, note the sample characteristics in the first table, including the number of points (N), study area size (A), and Lambda (N/A).

Second, note the calculations of Ripley’s K distribution for the vernal pool distribution in the table labeled as such.

- Column 1 represents the distance bins. These were selected arbitrarily to ensure an adequate number of points in each distance class on average.
- Column 2 gives the actual distance for the bin. Thus, the first bin includes all point-to-point distances that are less than 36.6 m. The second bin includes all point-to-point distances that fall between 36.6 and 73.2 m, and so on.
- Column 3, $p(d)$, gives the expected number of points within a distance d_s from an arbitrary point; this is equivalent to the $E(d_s)$ or the what’s in the double summation in the $K(d)$ formulas given above. Note the formula in the corresponding cell.
- Column 4, $K(d)$, gives the Ripley’s K for the corresponding distance class. Note the formula in the corresponding cell.
- Column 5, $L(d)$, gives the standard transformed Ripley’s K for the corresponding distance class. Note the formula in the corresponding cell; this is actually $L(d)-d$, which is the standard transform, although it is often simply labeled as $L(d)$, as is done here.

Third, note the statistics given in the corresponding table for the random point simulation. This table contains the Ripley’s K distribution, given in its standard transform, $L(d)$, summarized for 1,000 random point distributions generated for the same number of points and study area size as the observed point distribution. The columns represent selected percentiles of the empirical distribution.

Lastly, examine the plot of the observed Ripley’s K for the vernal pools and the 2.5 and 97.5 percentiles of the random distribution.

Questions

1. What can you say about the point pattern from this ‘second-order’ analysis?
2. What are some possible limitations or considerations in the use of Ripley’s K to describe spatial point patterns?
3. What are the ecological ramifications of the results?

Part 2: Continuous Data Analysis - Spatial Autocorrelation and Semi-variance

This data set consists of a snag transect data representing the frequency of snags (of all size and decay classes) within a 20x20 m plot centered on each meter along a 1,522 m transect in mature unmanaged forest in the Oregon Coast Range.

First, review the data and calculations given in the worksheet titled “geostatistics”:

- Column 1 gives the position along the transect in 1 meter intervals.
- Column 2 gives the number of snags within a 20x20 m plot centered on the corresponding meter along the transect. Note, the sample plots overlap up to a distance of 20 m. This “sliding window” approach allows for a 1-m resolution in lag distance; however, the autocorrelation for lag distances < 20 m is systematically inflated and should not be interpreted.
- Column 3 gives the lag distance, which is equivalent to the position (column 1) in this case because position is given in 1 meter increments.
- Column 4 gives the semivariance for the corresponding lag distance. Note the formula in the cell – this would be cumbersome to calculate by hand for so many points.
- Column 5 gives the autocorrelation (Moran’s I) for the corresponding lag distance. Note the formula in the cell – again, not an easy calculation to do by hand.

Second, examine the plots provided in the worksheet. The first plot depicts the snag counts within the 20x20 m window against transect position. The second plot is a semivariogram, which is a plot of semivariance against lag distance. The third plot is a correlogram, which is plot of Moran’s I against lag distance.

Questions

1. How does snag frequency behave as a function of distance along the transect?
2. Examine the semivariogram.
 - a. Is there an identifiable nugget? Range? Sill?
 - b. Does the regionalized variable (snag counts) exhibit spatial dependence?
3. Examine the autocorrelogram.
 - a. Is there spatial autocorrelation?
 - b. How would you describe the nature of the spatial variation?
 - c. Does the pattern consist of patches? Noise? A dominant trend?
 - d. Is there periodicity in the data?
4. What are the ecological ramifications of the results?

Part 3: Categorical Map Pattern Analysis - Lacunarity

This data set consists of four binary categorical map patterns shown in Appendix B for a 160x160 m area (grid dimensions 16x16; 10 m cell size). For purposes of this exercise, the pattern can be interpreted as representing the distribution of suitable habitat for a species of interest.

First, review the binary maps depicted in the worksheet titled “mosaics” (these are the same as given in Appendix B). Note, all maps contain the same proportion of habitat, $P = 0.5$, and only differ in their spatial configuration.

Second, calculate the Lacunarity for the random map at the scale of 20 m (20x20 m window size). To do this, pass a moving window 2 cells by 2 cells across the map starting in the upper left corner and tabulate the frequency of box masses. Box mass for a 2x2 window can take on values 0,1,2,3, and 4.

Third, check your frequencies against those given in the corresponding table in the worksheet titled “lacunarity”. Specifically, for the “random” table, box size is given by the parameter r , box mass is given by the parameter S , and the frequencies of box masses are given in the column $n(S,r)$. Compare your counts against these.

Fourth, examine the remaining calculations given in the lacunarity table by evaluating the formulas in the cells.

Fifth, examine the lacunarity summary tables provided. The top table contains the raw lacunarity values for box sizes 1 through 8 for each of the binary maps. This table was generated with a computer software program. Note that the results for the box size of 2 differ slightly from the results calculated by hand. The differences occur because the computer algorithm uses an approximation procedure for computational efficiency. The differences do not qualitatively change the results. The bottom table contains the log-transformed values. Specifically, the customary lacunarity plot is the natural log of lacunarity against the natural log of box size (given in number of cells along one side of the box, not meters).

Lastly, examine the lacunarity plot provided.

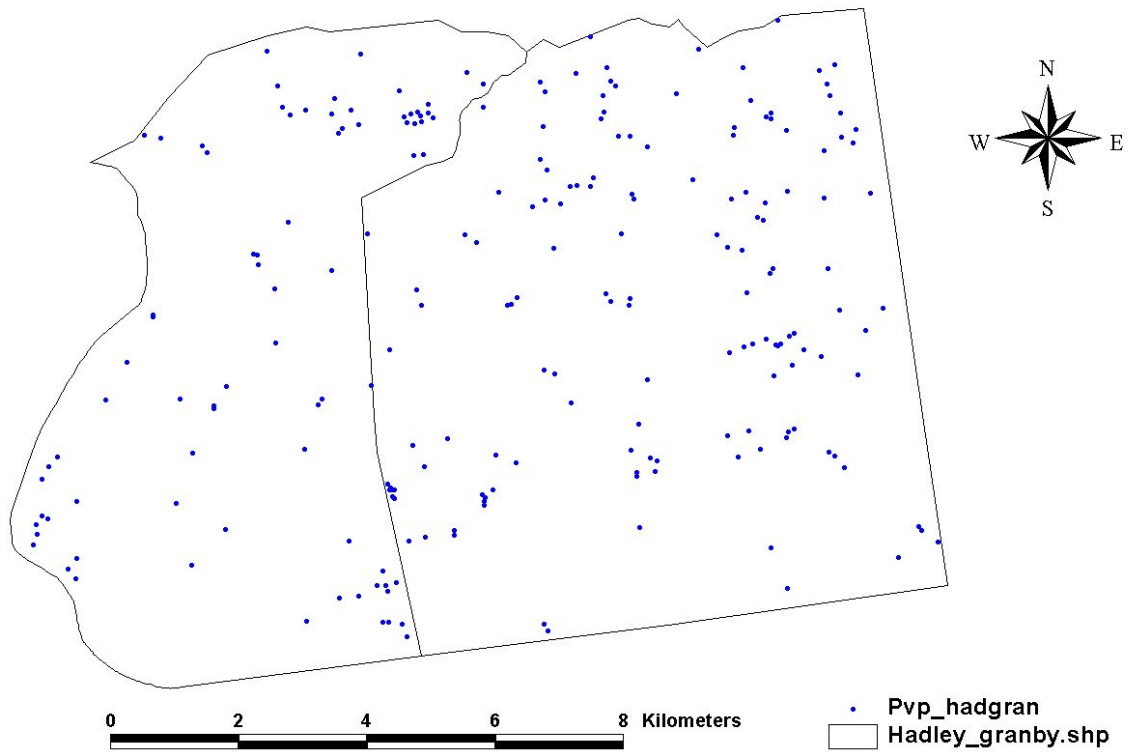
Questions

1. What can you say about the differences in contagion among landscapes?
2. What kind of lacunarity curve would you expect for a perfectly self-similar pattern?
3. What would you expect for a perfectly regular (uniform) pattern?

Appendix A

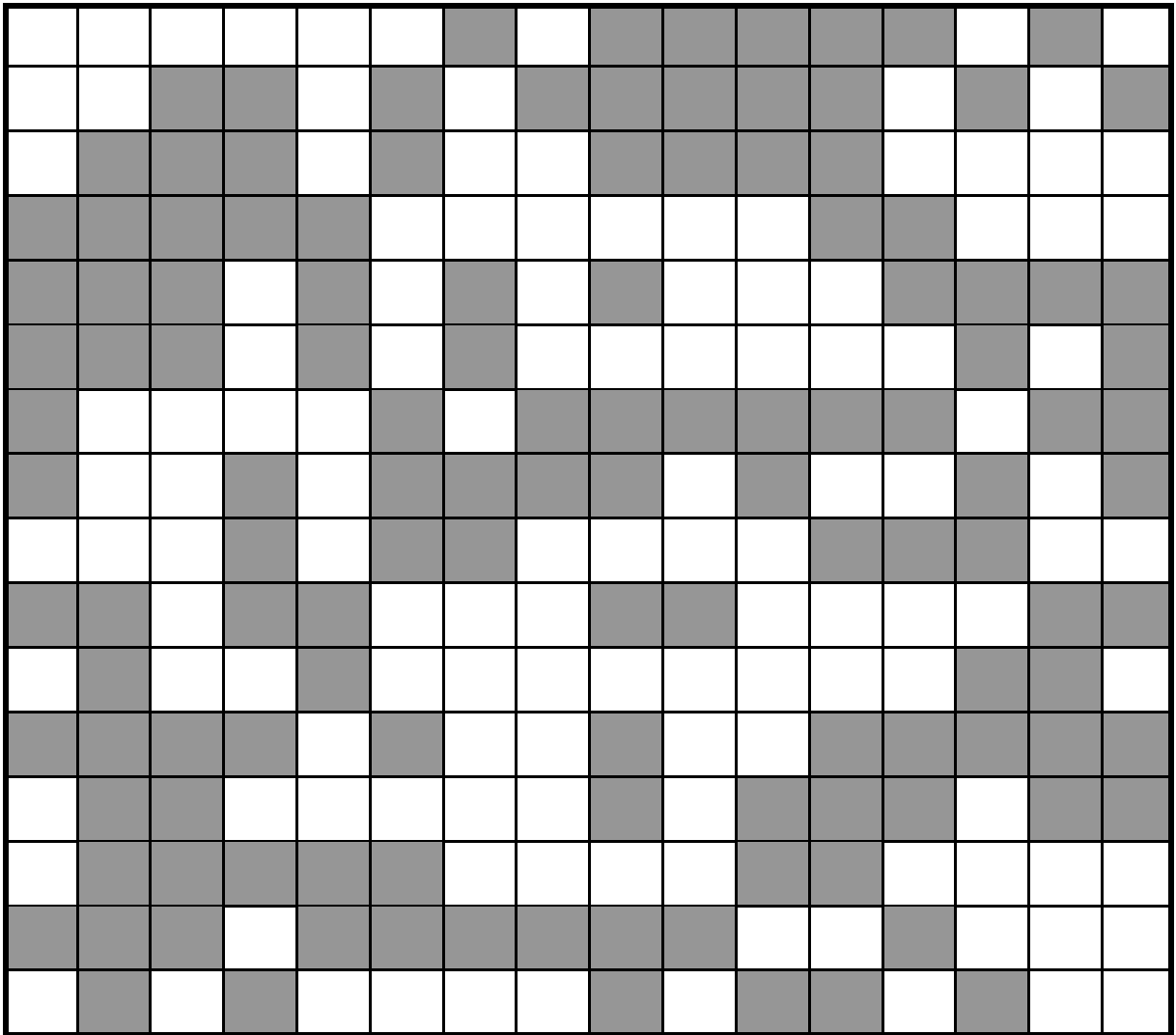
Map of 225 potential vernal pools in South Hadley and Granby Massachusetts.

Potential Vernal Pools (South Hadley & Granby MA)

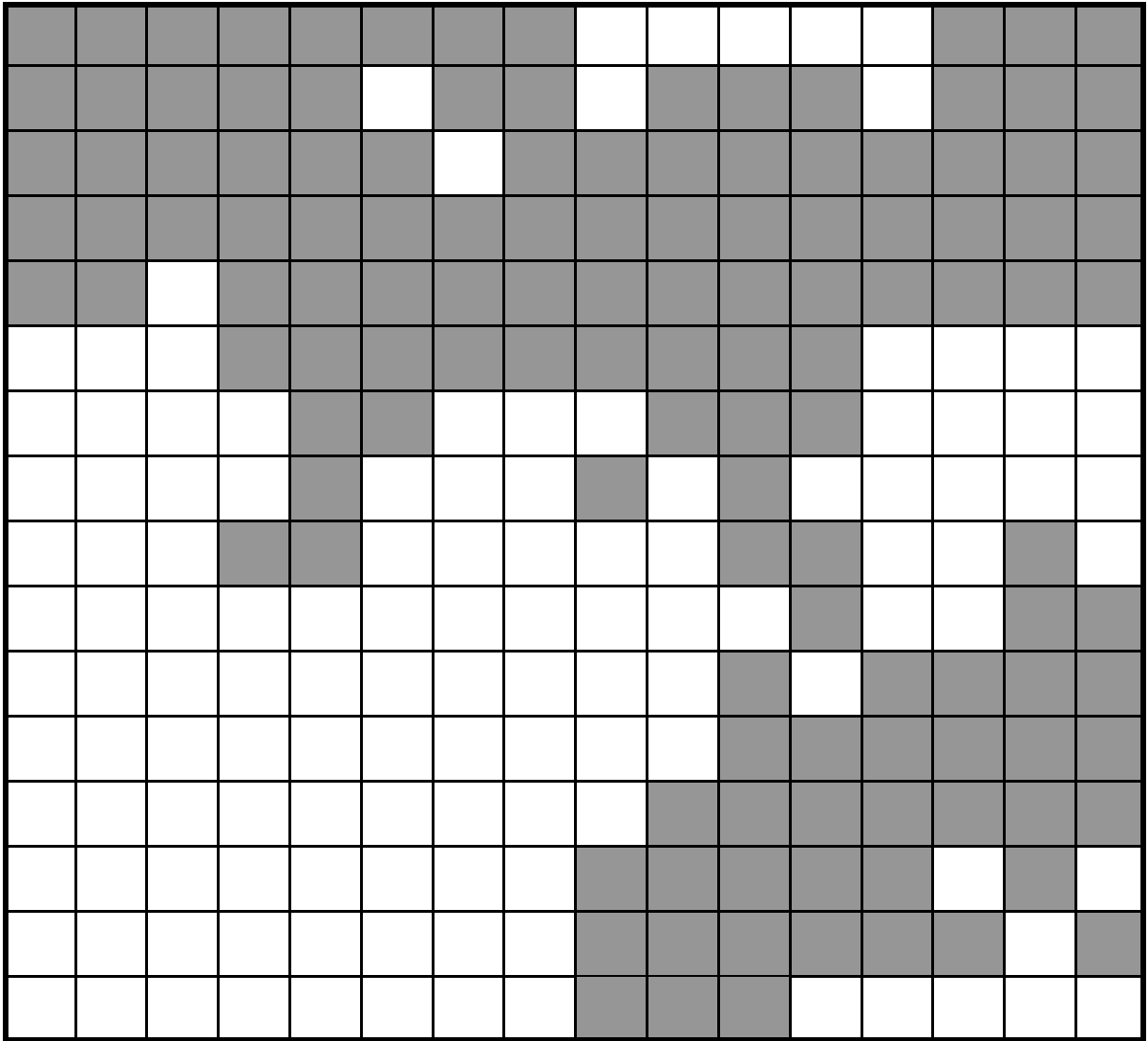


Appendix B

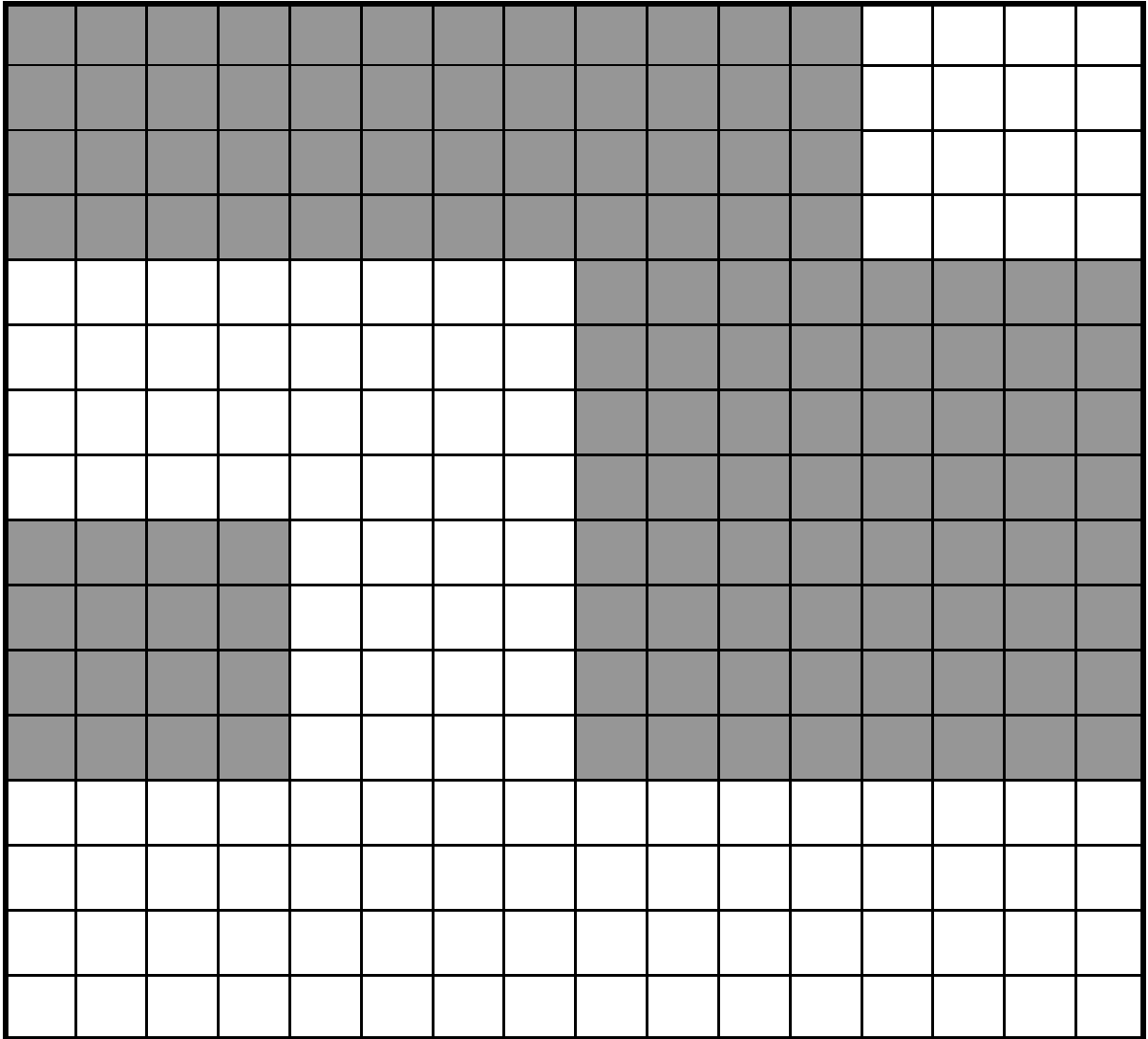
Random Habitat Map.--Map pattern for a random distribution of habitat across a 160x160 m landscape, where the proportion of the landscape occupied by habitat (shown in dark) is 0.5 and the cells are 10x10 m.



Contagious (Fractal) Habitat Map.--Map pattern for a highly contagious fractal ($H = .8$) distribution of habitat across a 160x160 m landscape, where the proportion of the landscape occupied by habitat (shown in dark) is 0.5 and the cells are 10x10 m.



Hierarchical ('Top') Habitat Map.--Map pattern for a two-level hierarchical (curdled) distribution of habitat (shown in dark) across a 160x160 m landscape where the cells are 10x10 m. First level--size = 4 (i.e., 4x4) with 0.5 probability of habitat; second level--size = 2 (i.e., 2x2) with 1.0 probability of habitat. The overall probability of habitat is 0.5 (i.e., 0.5x1.0).



Hierarchical ('Bottom') Habitat Map.--Map pattern for a two-level hierarchical (curdled) distribution of habitat (shown in dark) across a 160x160 m landscape where the cells are 10x10 m. First level--size = 4 (i.e., 4x4) with 1.0 probability of habitat; second level--size = 2 (i.e., 2x2) with 0.5 probability of habitat. The overall probability of habitat is 0.5 (i.e., 1.0x0.5).

