Excerpt from Chapter 13 of:

Legendre, P. and L. Legendre. 1998. Numerical ecology, 2nd English edition. Elsevier Science BV, Amsterdam. xv + 853 pages.

13.1 Structure functions

Ecologists are interested in describing spatial structures in quantitative ways and testing for the presence of spatial autocorrelation in data. The primary objective is to:

- either support the null hypothesis that no significant spatial autocorrelation is present in a data set, or that none remains after detrending (Subsection 13.2.1) or after

Table 13.1 Surface pattern analysis: research objectives and related numerical methods. Modified from Legendre & Fortin (1989).

Research objective	Numerical methods
1) Description of spatial structures and testing for the presence of spatial autocorrelation (Descriptions using structure functions should always be complemented by maps.)	Univariate structure functions: correlogram, variogram, etc. (Section 13.1)
	Multivariate structure functions: Mantel correlogram (Section 13.1)
	Testing for a gradient in multivariate data: (1) constrained (canonical) ordination between the multivariate data and the geographic coordinates of the sites (Section 13.4). (2) Mantel test between ecological distances (computed from the multivariate data) and geographic distances (Subsection 10.5.1)
2) Mapping; estimation of values at given locations	Univariate data: local interpolation map; trend-surface map (global statistical model) (Sect. 13.2)
	Multivariate data: clustering with spatial contiguity constraint, search for boundaries (Section 13.3); interpolated map of the 1st (2nd, etc.) ordination axis (Section 13.4); multivariate trend-surface map obtained by constrained ordination (canonical analysis) (Section 13.4)
3) Modelling species-environment relationships while taking spatial structures into account	Raw data tables: partial canonical analysis (Section 13.5)
	Distance matrices: partial Mantel analysis (Section 13.6)
4) Performing valid statistical tests on autocorrelated data	Subsection 1.1.1

controlling for the effect of explanatory (e.g., environmental) variables, thus insuring valid use of the standard univariate or multivariate statistical tests of hypotheses.

• or reject the null hypothesis and show that significant spatial autocorrelation is present in the data, in order to use it in conceptual or statistical models.

Tests of spatial autocorrelation coefficients may only support or reject the null hypothesis of the absence of significant spatial structure. When significant spatial structure is found, it may correspond, or not, to spatial autocorrelation (Section 1.1, model b) — depending on the hypothesis of the investigator.

Map

Spatial structures may be described through *structure functions*, which allow one to quantify the spatial dependence and partition it amongst distance classes. Interpretation of this description is usually supported by maps of the univariate or multivariate data (Sections 13.2 to 13.4). The most commonly used structure functions are correlograms, variograms, and periodograms.

Spatial correlogram

A *correlogram* is a graph in which autocorrelation values are plotted, on the ordinate, against *distance classes* among sites on the abscissa. Correlograms (Cliff & Ord 1981) can be computed for single variables (Moran's I or Geary's c autocorrelation coefficients, Subsection 1) or for multivariate data (Mantel correlogram, Subsection 5); both types are described below. In all cases, a test of significance is available for each individual autocorrelation coefficient plotted in a correlogram.

Variogram

Similarly, a *variogram* is a graph in which semi-variance is plotted, on the ordinate, against *distance classes* among sites on the abscissa (Subsection 3). In the geostatistical tradition, semi-variance statistics are not tested for significance, although they could be through the test developed for Geary's c, when the condition of second-order stationarity is satisfied (Subsection 13.1.1). Statistical models may be fitted to variograms (linear, exponential, spherical, Gaussian, etc.); they allow the investigator to relate the observed structure to hypothesized generating processes or to produce interpolated maps by kriging (Subsection 13.2.2).

Because they measure the relationship between pairs of observation points located a certain distance apart, correlograms and variograms may be computed either for preferred geographic directions or, when the phenomenon is assumed to be isotropic in space, in an all-directional way.

2-D periodogram

A *two-dimensional Schuster* (1898) *periodogram* may be computed when the structure under study is assumed to consist of a combination of sine waves propagated through space. The basic idea is to fit sines and cosines of various periods, one period at a time, and to determine the proportion of the series' variance (r^2) explained by each period. In periodograms, the abscissa is either a period or its inverse, a frequency; the ordinate is the proportion of variance explained. Two-dimensional periodograms may be plotted for all combinations of directions and spatial frequencies. The technique is described Priestley (1964), Ripley (1981), Renshaw and Ford (1984) and Legendre & Fortin (1989). It is not discussed further in the present book.

1 — Spatial correlograms

For quantitative variables, spatial autocorrelation may be measured by either Moran's I (1950) or Geary's c (1954) spatial autocorrelation statistics (Cliff & Ord, 1981):

Moran's *I*:
$$I(d) = \frac{\frac{1}{W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - \bar{y})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{for } h \neq i \qquad \textbf{(13.1)}$$

Geary's *c*:
$$c(d) = \frac{\frac{1}{2W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - y_i)^2}{\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{for } h \neq i \qquad \textbf{(13.2)}$$

The y_h's and y_i's are the values of the observed variable at sites h and i. Before computing spatial autocorrelation coefficients, a matrix of geographic distances $\mathbf{D} = [D_{hi}]$ among observation sites must be calculated. In the construction of a correlogram, spatial autocorrelation coefficients are computed, in turn, for the various distance classes d. The weights w_{hi} are Kronecker deltas (as in eq. 7.20); the weights take the value $w_{hi} = 1$ when sites h and i are at distance d and $w_{hi} = 0$ otherwise. In this way, only the pairs of sites (h, i) within the stated distance class (d) are taken into account in the calculation of any given coefficient. This approach is illustrated in Fig. 13.3. W is the sum of the weights w_{hi} for the given distance class, i.e. the number of pairs used to calculate the coefficient. For a given distance class, the weights w_{ij} are written in a $(n \times n)$ matrix **W**. Jumars *et al.* (1977) present ecological examples where the distance^{-1} or distance^{-2} among adjacent sites is used for weight instead of 1's.

The numerators of eqs. 13.1 and 13.2 are written with summations involving each pair of objects twice; in eq. 13.2 for example, the terms $(y_h - y_i)^2$ and $(y_i - y_h)^2$ are both used in the summation. This allows for cases where the distance matrix **D** or the weight matrix **W** is asymmetric. In studies of the dispersion of pollutants in soil, for instance, drainage may make it more difficult to go from A to B than from B to A; this may be recorded as a larger distance from A to B than from B to A. In spatio-temporal analyses, an observed value may influence a later value at the same or a different site, but not the reverse. An impossible connection may be coded by a very large value of distance. In most applications, however, the geographic distance matrix among sites is symmetric and the coefficients may be computed from the half-matrix of distances; the formulae remain the same, in that case, because W, as well as the sum in the numerator, are half the values computed over the whole distance matrix **D** (except $h = i$).

One may use distances along a network of connections (Subsection 13.3.1) instead of straight-line geographic distances; this includes the "chess moves" for regularly-spaced points as obtained from systematic sampling designs: rook's, bishop's, or king's connections (see Fig. 13.19). For very broad-scale studies, involving a whole ocean for instance, "great-circle distances", i.e. distances along earth's curved surface, should be used instead of straight-line distances through the earth crust.

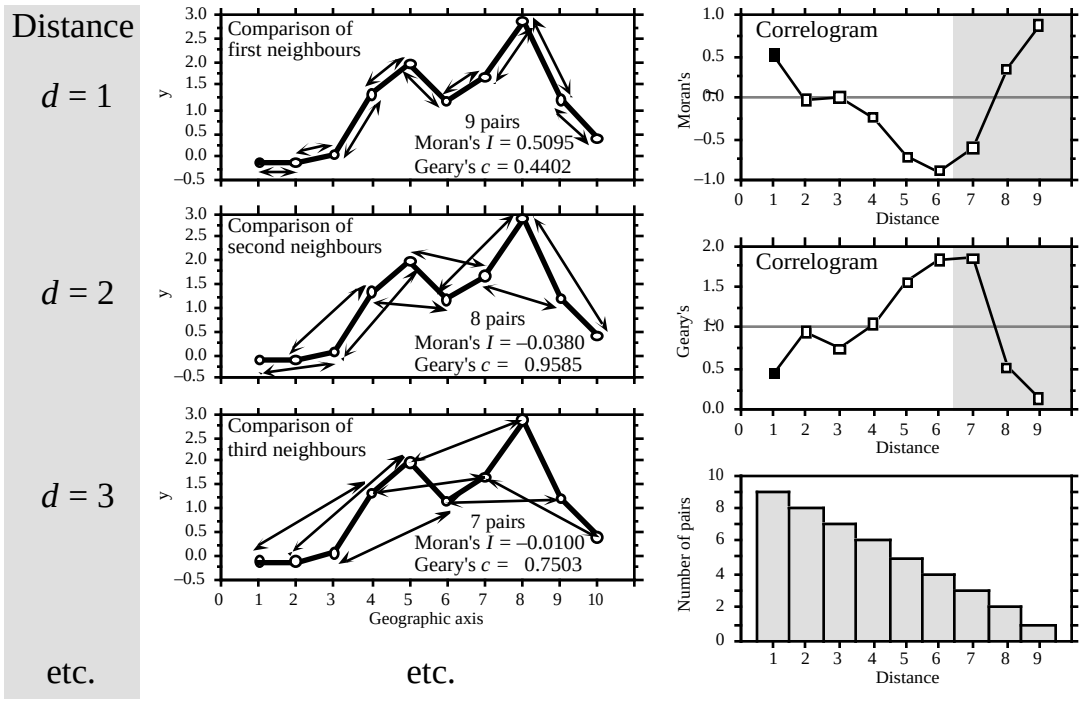


Figure 13.3 Construction of correlograms. Left: data series observed along a single geographic axis (10 equispaced observations). Moran's I and Geary's c statistics are computed from pairs of observations found at preselected distances ($d = 1$, $d = 2$, $d = 3$, etc.). Right: correlograms are graphs of the autocorrelation statistics plotted against distance. Dark squares: significant autocorrelation statistics ($p \leq 0.05$). Lower right: histogram showing the number of pairs in each distance class. Coefficients for the larger distance values (grey zones in correlograms) should not be considered in correlograms, nor interpreted, because they are based on a small number of pairs (test with low power) and only include the pairs of points bordering the series or surface.

Moran's I formula is related to Pearson's correlation coefficient; its numerator is a covariance, comparing the values found at all pairs of points in turn, while its denominator is the maximum-likelihood estimator of the variance (i.e. division by n instead of $n - 1$); in Pearson's r, the denominator is the product of the standard deviations of the two variables (eq. 4.7), whereas in Moran's I there is only one variable involved. Moran's I mainly differs from Pearson's r in that the sums in the numerator and denominator of eq. 13.1 do not involve the same number of terms; only the terms corresponding to distances within the given class are considered in the numerator whereas all pairs are taken into account in the denominator. Moran's I usually takes values in the interval [–1, +1] although values lower than –1 or higher than +1 may occasionally be obtained. Positive autocorrelation in the data translates into positive values of I; negative autocorrelation produces negative values.

Readers who are familiar with correlograms in time series analysis will be reassured to know that, when a problem involves equispaced observations along a single physical dimension, as in Fig. 13.3, calculating Moran's I for the different distance classes is nearly the same as computing the autocorrelation coefficient of time series analysis (Fig. 12.5, eq. 12.6); a small numeric difference results from the divisions by $(n-k-1)$ and $(n-1)$, respectively, in the numerator and denominator of eq. 12.6, whereas division is by $(n-k)$ and (n), respectively, in the numerator and denominator of Moran's I formula (eq. 13.1).

Geary's c coefficient is a distance-type function; it varies from 0 to some unspecified value larger than 1. Its numerator sums the squared differences between values found at the various pairs of sites being compared. A Geary's c correlogram varies as the reverse of a Moran's I correlogram; strong autocorrelation produces high values of I and low values of c (Fig. 13.3). Positive autocorrelation translates in values of c between 0 and 1 whereas negative autocorrelation produces values larger than 1. Hence, the reference 'no correlation' value is $c = 1$ in Geary's correlograms.

For sites lying on a surface or in a volume, geographic distances do not naturally fall into a small number of values; this is true for regular grids as well as random or other forms of irregular sampling designs. Distance values must be grouped into distance classes; in this way, each spatial autocorrelation coefficient can be computed using several comparisons of sampling sites.

Numerical example. In Fig. 13.4 (artificial data), 10 sites have been located at random into a 1-km^2 sampling area. Euclidean (geographic) distances were computed among sites. The number of classes is arbitrary and left to the user's decision. A compromise has to be made between resolution of the correlogram (more resolution when there are more, narrower classes) and power of the test (more power when there are more pairs in a distance class). Sturge's rule is often used to decide about the number of classes in histograms; it was used here and gave:

$$\text{Number of classes} = 1 + 3.322 \log_{10}(m) = 1 + 3.3 \log_{10}(45) = 6.46 \qquad \textbf{(13.3)}$$

where m is the number of distances in the upper triangular matrix and 3.322 is $1/\log_{10}2$; the number was rounded to the nearest integer (i.e. 6). The distance matrix was thus recoded into 6 classes, ascribing class numbers (1 to 6) to all distances within a class of the histogram.

An alternative to distance classes with equal widths would be to create distance classes containing the same number of pairs (notwithstanding tied values); distance classes formed in this way are of unequal widths. The advantage is that the tests of significance have the same power across all distance classes because they are based upon the same number of pairs of observations. The disadvantages are that limits of the distance classes are more difficult to find and correlograms are harder to draw.

Spatial autocorrelation coefficients can be tested for significance and confidence intervals can be computed. With proper correction for multiple testing, one can determine whether a significant spatial structure is present in the data and what are the distance classes showing significant positive or negative autocorrelation. Tests of significance require, however, that certain conditions specified below be fulfilled.

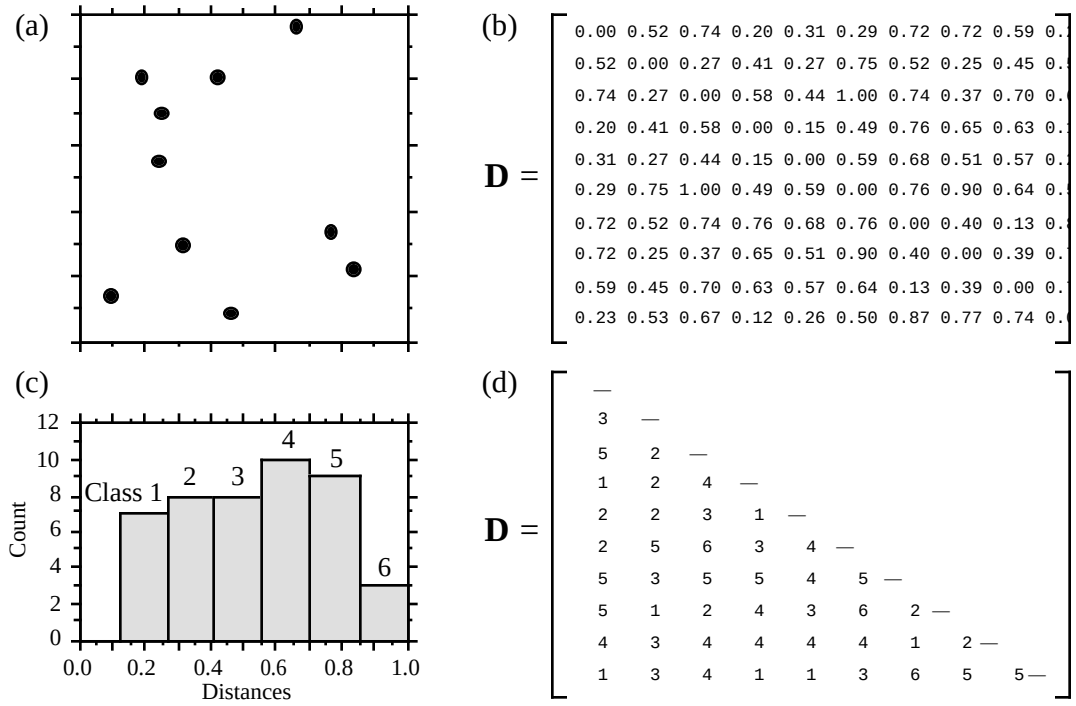


Figure 13.4 Calculation of distance classes, artificial data. (a) Map of 10 sites in a 1-km^2 sampling area. (b) Geographic distance matrix (**D**, in km). (c) Frequency histogram of distances (classes 1 to 6) for the upper (or lower) triangular portion of **D**. (d) Distances recoded into 6 classes.

Second-order stationarity

The tests require that the condition of *second-order stationarity* be satisfied. This rather strong condition states that the expected value (mean) and spatial covariance (numerator of eq. 13.1) of the variable is the same all over the study area, and the variance (denominator of eq. 13.1) is finite. The value of the autocorrelation function depends only on the length and orientation of the vector between any two points, not on its position in the study area (David, 1977).

Intrinsic assumption

A relaxed form of stationarity hypothesis, the *intrinsic assumption*, states that the differences $(y_h - y_i)$ for any distance d (in the numerator of eq. 13.2) must have zero mean and constant and finite variance over the study area, independently of the location where the differences are calculated. Here, one considers the *increments* of the values of the regionalized variable instead of the values themselves (David, 1977). As shown below, the variance of the increments is the variogram function. In layman's terms, this means that a single autocorrelation function is adequate to describe the entire surface under study. An example where the intrinsic assumption does not hold is

a region which is half plain and half mountains; such a region should be divided in two subregions in which the variable "altitude" could be modelled by separate autocorrelation functions. This condition must always be met when variograms or correlograms (including multivariate Mantel correlograms) are computed, even for descriptive purpose.

Cliff & Ord (1981) describe how to compute confidence intervals and test the significance of spatial autocorrelation coefficients. For any normally distributed statistic *Stat*, a confidence interval at significance level α is obtained as follows:

$$Pr\,(Stat - z_{\alpha/2}\sqrt{\mathrm{Var}\,(Stat)} < Stat_{\mathrm{Pop}} < Stat + z_{\alpha/2}\sqrt{\mathrm{Var}\,(Stat)}) \;=\; 1-\alpha \qquad \textbf{(13.4)}$$

For significance testing with large samples, a one-tailed critical value $Stat_\alpha$ at significance level α is obtained as follows:

$$Stat_\alpha = z_\alpha\sqrt{\mathrm{Var}\,(Stat)} + \text{Expected value of } Stat \text{ under } \mathrm{H}_0 \qquad \textbf{(13.5)}$$

It is possible to use this approach because both I and c are asymptotically normally distributed for data sets of moderate to large sizes (Cliff & Ord, 1981). Values $z_{\alpha/2}$ or z_α are found in a table of standard normal deviates. Under the hypothesis (H_0) of random spatial distribution of the observed values y_i, the expected values (E) of Moran's I and Geary's c are:

$$\mathrm{E}(I) = -(n-1)^{-1} \quad \text{and} \quad \mathrm{E}(c) = 1 \qquad \textbf{(13.6)}$$

Under the null hypothesis, the expected value of Moran's I approaches 0 as n increases. The variances are computed as follows under a randomization assumption, which simply states that, under H_0, the observations y_i are independent of their positions in space and, thus, are exchangeable:

$$\mathrm{Var}\,(I) \;=\; \mathrm{E}(I^2) - [\mathrm{E}(I)]^2 \qquad \textbf{(13.7)}$$

$$\mathrm{Var}\,(I) \;=\; \frac{n\,[(n^2-3n+3)\,S_1 - nS_2 + 3W^2] - b_2\,[(n^2-n)\,S_1 - 2nS_2 + 6W^2]}{(n-1)\,(n-2)\,(n-3)\,W^2} - \frac{1}{(n-1)^2}$$

$$\mathrm{Var}\,(c) \;=\; \frac{(n-1)\,S_1\,[n^2-3n+3-(n-1)\,b_2]}{n\,(n-2)\,(n-3)\,W^2} \qquad \textbf{(13.8)}$$

$$+\;\frac{-0.25\,(n-1)\,S_2\,[n^2+3n-6-(n^2-n+2)\,b_2] + W^2\,[n^2-3+(-(n-1)^2)\,b_2]}{n\,(n-2)\,(n-3)\,W^2}$$

In these equations,

- $S_1 = \frac{1}{2}\sum_{h=1}^{n}\sum_{i=1}^{n}(w_{hi} + w_{ih})^2$ (there is a term of this sum for *each cell* of matrix **W**);
- $S_2 = \sum_{i=1}^{n}(w_{i+} + w_{+i})^2$ where w_{i+} and w_{+i} are respectively the sums of row i and column i of matrix **W**;
- $b_2 = n\sum_{i=1}^{n}(y_i - \bar{y})^4 / \left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^2$ measures the kurtosis of the distribution;
- W is as defined in eqs. 13.1 and 13.2.

In most cases in ecology, tests of spatial autocorrelation are one-tailed because the sign of autocorrelation is stated in the ecological hypothesis; for instance, contagious biological processes such as growth, reproduction, and dispersal, all suggest that ecological variables are positively autocorrelated at short distances. To carry out an approximate test of significance, select a value of α (e.g. $\alpha = 0.05$) and find z_α in a table of the standard normal distribution (e.g. $z_{0.05} = +1.6452$). Critical values are found as in eq. 13.5, with a correction factor that becomes important when n is small:

- $I_\alpha = z_\alpha\sqrt{\mathrm{Var}(I)} - k_\alpha(n-1)^{-1}$ in all cases, using the value in the upper tail of the z distribution when testing for positive autocorrelation (e.g. $z_{0.05} = +1.6452$) and the value in the lower tail in the opposite case (e.g. $z_{0.05} = -1.6452$).

- $c_\alpha = z_\alpha\sqrt{\mathrm{Var}(c)} + 1$ when $c < 1$ (positive autocorrelation), using the value in the lower tail of the z distribution (e.g. $z_{0.05} = -1.6452$).

- $c_\alpha = z_\alpha\sqrt{\mathrm{Var}(c)} + 1 - k_\alpha(n-1)^{-1}$ when $c > 1$ (negative autocorrelation), using the value in the upper tail of the z distribution (e.g. $z_{0.05} = +1.6452$).

The value taken by the correction factor k_α depends on the values of n and W. If $4(n - \sqrt{n}) < W \leq 4(2n - 3\sqrt{n} + 1)$, then $k_\alpha = \sqrt{10\alpha}$; otherwise, $k_\alpha = 1$. If the test is two-tailed, use $\alpha^* = \alpha/2$ to find z_{α^*} and k_{α^*} before computing critical values. These corrections are based upon simulations reported by Cliff & Ord (1981, section 2.5).

Other formulas are found in Cliff & Ord (1981) for conducting a test under the assumption of normality, where one assumes that the y_i's result from n independent draws from a normal population. When n is very small, tests of I and c should be conducted by randomization (Section 1.2).

Moran's I and Geary's c are sensitive to extreme values and, in general, to asymmetry in the data distributions, as are the related Pearson's r and Euclidean distance coefficients. Asymmetry increases the variance of the data. It also increases the kurtosis and hence the variance of the I and c coefficients (eqs. 13.7 and 13.8); this

makes it more difficult to reach significance in statistical tests. So, practitioners usually attempt to normalize the data before computing correlograms and variograms.

Statistical testing in correlograms implies multiple testing since a test of significance is carried out for each autocorrelation coefficient. Oden (1984) has developed a Q statistic to test the global significance of spatial correlograms; his test is an extension of the Portmanteau Q-test used in time series analysis (Box & Jenkins, 1976). An alternative global test is to check whether the correlogram contains at least one autocorrelation statistic which is significant at the Bonferroni-corrected significance level (Box 1.3). Simulations in Oden (1984) show that the power of the Q-test is not appreciably greater than the power of the Bonferroni procedure, which is computationally a lot simpler. A practical question remains, though: how many distance classes should be created? This determines the number of simultaneous tests that are carried out. More classes mean more resolution but fewer pairs per class and, thus, less power for each test; more classes also mean a smaller Bonferroni-corrected α' level, which makes it more difficult for a correlogram to reach global significance.

When the overall test has shown global significance, one may wish to identify the individual autocorrelation statistics that are significant, in order to reach an interpretation (Subsection 2). One could rely on Bonferroni-corrected tests for all individual autocorrelation statistics, but this approach would be too conservative; a better solution is to use Holm's correction procedure (Box 1.3). Another approach is the *progressive Bonferroni correction* described in Subsection 12.4.2; it is only applicable when the ecological hypothesis indicates that significant autocorrelation is to be expected in the smallest distance classes and the purpose of the analysis is to determine the extent of the autocorrelation (i.e. which distance class it reaches). With the progressive Bonferroni approach, the likelihood of emergence of significant values decreases as one proceeds from left to right, i.e. from the small to the large distance classes of the correlogram. One does not have to limit the correlogram to a small number of classes to reduce the effect of the correction, as it is the case with Oden's overall test and with the Bonferroni and Holm correction methods. This approach will be used in the examples that follow.

Autocorrelation coefficients and tests of significance also exist for qualitative (nominal) variables (Cliff & Ord 1981); they have been used to analyse for instance spatial patterns of sexes in plants (Sakai & Oden 1983; Sokal & Thomson 1987). Special types of spatial autocorrelation coefficients have been developed to answer specific problems (e.g. Galiano 1983; Estabrook & Gates 1984). The paired-quadrat variance method, developed by Goodall (1974) to analyse spatial patterns of ecological data by random pairing of quadrats, is related to correlograms.

2 — Interpretation of all-directional correlograms

Isotropy
Anisotropy

When the autocorrelation function is the same for all geographic directions considered, the phenomenon is said to be *isotropic*. Its opposite is *anisotropy*. When a variable is isotropic, a single correlogram may be computed over all directions of the

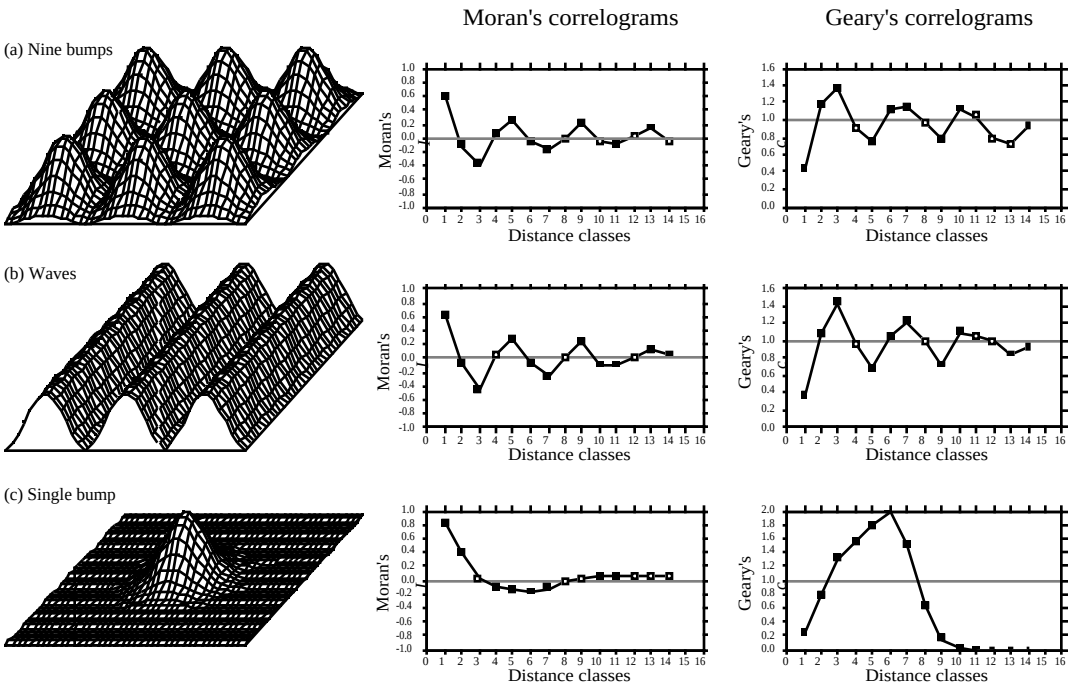


Figure 13.5 Spatial autocorrelation analysis of artificial spatial structures shown on the left: (a) nine bumps; (b) waves; (c) a single bump. Centre and right: all-directional correlograms. Dark squares: autocorrelation statistics that remain significant after progressive Bonferroni correction ($\alpha = 0.05$); white squares: non-significant values.

study area. The correlogram is said to be *all-directional* or *omnidirectional*. Directional correlograms, which are computed for a single direction of space, are discussed together with anisotropy and directional variograms in Subsection 3.

Correlograms are analysed mostly by looking at their shapes. Examples will help clarify the relationship between spatial structures and all-directional correlograms. The important message is that, although correlograms may give clues as to the underlying spatial structure, the information they provide is not specific; a blind interpretation may often be misleading and should always be supported by maps (Section 13.2).

Numerical example. Artificial data were generated that correspond to a number of spatial patterns. The data and resulting correlograms are presented in Fig. 13.5.

• Nine bumps — The surface in Fig. 13.5a is made of nine bi-normal curves. 225 points were sampled across the surface using a regular 15×15 grid (Fig. 13.5f). The "height" was noted at each sampling point. The 25 200 distances among points found in the upper-triangular portion of the distance matrix were divided into 16 distance classes, using Sturge's rule (eq. 13.3), and

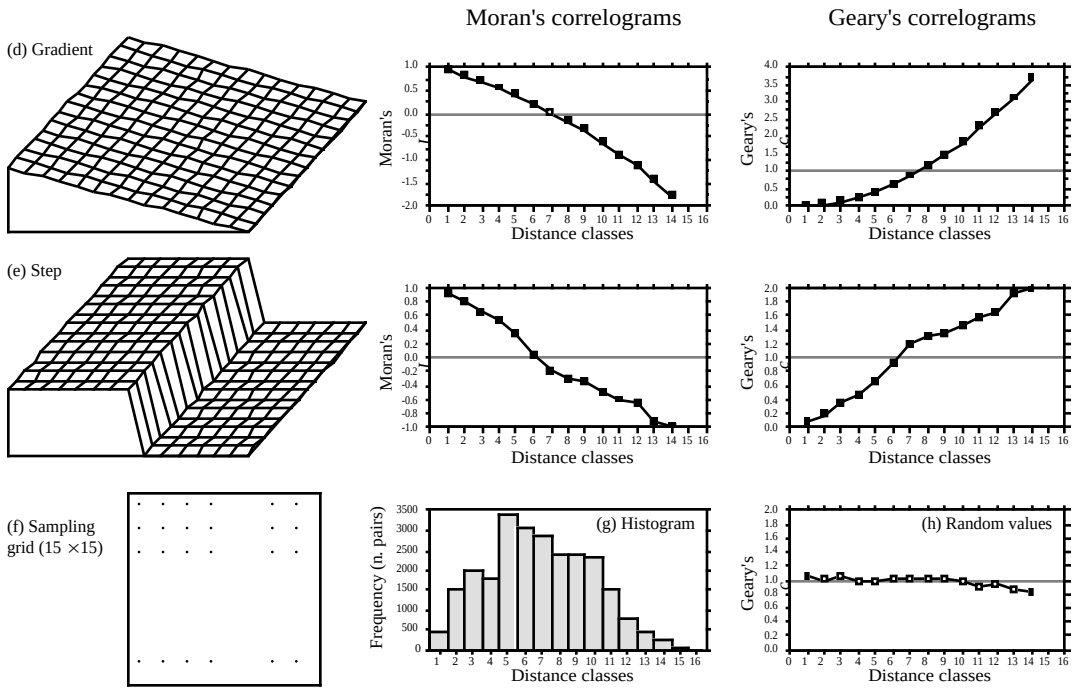


Figure 13.5 **(continued)** Spatial autocorrelation analysis of artificial spatial structures shown on the left: (d) gradient; (e) step. (h) All-directional correlogram of random values. (f) Sampling grid used on each of the artificial spatial structures to obtain 225 "observed values" for spatial autocorrelation analysis. (g) Histogram showing the number of pairs in each distance class. Distances, from 1 to 19.8 in units of the sampling grid, were grouped into 16 distance classes. Spatial autocorrelation statistics (I or c) are not shown for distance classes 15 and 16; see text.

correlograms were computed. According to Oden's test, the correlograms are globally significant at the $\alpha = 5\%$ level since several individual values are significant at the Bonferroni-corrected level $\alpha' = 0.05/16 = 0.00312$. In each correlogram, the progressive Bonferroni correction method was applied to identify significant spatial autocorrelation coefficients: the coefficient for distance class 1 was tested at the $\alpha = 0.05$ level; the coefficient for distance class 2 was tested at the $\alpha' = 0.05/2$ level; and, more generally, the coefficient for distance class k was tested at the $\alpha' = 0.05/k$ level. Spatial autocorrelation coefficients are not reported for distance classes 15 and 16 (60 and 10 pairs, respectively) because they only include the pairs of points bordering the surface, to the exclusion of all other pairs.

There is a correspondence between individual significant spatial autocorrelation coefficients and the main elements of the spatial structure. The correspondence can clearly be seen in this example, where the data generating process is known. This is not the case when analysing field data, in which case the existence and nature of the spatial structures must be confirmed by mapping the data. The presence of several equispaced patches produces an alternation of

significant positive and negative values along the correlograms. The first spatial autocorrelation coefficient, which is above 0 in Moran's correlogram and below 1 in Geary's, indicates positive spatial autocorrelation in the first distance class; the first class contains the 420 pairs of points that are at distance 1 of each other on the grid (i.e. the first neighbours in the N-S or E-W directions of the map). Positive and significant spatial autocorrelation in the first distance class confirms that the distance between first neighbours is smaller than the patch size; if the distance between first neighbours in this example was larger than the patch size, first neighbours would be dissimilar in values and autocorrelation would be negative for the first distance class. The next peaking positive autocorrelation value (which is smaller than 1 in Geary's correlogram) occurs at distance class 5, which includes distances from 4.95 to 6.19 in grid units; this corresponds to positive autocorrelation between points located at similar positions on neighbouring bumps, or neighbouring troughs; distances between successive peaks are 5 grid units in the E-W or N-S directions. The next peaking positive autocorrelation value occurs at distance class 9 (distances from 9.90 to 11.14 in grid units); it includes value 10, which is the distance between second-neighbour bumps in the N-S and E-W directions. Peaking negative autocorrelation values (which are larger than 1 in Geary's correlogram) are interpreted in a similar way. The first such value occurs at distance class 3 (distances from 2.48 to 3.71 in grid units); it includes value 2.5, which is the distance between peaks and troughs in the N-S and E-W directions on the map. If the bumps were unevenly spaced, the correlograms would be similar for the small distance classes, but there would be no other significant values afterwards.

The main problem with all-directional correlograms is that the diagonal comparisons are included in the same calculations as the N-S and E-W comparisons. As distances become larger, diagonal comparisons between, say, points located near the top of the nine bumps tend to fall in different distance classes than comparable N-S or E-W comparisons. This blurs the signal and makes the spatial autocorrelation coefficients for larger distance classes less significant and interpretable.

• Wave (Fig. 13.5b) — Each crest was generated as a normal curve. Crests were separated by five grid units; the surface was constructed in this way to make it comparable to Fig. 13.5a. The correlograms are nearly indistinguishable from those of the nine bumps. All-directional correlograms alone cannot tell apart regular bumps from regular waves; directional correlograms or maps are required.

• Single bump (Fig. 13.5c) — One of the normal curves of Fig. 13.5a was plotted alone at the centre of the study area. Significant negative autocorrelation, which reaches distance classes 6 or 7, delimits the extent of the "range of influence" of this single bump, which covers half the study area. It is not limited here by the rise of adjacent bumps, as this was the case in (a).

• Linear gradient (Fig. 13.5d) — The correlogram is monotonic decreasing. Nearly all autocorrelation values in the correlograms are significant.

True, false gradient

There are actually two kinds of gradients (Legendre, 1993). "True gradients", on the one hand, are deterministic structures. They correspond to generating model 2 of Subsection 1.1.1 (eq. 1.2) and can be modelled using trend-surface analysis (Subsection 13.2.1). The observed values have independent error terms, i.e. error terms which are not autocorrelated. "False gradients", on the other hand, are structures that may look like gradients, but actually correspond to autocorrelation generated by some spatial process (model 1 of Subsection 1.1.1; eq. 1.1). When the sampling area is small relative to the range of influence of the generating process, the data generated by such a process may look like a gradient.

In the case of "true gradients", spatial autocorrelation coefficients should not be tested for significance because the condition of second-order stationarity is not satisfied (definition in previous Subsection); the expected value of the mean is not the same over the whole study area. In the case of "false gradients", however, tests of significance are warranted. For descriptive purposes, correlograms may still be computed for "true gradients" (without tests of significance) because the intrinsic assumption is satisfied. One may also choose to extract a "true gradient" using trend-surface analysis, compute residuals, and look for spatial autocorrelation among the residuals. This is equivalent to trend extraction prior to time series analysis (Section 12.2).

How does one know whether a gradient is "true" or "false"? This is a moot point. When the process generating the observed structure is known, one may decide whether it is likely to have generated spatial autocorrelation in the observed data, or not. Otherwise, one may empirically look at the *target population* of the study. In the case of a spatial study, this is the population of potential sites in the larger area into which the study area is embedded, the study area representing the *statistical population* about which inference can be made. Even from sparse or indirect data, a researcher may form an opinion as to whether the observed gradient is deterministic ("true gradient") or is part of a landscape displaying autocorrelation at broader spatial scale ("false gradient").

• Step (Fig. 13.5e) — A step between two flat surfaces is enough to produce a correlogram which is indistinguishable, for all practical purposes, from that of a gradient. Correlograms alone cannot tell apart regular gradients from steps; maps are required. As in the case of gradients, there are "true steps" (deterministic) and "false steps" (resulting from an autocorrelated process), although the latter is rare. The presence of a sharp discontinuity in a surface generally indicates that the two parts should be subjected to separate analyses. The methods of boundary detection and constrained clustering (Section 13.3) may help detect such discontinuities and delimit homogeneous areas prior to spatial autocorrelation analysis.

• Random values (Fig. 13.5h) — Random numbers, drawn from a standard normal distribution, were generated for each point of the grid and used as the variable to be analysed. Random data are said to represent a "pure nugget effect" in geostatistics. The autocorrelation coefficients were small and non-significant at the 5% level. Only the Geary correlogram is presented.

Sokal (1979) and Cliff & Ord (1981) describe, in general terms, where to expect significant values in correlograms, for some spatial structures such as gradients and large or small patches. Their summary tables are in agreement with the test examples above. The absence of significant coefficients in a correlogram must be interpreted with caution, however:

• It may indicate that the surface under study is free of spatial autocorrelation at the study scale. Beware: this conclusion is subject to type II (or β) error. Type II error depends on the power of the test which is a function of (1) the α significance level, (2) the size of effect (i.e. the minimum amount of autocorrelation) one wants to detect, (3) the number of observations (n), and (4) the variance of the sample (Cohen, 1988):

$$\text{Power} = (1 - \beta) = f(\alpha, \text{size of effect}, n, s_x)$$

Is the test powerful enough to warrant such a conclusion? Are there enough observations to reach significance? The easiest way to increase the power of a test, for a given variable and fixed α, is to increase n.

• It may indicate that the sampling design is inadequate to detect the spatial autocorrelation that may exist in the system. Are the grain size, extent and sampling interval (Section 13.0) adequate to detect the type of autocorrelation one can hypothesize from knowledge about the biological or ecological process under study?

Ecologists can often formulate hypotheses about the mechanism or process that may have generated a spatial phenomenon and deduct the shape that the resulting surface should have. When the model specifies a value for each geographic position (e.g. a spatial gradient), data and model can be compared by correlation analysis. In other instances, the biological or ecological model only specifies process generating the spatial autocorrelation, not the exact geographic position of each resulting value. Correlograms may be used to support or reject the biological or ecological hypothesis. As in the examples of Fig. 13.5, one can construct an artificial model-surface corresponding to the hypothesis, compute a correlogram of that surface, and compare the correlograms of the real and model data. For instance, Sokal *et al.* (1997a) generated data corresponding to several gene dispersion mechanisms in populations and showed the kind of spatial correlogram that may be expected from each model. Another application concerning phylogenetic patterns of human evolution in Eurasia and Africa (space-time model) is found in Sokal *et al.* (1997b).

Bjørnstad & Falck (1997) and Bjørnstad *et al.* (1998) proposed a spline correlogram which provides a continuous and model-free function for the spatial covariance. The spline correlogram may be seen as a modification of the nonparametric covariance function of Hall and co-workers (Hall & Patil, 1994; Hall *et al.*, 1994). A bootstrap algorithm estimates the confidence envelope of the entire correlogram or derived statistics. This method allows the statistical testing of the similarity between correlograms of real and simulated (i.e. model) data.

Ecological application 13.1a

During a study of the factors potentially responsible for the choice of settling sites of *Balanus crenatus* larvae (Cirripedia) in the St. Lawrence Estuary (Hudon *et al.*, 1983), plates of artificial substrate (plastic laminate) were subjected to colonization in the infralittoral zone. Plates were positioned vertically, parallel to one another. A picture was taken of one of the plates after a 3-month immersion at a depth of 5 m below low tide, during the summer 1978. The picture was divided into a (10×15) grid, for a total of 150 pixels of 1.7×1.7 cm. Barnacles were counted by C. Hudon and P. Legendre for the present Ecological application (Fig. 13.6a; unpublished *in op. cit.*). The hypothesis to be tested is that barnacles have a patchy distribution. Barnacles are gregarious animals; their larvae are chemically attracted to settling sites by arthropodine secreted by settled adults (Gabbott & Larman, 1971).

A gradient in larval concentration was expected in the top-to-bottom direction of the plate because of the known negative phototropism of barnacle larvae at the time of settlement (Visscher, 1928). Some kind of border effect was also expected because access to the centre of the plates located in the middle of the pack was more limited than to the fringe. These large-scale effects create violations to the condition of second-order stationarity. A trend-surface equation (Subsection 13.2.1) was computed to account for it, using only the Y coordinate (top-

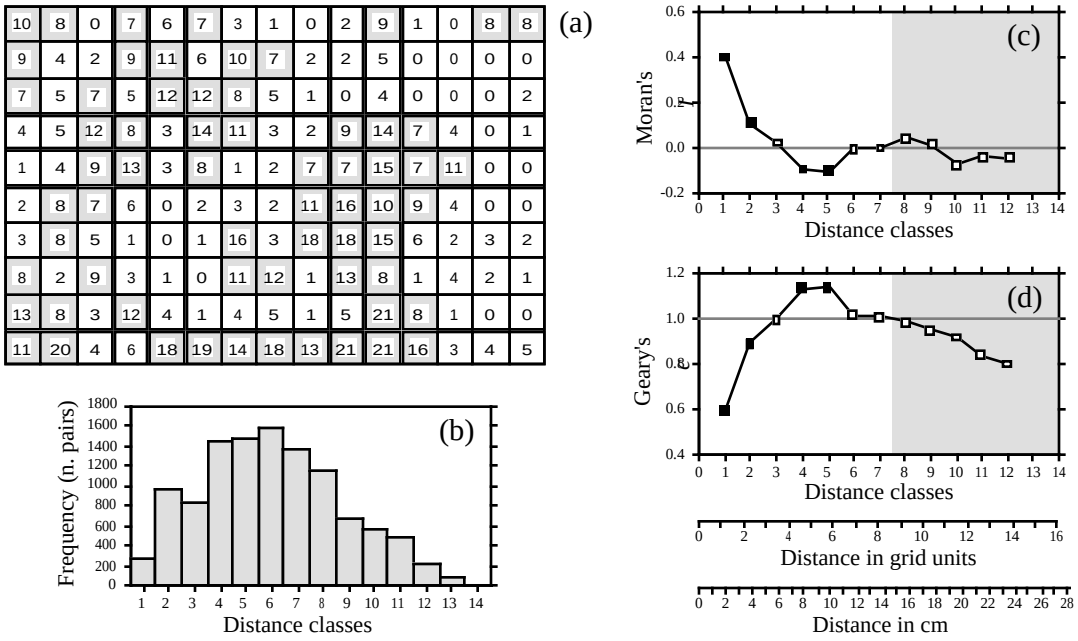


Figure 13.6 (a) Counts of adult barnacles in 150 (1.7×1.7 cm) pixels on a plate of artificial substrate (17×25.5 cm). The mean concentration is 6.17 animals per pixel; pixels with counts ≥ 7 are shaded to display the aggregates. (b) Histogram of the number of pairs in each distance class. (c) Moran's correlogram. (d) Geary's correlogram. Dark squares: autocorrelation statistics that remain significant after progressive Bonferroni correction ($\alpha = 0.05$); white squares: non-significant values. Coefficients for distance classes 13 and 14 are not given because they only include the pairs of points bordering the surface. Distances are also given in grid units and cm.

to-bottom axis). Indeed, a significant trend surface was found, involving Y and Y^2, that accounted for 10% of the variation. It forecasted high barnacle concentration in the bottom part of the plate and near the upper and lower margins. Residuals from this equation were calculated and used in spatial autocorrelation analysis.

Euclidean distances were computed among pixels; following Sturge's rule (eq. 13.3), the distances were divided into 14 classes (Fig. 13.6b). Significant positive autocorrelation was found in the first distance classes of the correlograms (Fig. 13.6c, d), supporting the hypothesis of patchiness. The size of the patches, or "range of influence" (i.e. the distance between zones of high and low concentrations), is indicated by the distance at which the first maximum negative autocorrelation value is found. This occurs in classes 4 and 5, which corresponds to a distance of about 5 in grid units, or 8 to 10 cm. The patches of high concentration are shaded on the map of the plate of artificial substrate (Fig. 13.6a).

In anisotropic situations, directional correlograms should be computed in two or several directions. Description of how the pairs of points are chosen is deferred to Subsection 3 on variograms. One may choose to represent either a single, or several of

these correlograms, one for each of the aiming geographic directions, as seems fit for the problem at hand. A procedure for representing in a single figure the directional correlograms computed for several directions of a plane has been proposed by Oden & Sokal (1986); Legendre & Fortin (1989) give an example for vegetation data. Another method is illustrated in Rossi *et al.* (1992).

Another way to approach anisotropic problems is to compute two-dimensional spectral analysis. This method, described by Priestley (1964), Rayner (1971), Ford (1976), Ripley (1981) and Renshaw & Ford (1984), differs from spatial autocorrelation analysis in the structure function it uses. As in time-series spectral analysis (Section 12.5), the method assumes the data to be stationary (second-order stationarity; i.e. no "true gradient" in the data) and made of a combination of sine patterns. An autocorrelation function $r_{dX,dY}$ for all combinations of lags (dX, dY) in the two geographic axes of a plane, as well as a periodogram with intensity I for all combinations of frequencies in the two directions of the plane, are computed. Details of the calculations are also given in Legendre & Fortin (1989), with an example.

3 — Variogram

Like correlograms, semi-variograms (called *variograms* for simplicity) decompose the spatial (or temporal) variability of observed variables among distance classes. The structure function plotted as the ordinate, called *semi-variance*, is the numerator of eq. 13.2:

$$\gamma(d) = \frac{1}{2W} \sum_{h=1}^{n} \sum_{i=1}^{n} w_{hi} (y_h - y_i)^2 \quad \text{for } h \neq i \qquad \textbf{(13.9)}$$

or, for symmetric distance and weight matrices,

$$\gamma(d) = \frac{1}{2W} \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} w_{hi} (y_h - y_i)^2 \qquad \textbf{(13.10)}$$

$\gamma(d)$ is thus a non-standardized form of Geary's c coefficient. γ may be seen as a measure of the error mean square of the estimate of y_i using a value y_h distant from it by d. The two forms lead to the same numerical value in the case of symmetric distance and weight matrices. The calculation is repeated for different values of d. This provides the *sample variogram*, which is a plot of the empirical values of variance $\gamma(d)$ as a function of distance d.

The equations usually found in the geostatistical literature look a bit different, but they correspond to the same calculations:

$$\gamma(d) = \frac{1}{2W(d)} \sum_{i=1}^{W(d)} (y_i - y_{i+d})^2 \qquad \text{or} \qquad \gamma(d) = \frac{1}{2W(d)} \sum_{(h,\,i)\,|\,d_{hi} \approx d}^{W(d)} (y_h - y_i)^2$$

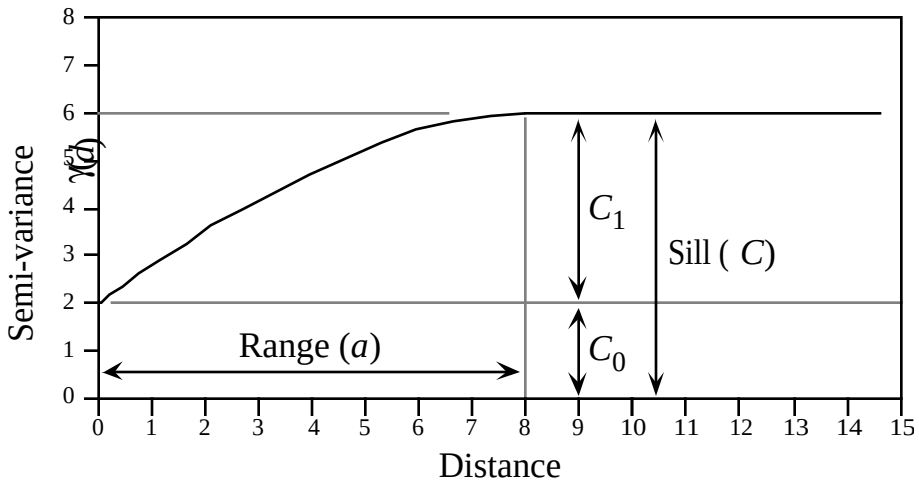


Figure 13.7 Spherical variogram model showing characteristic features: nugget effect ($C_0 = 2$ in this example), spatially structured component ($C_1 = 4$), sill ($C = C_0 + C_1 = 6$), and range ($a = 8$).

Both of these expressions mean that pairs of values are selected to be at distance d of each other; there are $W(d)$ such pairs for any given distance class d. The condition $d_{hi} \approx d$ means that distances may be grouped into distance classes, placing in class d the individual distances d_{hi} that are approximately equal to d. In directional variograms (below), d is a directional measure of distance, i.e. taken in a specified direction only. The semi-variance function is often called the variogram in the geostatistical literature. When computing a variogram, one assumes that the autocorrelation function applies to the entire surface under study (intrinsic hypothesis, Subsection 13.1.1).

Generally, variograms tend to level off at a *sill* which is equal to the variance of the variable (Fig. 13.7); the presence of a sill implies that the data are second-order stationary. The distance at which the variance levels off is referred to as the *range* (parameter a); beyond that distance, the sampling units are not spatially correlated. The discontinuity at the origin (non–zero intercept) is called the *nugget effect*; the geostatistical origin of the method transpires in that name. It corresponds to the local variation occurring at scales finer than the sampling interval, such as sampling error, fine-scale spatial variability, and measurement error. The nugget effect is represented by the error term ε_{ij} in spatial structure model 1b of Subsection 1.1.1. It describes a portion of variation which is not autocorrelated, or is autocorrelated at a scale finer than can be detected by the sampling design. The parameter for the nugget effect is C_0 and the spatially structured component is represented by C_1; the sill, C, is equal to $C_0 + C_1$. The *relative nugget effect* is $C_0/(C_0 + C_1)$.

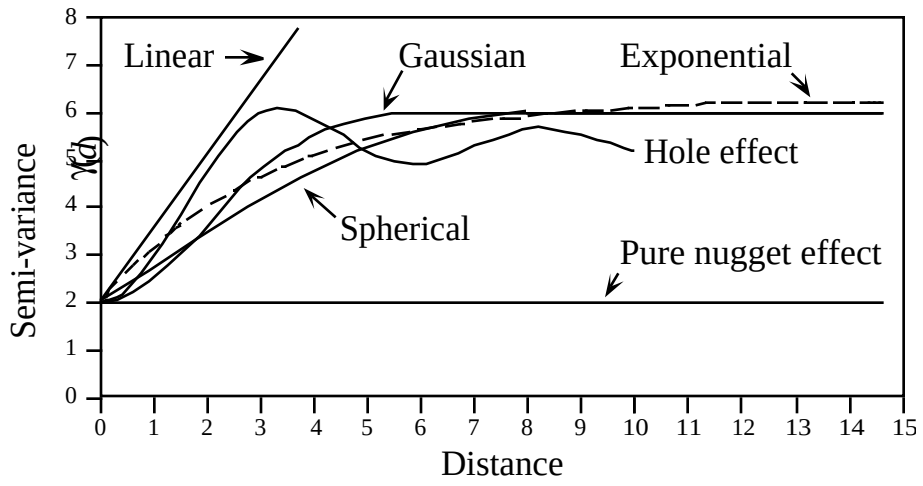


Figure 13.8 Commonly used variogram models.

Although a sample variogram is a good descriptive summary of the spatial contiguity of a variable, it does not provide all the semi-variance values needed for kriging (Subsection 13.2.2). A model must be fitted to the sample variogram; the model will provide values of semi-variance for all the intermediate distances. The most commonly used models are the following (Fig. 13.8):

- Spherical model: $\gamma(d) = C_0 + C_1\left[1.5\frac{d}{a} - 0.5\left(\frac{d}{a}\right)^3\right]$ if $d \le a$; $\gamma(d) = C$ if $d > a$.

- Exponential model: $\gamma(d) = C_0 + C_1\left[1 - \exp\left(-3\,\frac{d}{a}\right)\right]$.

- Gaussian model: $\gamma(d) = C_0 + C_1\left[1 - \exp\left(-3\,\frac{d^2}{a^2}\right)\right]$.

- Hole effect model: $\gamma(d) = C_0 + C_1\left[1 - \frac{\sin(ad)}{ad}\right]$. An equivalent form is $\gamma(d) = C_0 + C_1\left[1 - \frac{a'\sin(d/a')}{d}\right]$ where $a' = 1/a$. $(C_0 + C_1)$ represents the value of γ towards which the dampening sine function tends to stabilize. This equation would adequately model a variogram of the periodic structures in Fig. 13.5a-b (variograms only differ from Geary's correlograms by the scale of the ordinate).

• Linear model: $\gamma(d) = C_0 + bd$ where b is the slope of the variogram model. A linear model with sill is obtained by adding the specification: $\gamma(d) = C$ if $d \geq a$.

• Pure nugget effect model: $\gamma(d) = C_0$ if $d > 0$; $\gamma(d) = 0$ if $d = 0$. The second part applies to a point estimate. In practice, observations have the size of the sampling grain (Section 13.0); the error at that scale is always larger than 0.

Other less-frequently encountered models are described in geostatistics textbooks. A model is usually chosen on the basis of the known or assumed process having generated the spatial structure. Several models may be added up to fit any particular sample variogram. Parameters are fitted by weighted least squares; the weights are function of the distance and the number of pairs in each distance class (Cressie, 1991).

Anisotropy

As mentioned at the beginning of Subsection 2, anisotropy is present in data when the autocorrelation function is not the same for all geographic directions considered (David, 1977; Isaaks & Srivastava, 1989). In *geometric anisotropy*, the variation to be expected between two sites distant by d in one direction is equivalent to the variation expected between two sites distant by $b \times d$ in another direction. The range of the variogram changes with direction while the sill remains constant. In a river for instance, the kind of variation expected in phytoplankton concentration between two sites 5 m apart across the current may be the same as the variation expected between two sites 50 m apart along the current even though the variation can be modelled by spherical variograms with the same sill in the two directions. Constant b is called the *anisotropy ratio* ($b = 50/5 = 10$ in the river example). This is equivalent to a change in distance units along one of the axes. The anisotropy ratio may be represented by an ellipse or a more complex figure on a map, its axes being proportional to the variation expected in each direction. In *zonal anisotropy*, the sill of the variogram changes with direction while the range remains constant. An extreme case is offered by a strip of land. If the long axis of the strip is oriented in the direction of a major environmental gradient, the variogram may correspond to a linear model (always increasing) or to a spherical model with a sill larger than the nugget effect, whereas the variogram in the direction perpendicular to it may show only random variation without spatial structure with a sill equal to the nugget effect.

Directional variogram and correlogram

Directional variograms and correlograms may be used to determine whether anisotropy (defined in Subsection 2) is present in the data; they may also be used to describe anisotropic surfaces or to account for anisotropy in kriging (Subsection 13.2.2). A direction of space is chosen (i.e. an angle θ, usually by reference to the geographic north) and a search is launched for the pairs of points that are within a given distance class d in that direction. There may be few such pairs perfectly aligned in the aiming direction, or none at all, especially when the observed sites are not regularly spaced on the map. More pairs can usually be found by looking within a small neighbourhood around the aiming line (Fig. 13.9). The neighbourhood is determined by an angular tolerance parameter φ and a parameter κ that sets the tolerance for distance classes along the aiming line. For each observed point $Ø_h$ in turn, one looks for other points $Ø_i$ that are at distance $d \pm \kappa$ from it. All points found

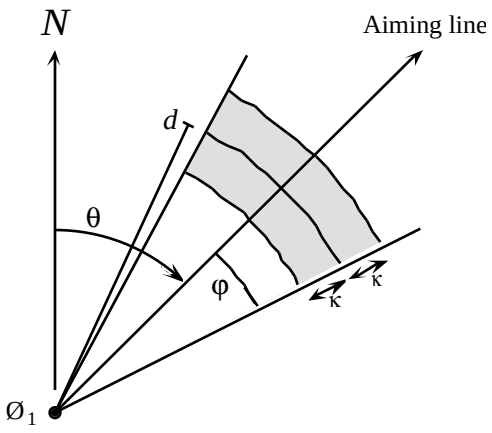


Figure 13.9 Search parameters for pairs of points in directional variograms and correlograms. From an observed study site $Ø_1$, an aiming line is drawn in the direction determined by angle θ (usually by reference to the geographic north, indicated by N). The angular tolerance parameter φ determines the search zone (grey) laterally whereas parameter κ sets the tolerance along the aiming line for each distance class d. Points within the search window (in gray) are included in the calculation of $I(d)$, $c(d)$ or $\gamma(d)$.

within the search window are paired with the reference point $Ø_h$ and included in the calculation of semi-variance or spatial autocorrelation coefficients for distance class d. In most applications, the search is bi-directional, meaning that one also looks for points within a search window located in the direction opposite (180°) the aiming direction. Isaaks & Srivastava (1989, Chapter 7) propose a way to assemble directional measures of semi-variance into a single table and to produce a contour map that describes the anisotropy in the data, if any; Rossi *et al.* (1992) have used the same approach for directional spatial correlograms.

Numerical example. An artificial data set was produced containing random autocorrelated data. The data were generated using the turning bands method (David, 1977; Journel & Huijbregts, 1978); random normal deviates were autocorrelated following a spherical model with a range of 5. Pure spatial autocorrelation, as described in the spatial structure model 1b of Subsection 1.1.1, generates continuity in the data (Fig. 13.10a). The variogram (without test of significance) and spatial correlograms (with tests) are presented in Figs. 13.10b-d. In this example, the data were standardized during data generation, prior to spatial autocorrelation analysis, so that the denominator of eq. 13.2 is 1; therefore, the variogram and Geary's correlogram are identical. The variogram suggests a spherical model with a range of 6 units and a small nugget effect (Fig. 13.10b).

Besides the description of spatial structures, variograms are used for several other purposes in spatial analysis. In Subsection 13.2.2, they will be the basis for interpolation by kriging. In addition, structure functions (variograms, spatial

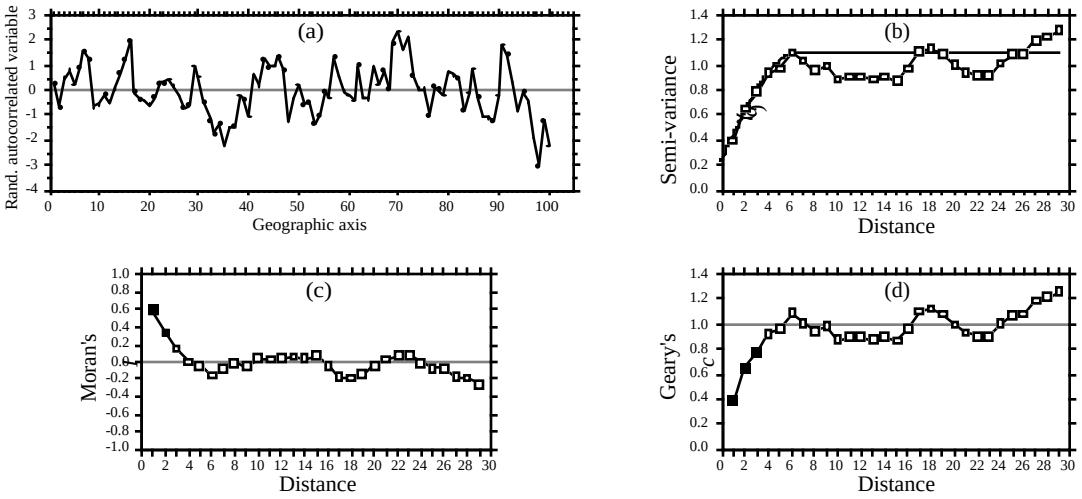


Figure 13.10 (a) Series of 100 equispaced random, spatially autocorrelated data. (b) Variogram, with spherical model superimposed (heavy line). Abscissa: distances between points along the geographic axis in (a). (c) and (d) Spatial correlograms. Dark squares: autocorrelation statistics that remain significant after progressive Bonferroni correction ($\alpha = 0.05$); white squares: non-significant values.

correlograms) may prove extremely useful to help determine the grain size of the sampling units and the sampling interval to be used in a survey, based upon the analysis of a pilot study. They may also be used to perform change-of-scale operations and predict the type of autocorrelation and variance that would be observed if the grain size of the sampling design was different from that actually used in a field study (Bellehumeur *et al.*, 1997).

4 — Spatial covariance, semi-variance, correlation, cross-correlation

This Subsection examines the relationships between spatial covariance, semi-variance and correlation (including cross-correlation), under the assumption of second-order stationarity, leading to the concept of cross-correlation. This assumption (Subsection 13.2.1) may be restated as follows:

- The first moment (mean of points *i*) of the variable exists:

$$\mathrm{E}\,[y_i] \;=\; \frac{1}{n}\sum_{i\,=\,1}^{n} y_i \;=\; m_i \qquad \textbf{(13.11)}$$

Its value does not depend on position in the study area.

- The second moment (covariance, numerator of eq. 13.1) of the variable exists:

$$C(d) = \left[\frac{1}{W(d)} \sum_{(h,i)\,|\,d_{hi} \approx d}^{W(d)} y_h y_i \right] - m_h m_i \qquad \textbf{(13.12)}$$

$$C(d) = \mathrm{E}[y_h y_i] - m^2 \quad \text{for } h, i \,|\, d_{hi} \approx d \qquad \textbf{(13.13)}$$

The value of $C(d)$ depends only on d and on the orientation of the distance vector, but not on position in the study area. To understand eq. 13.12 as a measure of covariance, imagine the elements of the various pairs y_h and y_i written in two columns as if they were two variables. The equation for the covariance (eq. 4.4) may be written as follows, using a final division by n instead of $(n-1)$ (maximum-likelihood estimate of the covariance, which is standard in geostatistics):

$$s_{y_h y_i} = \frac{\sum y_h y_i}{n} - \frac{\sum y_h \sum y_i}{n\ n} = \frac{\sum y_h y_i}{n} - m_h m_i$$

The overall variance (Var, with division by n instead of $n-1$) also exists since it is the covariance calculated for $d = 0$:

$$\mathrm{Var}[y_i] = \mathrm{E}[y_i - m_i]^2 = C(0) \qquad \textbf{(13.14)}$$

When computing the semi-variance, one only considers pairs of observations distant by d. Eqs. 13.9 and 13.10 are re-written as follows:

$$\gamma(d) = \frac{1}{2}\mathrm{E}[y_h - y_i]^2 \quad \text{for } h, i \,|\, d_{hi} \approx d \qquad \textbf{(13.15)}$$

A few lines of algebra obtain the following formula:

$$\gamma(d) = \frac{\sum y_i^2 - \sum y_h y_i}{W(d)} = C(0) - C(d) \quad \text{for } h, i \,|\, d_{hi} \approx d \qquad \textbf{(13.16)}$$

Two properties are used in the derivation: (1) $\sum y_h = \sum y_i$, and (2) the variance (Var, eq. 13.14) can be estimated using any subset of the observed values if the hypothesis of second-order stationarity is verified.

The correlation is the covariance divided by the product of the standard deviations (eq. 4.7). For a spatial process, the (auto)correlation is written as follows (leading to eq. 13.1):

$$r(d) = \frac{C(d)}{s_h s_i} = \frac{C(d)}{\mathrm{Var}[y_i]} = \frac{C(d)}{C(0)} \qquad \textbf{(13.17)}$$

Consider the formula for Geary's c (eq. 13.2), which is the semi-variance divided by the overall variance. The following derivation:

$$c(d) = \frac{\gamma(d)}{\mathrm{Var}[y_i]} = \frac{C(0) - C(d)}{C(0)} = 1 - \frac{C(d)}{C(0)} = 1 - r(d)$$

leads to the conclusion that Geary's c is one minus the coefficient of spatial (auto)correlation. In a graph, the semi-variance and Geary's c coefficient have exactly the same shape (e.g. Figs. 13.10b and d); only the ordinate scales may differ. An autocorrelogram plotted using $r(d)$ has the exact reverse shape as a Geary correlogram. An important conclusion is that the plots of semi-variance, covariance, Geary's c coefficient, and $r(d)$, are equivalent to characterize spatial structures under the hypothesis of second-order stationarity (Bellehumeur & Legendre, 1998).

Cross-covariances may also be computed from eq. 13.12, using values of *two different variables* observed at locations distant by d (Isaaks & Srivastava, 1989). Eq. 13.17 leads to a formula for cross-correlation which may be used to plot cross-correlograms; the construction of the correlation statistic is the same as for time series (eq. 12.10). With transect data, the result is similar to that of eq. 12.10. However, the programs designed to compute spatial cross-correlograms do not require the data to be equispaced, contrary to programs for time-series analysis. The theory is presented by Rossi *et al.* (1992), as well as applications to ecology.

Ecological application 13.1b

A survey was conducted on a homogeneous sandflat in the Manukau Harbour, New Zealand, to identify the scales at which spatial heterogeneity could be detected in the distribution of adult and juvenile bivalves (*Macomona liliana* and *Austrovenus stutchburyi*), as well as indications of adult-juvenile interactions within and between species. The results were reported by Hewitt *et al.* (1997); see also Ecological application 13.2. Sampling was conducted along transects established at three sites located within a 1-km^2 area; there were two transects at each site, forming a cross. Sediment cores (10 cm diam., 13 cm deep) were collected using a nested sampling design; the basic design was a series of cores 5 m apart, but additional cores were taken 1 m from each of the 5-m-distant cores. This design provided several comparison in the short distance classes (1, 4, 5, and 6 m). Using transects instead of rectangular areas allowed relatively large distances (150 m) to be studied, given the allowable sampling effort. Nested sampling designs have also been advocated by Fortin *et al.* (1989) and by Bellehumeur & Legendre (1998).

Spatial correlograms were used to identify scales of variation in bivalve concentrations. The Moran correlogram for juvenile *Austrovenus*, computed for the three transects perpendicular to the direction of tidal flow, displayed significant spatial autocorrelation at distances of 1 and 5 m (Fig. 13.11a). The same pattern was found in the transects parallel to tidal flow. Figure 13.11a also indicates that the range of influence of autocorrelation was about 15 m. This was confirmed by plotting bivalve concentrations along the transects: LOWESS smoothing of the graphs (Subsection 10.3.8) showed patches of about 25-30 m in diameter (Hewitt *et al.*, 1997, Figs. 3 and 4).

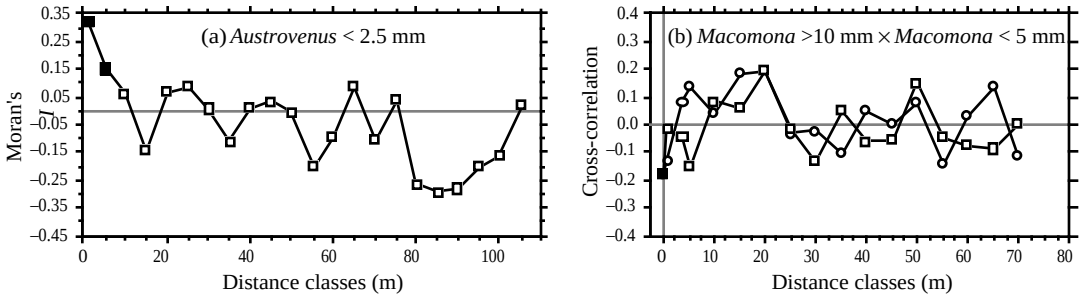


Figure 13.11 (a) Spatial autocorrelogram for juvenile *Austrovenus* densities. (b) Cross-correlogram for adult-juvenile *Macomona* interactions, folded about the ordinate: circles = positive lags, squares = negative lags. Dark symbols: correlation statistics that are significant after progressive Bonferroni correction ($\alpha = 0.05$). Redrawn from Hewitt *et al*. (1997).

Cross-correlograms were computed to detect signs of adult-juvenile interactions. In the comparison of adult (> 10 mm) to juvenile *Macomona* (< 5 mm), a significant negative cross-correlation was identified at 0 m in the direction parallel to tidal flow (Fig. 13.11b); correlation was not significant for the other distance classes. As in time series analysis, the cross-correlation function is not symmetrical; the correlation obtained by comparing values of $\mathbf{y}_1$ to values of $\mathbf{y}_2$ located at distance d on their right is not the same as when values of $\mathbf{y}_2$ are compared to values of $\mathbf{y}_1$ located at distance d on their right, except for $d = 0$. In Fig. 13.11b, the cross-correlogram is folded about the ordinate (compare to Fig. 12.9). Contrary to time series analysis, it is not useful in spatial analysis to discuss the direction of lag of a variable with respect to the other unless one has a specific hypothesis to test.

5 — Multivariate Mantel correlogram

Sokal (1986) and Oden & Sokal (1986) found an ingenious way to compute a correlogram for multivariate data, using the normalized Mantel statistic r_M and test of significance (Subsection 10.5.1). This method is useful, in particular, to describe the spatial structure of species assemblages.

The principle is to quantify the ecological relationships among sampling sites by means of a matrix $\mathbf{Y}$ of multivariate similarities or distances (using, for instance, coefficients S_{17} or D_{14} in the case of species abundance data), and compare $\mathbf{Y}$ to a *model matrix* $\mathbf{X}$ (Subsection 10.5.1) which is different for each geographic distance class (Fig. 13.12).

- For distance class 1 for instance, pairs of neighbouring stations (that belong to the first class of geographic distances) are coded 1, whereas the remainder of matrix $\mathbf{X}_1$ contains zeros. A first Mantel statistic (r_{M1}) is calculated between $\mathbf{Y}$ and $\mathbf{X}_1$.

- The process is repeated for the other distance classes d, building each time a model-matrix $\mathbf{X}_d$ and recomputing the normalized Mantel statistic. Matrix $\mathbf{X}_d$ may contain 1's

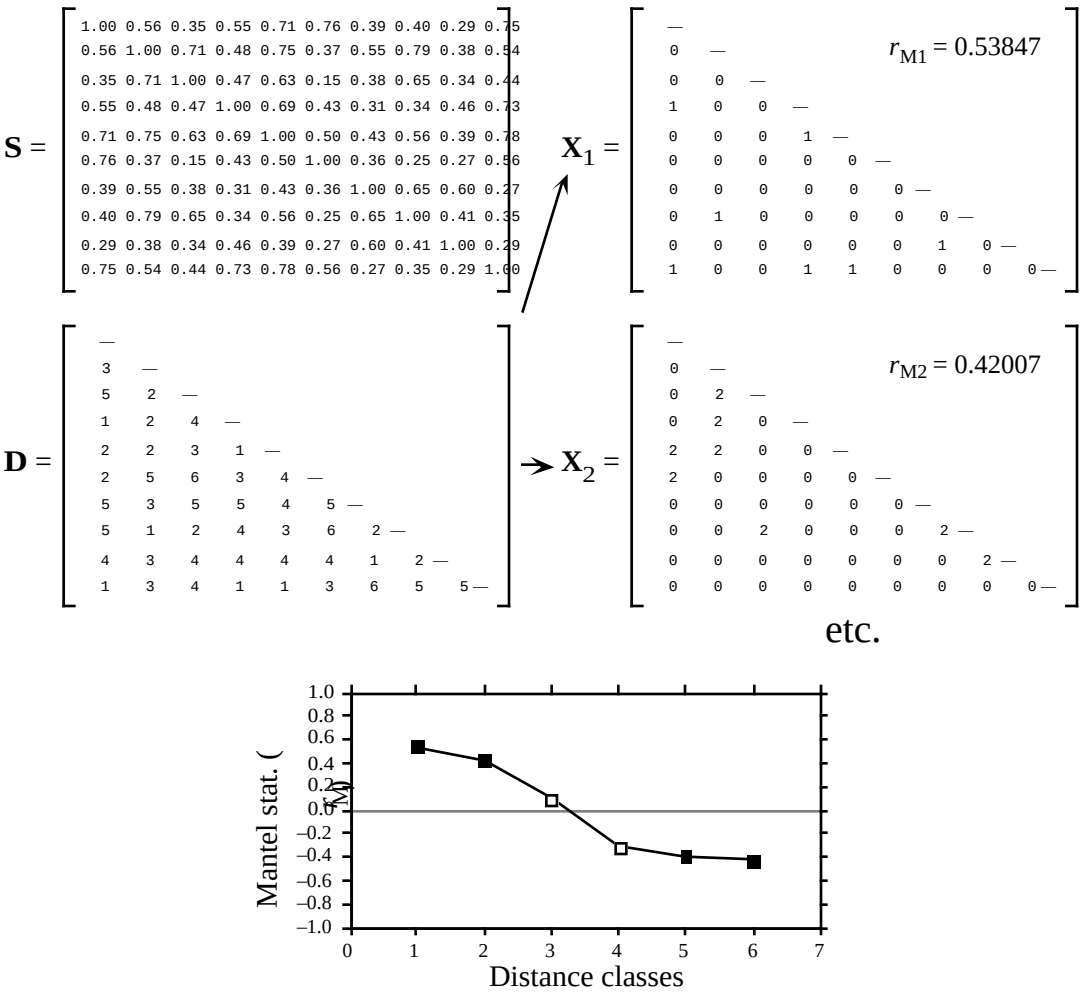


Figure 13.12 Construction of a Mantel correlogram for a similarity matrix **S** ($n = 10$ sites). The matrix of geographic distance classes **D**, from Fig. 13.4, gives rise to model matrices $\mathbf{X}_1$, $\mathbf{X}_2$, etc. for the various distance classes d. These are compared, in turn, to matrix $\mathbf{Y} = \mathbf{S}$ using standardized Mantel statistics (r_{Md}). Dark symbols in the correlogram: Mantel statistics that are significant after progressive Bonferroni correction ($\alpha = 0.05$).

for pairs that are in the given distance class, or the code value for that distance class (d), or any other value different from zero; all coding methods lead to the same value of the normalized Mantel statistic r_M.

The Mantel statistics, plotted against distance classes, produce a multivariate correlogram. Each value is tested for significance in the usual way, using either

permutations or Mantel's normal approximation (Box 10.2). Computation of standardized Mantel statistics assumes second-order stationarity. As in the case of univariate correlograms (above), one is advised to use some form of correction for multiple testing before interpreting Mantel correlograms.

Numerical example. Consider again the 10 sampling sites of Fig. 13.4. Assume that species assemblage data were available and produced similarity matrix **S** of Fig. 13.12. Matrix **S** played here the role of **Y** in the computation of Mantel statistics. Were the species data autocorrelated? Distance matrix **D**, already divided into 6 classes in Fig. 13.4, was recoded into a series of model matrices $\mathbf{X}_d$ (d = 1, 2, etc.). In each of these, the pairs of sites that were in the given distance class received the value d, whereas all other pairs received the value 0. Mantel statistics were computed between **S** and each of the $\mathbf{X}_d$ matrices in turn; positive and significant Mantel statistics indicate positive autocorrelation in the present case. The statistics were tested for significance using 999 permutations and plotted against distance classes d to form the Mantel correlogram. The progressive Bonferroni method was used to account for multiple testing because interest was primarily in detecting autocorrelation in the first distance classes.

Before computing the Mantel correlogram, one must assume that the condition of second-order stationarity is satisfied. This condition is more difficult to explain in the case of multivariate data; it means essentially that the surface is uniform in (multivariate) mean and variance at broad scale. The correlogram illustrated in Fig. 13.12 suggests the presence of a gradient. If the condition of second-order-stationarity is satisfied, this means that the gradient detected by this analysis is a part of a larger, autocorrelated spatial structure. This was called a "false gradient" in the numerical example of Subsection 2, above.

When **Y** is a similarity matrix and distance classes are coded as described above, positive Mantel statistics correspond to positive autocorrelation; this is the case in the numerical example. When the values in **Y** are distances instead of similarities, or if the 1's and 0's are interchanged in matrix **X**, the signs of all Mantel statistics are changed. One should always specify whether positive autocorrelation is expressed by positive or negative values of the Mantel statistics when presenting Mantel correlograms. The method was applied to vegetation data by Legendre & Fortin (1989).

13.2 Maps

The most basic step in spatial pattern analysis is the production of maps displaying the spatial distributions of values of the variable(s) of interest. Furthermore, maps are essential to help interpret spatial structure functions (Section 13.1).

Several methods are available in mapping programs. The final product of modern computer programs may be a contour map, a mesh map (such as Figs. 13.13b and 13.16b), a raised contour map, a shaded relief map, and so on. The present Section is not concerned with the graphic representation of maps but instead with the way the mapped values are obtained. Spatial interpolation methods have been reviewed by Lam (1983).

Geographic information systems (GIS) are widely used nowadays, especially by geographers, to manage complex data corresponding to points, lines and surfaces in space. The present Section is not an introduction to these complex systems. It only aims at presenting the most widespread methods for mapping univariate data (i.e. a single variable *y*). The spatial analysis of multivariate data (multivariate matrix **Y**) is deferred to Sections 13.3 to 13.5.

Beware of non-additive variables such as pH, logarithms of counts of organisms, diversity measures, and the like (Subsection 1.4.2). Maps of such variables, produced by trend-surface analysis or interpolation methods, should be interpreted with caution; the interpolated values only make sense by reference to sampling units of the same size as those used in the original sampling design. Block kriging (Subsection 2) for blocks representing surfaces or volumes that differ from the grain of the observed data simply does not make sense for non-additive variables.

1 — Trend-surface analysis

Trend-surface analysis is the oldest method for producing smoothed maps. In this method, estimates of the variable at given locations are not obtained by interpolation, as in the methods presented in Subsection 2, but through a regression model calibrated over the entire study area.

In 1914, Student proposed to express observed values as a polynomial function of time and mentioned that it could be done for spatial data as well. This is also one of the most powerful tools of spatial pattern analysis, and certainly the easiest to use. The objective is to express a response variable *y* as a nonlinear function of the geographic coordinates X and Y of the sampling sites where the variable was observed:

$$y = f(\mathrm{X},\ \mathrm{Y})$$

In most cases, a polynomial of X and Y with cross-product terms is used; trend-surface analysis is then an application of polynomial regression (Subsection 10.3.4) to spatially-distributed data. For example a relatively complex, but smooth surface might be fitted to a variable using a third-order polynomial with 10 parameters (b_0 to b_9):

$$\hat{y} = f(\mathrm{X}, \mathrm{Y}) = b_0 + b_1\mathrm{X} + b_2\mathrm{Y} + b_3\mathrm{X}^2 + b_4\mathrm{XY} + b_5\mathrm{Y}^2 + b_6\mathrm{X}^3 + b_7\mathrm{X}^2\mathrm{Y} + b_8\mathrm{XY}^2 + b_9\mathrm{Y}^3 \quad \textbf{(13.18)}$$

Note the distinction between the response variable *y*, which may represent a physical or biological variable, and the Cartesian geographic coordinate Y. Using polynomial regression, trend-surface analysis produces an equation which is linear in its parameters, although the response of *y* to the explanatory variables in matrix **X** = [X, Y] may be nonlinear. If variables *y*, X and Y have been centred on their respective means prior to model fitting, the model has an intercept of 0 by construct; therefore parameter b_0 does not have to be fitted and it can be removed from the model.

Numerical example. The data from Table 10.5 are used here to illustrate the method of trend-surface analysis. The dependent variable of the analysis, *y*, is Ma, which was the log-

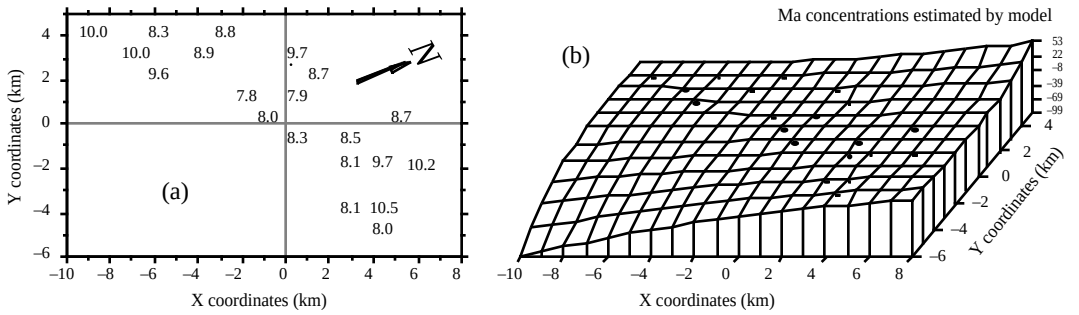


Figure 13.13 Variable Ma (log-transformed concentrations of aerobic heterotrophic bacteria growing on marine agar at salinity of 34 psu) at 20 sites in the Thau lagoon on 25 October 1988. (a) Map of the sampling sites with respect to arbitrary geographic coordinates X and Y. The observed values of Ma, from Table 10.5, are also shown. The *N* arrow points to the north. (b) Trend-surface map; the vertical axis gives the values of Ma estimated by the polynomial regression equation. Dots represent the sampling sites.

transformed (ln(x + 1)) concentrations of aerobic heterotrophic bacteria growing on marine agar at salinity of 34 psu. The explanatory variables are the X and Y geographic coordinates of the sampling sites (Fig. 13.13a). The steps of the calculations are the following:

• Centre the geographic coordinates on their respective means. The reason for centring X and Y is given in Subsection 10.3.4; the amount of variation explained by a trend-surface equation is not changed by a translation (centring) of the spatial coordinates across the map.

• Determine the order of the polynomial equation to be used. A first-degree regression equation of Ma as a function of the geographic coordinates X and Y alone would only represent the linear variation of Ma with respect to X and Y; in other words, a flat surface, possibly sloping with respect to X, Y, or both. With the present data, the first-degree regression equation was not significant (R^2 = 0.02), meaning that there was no significant linear geographic trend to be described in the data. A regression equation incorporating the second-degree monomials (X^2, XY and Y^2) together with X and Y would be appropriate to model a surface presenting a single large bump or trough. Again, this did not seem to be the case with the present data since the second-degree equation was not significant (R^2 = 0.39). An equation incorporating the third-degree, fourth-degree, etc. terms would be able to model structures of increasing complexity and refinement. The cost, however, is a loss of degrees of freedom for every new monomial in the equation; trend-surface analysis using high-order equations thus requires a large number of observed sampling sites. In the present example, the polynomial was limited to the third degree, for a total of 9 terms; this is a large number of terms, considering that the data only contained 20 sampling sites.

• Calculate the various terms of the third-degree polynomial, by combining the variables X and Y as follows: X^2, X×Y, Y^2, X^3, X^2×Y, X×Y^2, Y^3.

• Compute the multiple regression equation. The model obtained using all 9 regressors had $R^2 = 0.87$, but several of the partial regression coefficients were not significant.

• Remove nonsignificant terms. The linear terms may be important to express a linear gradient; the quadratic and cubic terms may be important to model more complex surfaces. Nonsignificant terms should not be left in the model, however, except when they are required for comparison purpose. Nonsignificant terms were removed one by one (backward elimination, Subsection 10.3.3) until all terms (monomials) in the polynomial equation were significant. The resulting trend-surface equation was highly significant ($R^2 = 0.81$, $p < 0.0001$):

$$\hat{y} = 8.13 - 0.16\ XY - 0.09\ Y^2 + 0.04\ X^2Y + 0.14\ XY^2 + 0.10\ Y^3$$

Remember, however, that tests of significance are too liberal with autocorrelated data, due to the non-independence of residuals (Subsection 1.1.1).

• Lay out a regular grid of points (X', Y') and, using the regression equation, compute forecasted values ($\hat{y}'$) for these points. Plot a map (Fig. 13.13b) using the file with (X', Y', and $\hat{y}'$). Values estimated by a trend-surface equation at the observed study sites do not coincide with the values observed at these sites; regression is not an exact interpolator, contrary to kriging (Subsection 2).

Different features could be displayed by rotating the Figure. The orientation chosen in Fig. 13.13b does not clearly show that the values along the long axis of the Thau lagoon are smaller near the centre than at the ends. It displays, however, the wavy structure of the data from the lower left-hand to the upper right-hand corner, which is roughly the south-to-north direction. The Figure also clearly indicates that one should refrain from interpreting extrapolated data values, i.e. values located outside the area that has actually been sampled. In the present example, the values forecasted by the model in the lower left-hand and the upper right-hand corners (–99 and +53, respectively) are meaningless for log bacterial concentrations. Within the area where real data are available, however, the trend-surface model provides a good visual representation of the broad-scale spatial variation of the response variable.

Examination of the residuals is essential to make sure that the model is not missing some salient feature of the data. If the trend-surface model has extracted all the spatially-structured variation of the data, given the scale of the study, residuals should look random when plotted on a map and a correlogram of residuals should be non-significant. With the present data, residuals were small and did not display any recognizable spatial pattern.

A cubic trend-surface model is often appropriate with ecological data. Consider an ecological phenomenon which starts at the mean value of the response variable *y* at the left-hand border of the sampled area, increases to a maximum, then goes down to a minimum, and comes back to the mean value at the right-hand border. The amount of space required for the phenomenon to complete a full cycle — whatever the shape it may take — is its extent (Section 13.0). Using trend-surface analysis, such a phenomenon would be correctly modelled by a third-degree trend surface equation. A polynomial equation is a more flexible mathematical model than sines or cosines, in that it does not require symmetry or strict periodicity.

The degree of the polynomial which is appropriate to model a phenomenon is predictable to a certain extent. If the extent is of the same order as the size of the study

area (say, in the X direction), the phenomenon will be correctly modelled by a polynomial of degree 3 which has two extreme values, a minimum and a maximum. If the extent is larger than the study area, a polynomial of degree less than 3 is sufficient; degree 2 if there is only one maximum, or one minimum, in the sampling window; and degree 1 if the study area is limited to the increasing, or decreasing, portion of the phenomenon. Conversely, if the scale of the phenomenon controlling the variable is smaller than the study area, more than two extreme values (minima and maxima) will be found, and a polynomial of order larger than 3 is required to model it correctly. The same reasoning applies to the X and Y directions when using a polynomial combining the X and Y geographic coordinates. So, using a polynomial of degree 3 acts as a filter: it is a way of looking for phenomena that are of the same extent, or larger, than the study area.

An assumption must be made when using the method of trend-surface analysis: that all observations form a single statistical population, subjected to one and the same generating process, and can consequently be modelled using a single polynomial equation of the geographic coordinates. Evidence to that effect may be available prior to the analysis. When this is not the case, the hypothesis of homogeneity may be supported by examining the regression residuals (Subsection 10.3.1). When there are indications that values in different regions of the geographic space obey different processes (e.g. different geology, action of currents or wind, or influence of other physical variables), the study area should be divided into regions, to be modelled by separate trend-surface equations.

Polynomial regression, used in the numerical example above, is a good first approach to fitting a model to a surface when the shape to be modelled is unknown, or known to be simple. In some instances, however, it may not provide a good fit to the data; trend-surface analysis must then be conducted using nonlinear regression (Subsection 10.3.6), which requires that an appropriate numerical model be provided to the estimation program. Consider the example of the effect of some human-generated environmental disturbance at a site, the indicator variable being the number of species. The response, in this case, is expected to be stronger near the impacted site, tapering off as one gets farther away from it. Assume that data were collected along a transect (a single geographic coordinate X) and that the impacted site is near the centre of the transect. A polynomial equation would not be appropriate to model an inverse-squared-distance diffusion process (Fig. 13.14a), whereas an equation of the form:

$$\hat{y} = b_0 + \frac{b_1 X^2}{b_2 X^2 + 1}$$

would provide a much better fit (Fig. 13.14b). The minimum of this equation is b_0; it is obtained when X = 0. The maximum, b_1/b_2, is reached asymptotically as X becomes large in either the positive or negative direction. For data collected in different directions around the impacted site, a nonlinear trend-surface equation with similar properties would be of the form:

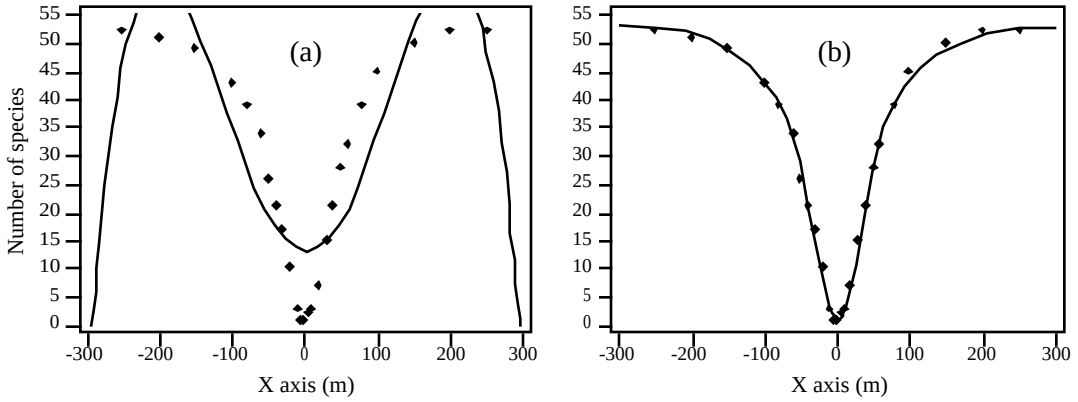


Figure 13.14 (a) Artificial data representing the number of species around the site of an environmental disturbance (located at X = 0) are not well-fitted by a 4th-order polynomial equation of the X coordinates ($R^2 = 0.7801$). (b) They are well-fitted by the following inverse-squared-distance diffusion equation: $\hat{y} = 1 + [0.0213X^2 / (0.0004X^2 + 1)]$ ($R^2 = 0.9975$).

$$\hat{y} = b_0 + \frac{b_1X^2 + b_2Y^2}{b_3X^2 + b_4Y^2 + 1}$$

where X and Y are the coordinates of the sites in geographic space.

Trend-surface analysis is appropriate for describing broad-scale spatial trends in data. It does not produce accurate fine-grained maps of the spatial variation of a variable, however. Other methods described in this Chapter may prove useful to model variation at finer scales. In some studies, the broad-scale trend itself is of interest; this is the case in the numerical example above and in Ecological application 13.2. In other instances, and especially in studies that cover large geographic expanses, the broad-scale trend may be already known and understood; students of geographic variation patterns may want to conduct analyses on detrended data, i.e. data from which the broad-scale trend has been removed. Detrending a variable may be achieved by computing the residuals from a trend-surface equation of sufficient order, as in time-series analysis (Section 12.2).

If there is replication at each point, it is possible to perform a test of goodness-of-fit of a trend-surface model (Draper and Smith, 1981; Legendre & McArdle, 1997). By comparing the observed error mean square after fitting the trend surface to the error mean square estimated by the among-replicate within-location variation, one can test if the model fits the data properly. The among-replicate within-location variation is computed from the deviations from the means at the various locations; it is actually the

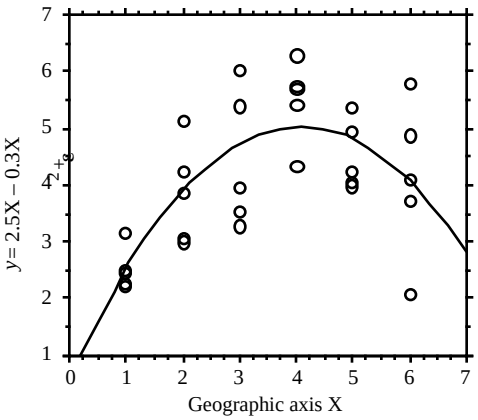


Figure 13.15 Artificial data representing sampling along a geographic axis X with 5 replicates at each site; $n = 30$. The F test of goodness-of-fit indicates that the trend-surface equation $\hat{y} = 0.562 + 2.184X - 0.267X^2$ ($R^2 = 0.4899$) fits the data properly.

residual mean square of an ANOVA among locations. These error mean squares are not much different if the trend surface goes through the expected values at the various locations, so that the F ratio of the two mean squares is not significant. If, on the contrary, the fitted surface does not follow the major features of the variation among locations, the deviations of the data from the fitted trend-surface values are likely to be larger than expected from our knowledge of the sampling error; the F statistic is then significantly larger than 1, indicating that the trend surface is misrepresenting the variation among locations.

Numerical example. Consider the artificial data in Fig. 13.15. Variable X represents a geographic axis along which sampling has taken place at 6 sites with replication. Variable y was constructed using equation $y = 2.5X - 0.3X^2 + \varepsilon$, where ε is a random standard normal deviate [N(0, 1)]. A quadratic trend-surface model of X was fitted to the data. The residual mean square, or "error mean square after fitting the trend surface", was $MS_1 = 0.84909$ ($\nu = 27$). An analysis of variance was conducted on y using the grouping into 6 sites as the classification criterion. The residual mean square obtained from the ANOVA was $MS_2 = 0.87199$ ($\nu = 24$). The ratio of these two mean squares gave an F statistic:

$$F = \frac{MS_1}{MS_2} = \frac{0.84909}{0.87199} = 0.97374$$

which was tested against $F_{\alpha=0.05\,(27,\ 24)} = 1.959$. The F statistic was not significantly different from 1 ($p = 0.5308$), which indicated that the model fitted the data properly.

The trend-surface analysis was recomputed using a linear model of X. The model obtained was $\hat{y} = 3.052 + 0.316X$ ($R^2 = 0.1941$). MS_1 in this case was 1.29358 ($\nu = 28$). The F ratio

$MS_1/MS_2 = 1.29358/0.87199 = 1.48348$. The reference value was $F_{0.05(28, 24)} = 1.952$. The probability associated with the F ratio, $p = 0.1651$, indicated that this model still fitted the data, which were constructed to contain a linear term (2.5X in the construction equation) as well as a quadratic trend (term $-0.3\,X^2$), but the fit was poorer than with the quadratic polynomial model which was capable of accounting for both the linear and quadratic trends.

This numerical example shows that trend-surface analysis may be applied to data collected along a transect; the "trend surface" is one-dimensional in that case. The numerical example at the end of Subsection 10.3.4 is another example of a trend-surface analysis of a dependent variable, salinity, with respect to a single geographic axis (Fig. 10.9). Trend-surface analysis may also be used to model data in three-dimensional geographic space (geographic coordinates X, Y and Z, where Z is either altitude or depth) or with one of the dimensions representing time. Section 13.5 will show how the analysis may be extended to a multivariate dependent data matrix **Y**.

Haining (1987) described alternative methods for estimating the parameters of a trend-surface model when the residuals are spatially autocorrelated; in that case, least-squares estimation of the parameters is inefficient and standard errors as well as tests of significance are biased. Haining's methods allow one to recognize three components of spatial variation corresponding to the site, local, and regional scales.

Ecological application 13.2

A survey was conducted at 200 locations within a fairly homogeneous 12.5 ha rectangular sandflat area in Manukau Harbour, New Zealand, to identify factors that control the spatial distributions of the two dominant bivalves, *Macomona liliana* Iredale and *Austrovenus stutchburyi* (Gray), and to look for evidence of adult-juvenile interactions within and between species. Results are reported in Legendre *et al.* (1997). Most of the broad-scale spatial structure detected in the bivalve counts (two species, several size classes) was explained by the physical and biological variables. Results of principal component analysis and spatial regression modelling suggested that different factors controlled the spatial distributions of adults and juveniles. Larger size classes of both species displayed significant spatial structures, with physical variables explaining some but not all of this variation; the spatial patterns of the two species differed, though. Smaller organisms were less strongly spatially structured; virtually all of their spatial structure was explained by physical variables.

Highly significant trend-surface equations were found for all bivalve species and size classes (log-transformed data), indicating that the spatial distributions of the organisms were not random, but highly organised at the scale of the study site. The trend-surface models for smaller animals had much smaller coefficients of determination (10-20%) than for larger animals (30-55%). The best models, i.e. the models with the highest coefficients of determination (R^2), were for the *Macomona* > 15 mm and *Austrovenus* > 10 mm. The coefficients of determination were consistently higher for *Austrovenus* than for *Macomona*, despite the fact that *Macomona* were usually far more numerous than *Austrovenus*. A map illustrating the trend-surface equation is presented for the largest *Macomona* size class (Fig. 13.16); the field counts are also given for comparison.

(a)

35 41 37 48 52 68 64 51 52 19 50 48 46 50 40 33 27 29 32 39
22 47 46 33 51 71 54 71 44 56 63 50 30 49 46 41 43 40 29 32
45 66 58 60 27 59 50 52 41 40 54 39 41 46 47 19 24 30 20 28
41 56 33 60 59 60 53 52 26 39 31 33 29 54 37 18 36 42 27 29
48 36 38 34 41 55 42 67 53 57 34 34 42 33 32 46 33 18 27 32
41 32 41 35 48 52 40 56 40 44 41 43 44 41 43 32 22 34 30 32
17 29 50 67 35 52 28 44 45 28 30 35 30 28 35 37 20 28 50 31
27 43 38 44 29 44 35 39 44 38 40 24 37 43 35 31 38 48 50 37
27 42 36 36 29 46 39 38 29 37 38 35 35 38 25 38 39 54 50 44
52 41 32 41 42 28 47 22 24 17 37 31 38 29 41 38 36 46 38 52

(b)

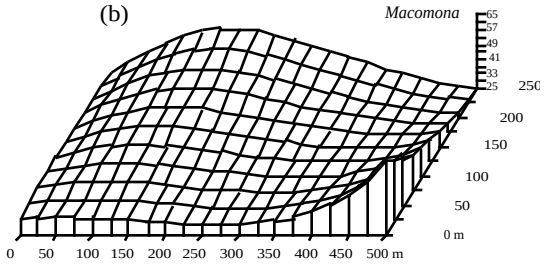


Figure 13.16 *Macomona* > 15 mm at 200 sites in Manukau Harbour, New Zealand, on 22 January 1994. (a) Actual counts at sampling sites in 200 regular grid cells; in the field, sites were not perfectly equispaced. (b) Map of the trend-surface equation explaining 32% of the spatial variation in the data. The values estimated from the trend-surface equation (log-transformed data) were back-transformed to raw counts before plotting. Modified from Legendre *et al*. (1997).

2 — Interpolated maps

In this family of methods, the value of the variable at a point location on a map is estimated by local interpolation, using only the observations available in the vicinity of the point of interest. In this respect, interpolation mapping differs from trend surface analysis (Subsection 1), where estimates of the variable at given locations were not obtained by interpolation, as in the present Subsection, but through a statistical model calibrated over the entire study area. Fig. 13.17 illustrates the principle of interpolation mapping. A regular grid of nodes (Fig. 13.17c) is defined over the area that contains the study sites $Ø_i$ (Fig. 13.17a, b). Interpolation assigns a value to each point of that grid. This is the single most important step in mapping. Following that, results may be represented in the form of contours (Fig. 13.17d) with or without colours or shades, or three-dimensional constructs such as Fig. 13.16b.

Assigning a value to each grid node may be done in different ways. Different interpolation methods may produce maps that look different; this is also the case when using different parameters with a same method (e.g. different exponents in inverse-distance weighting).

The most simple rule would be to give, to each node of the grid, the value of the observation which is the closest to it. The end result is a division of the map into Voronoï polygons (Subsection 13.3.1) displaying a “zone of influence” drawn around each observation. Another simple solution consists in dividing the map into Delaunay triangles (Subsection 13.3.1). There is an observed value y_i at each site $Ø_i$. A triangular portion of plane, adjusted to the points $Ø_i$ that form the apices (corners) of a Delaunay triangle, provides interpolated values for all points lying within the triangle. Maps obtained using these solutions are shown in Chapter 11 of Isaaks & Srivastava (1989).

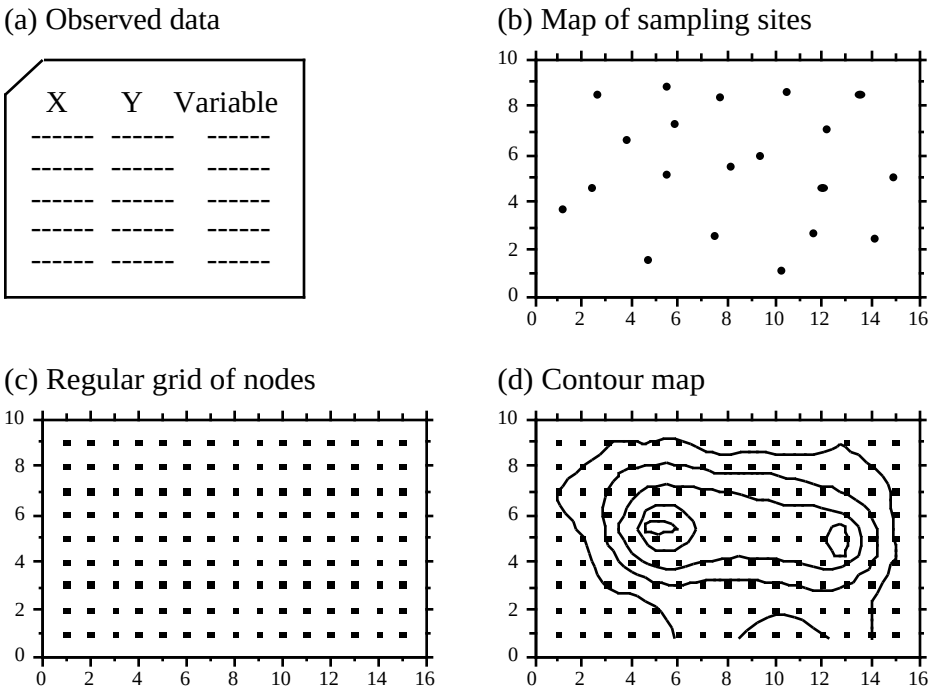


Figure 13.17 Summary of the interpolation procedure.

Alternatively, one may draw a "search circle" (or an ellipsoid for anisotropic data) around each grid node (Fig. 13.18). The radius of the circle may be determined in either of two ways. (1) One may fix a minimum number of observed points that must be included in the interpolation for each grid node; or (2) one may use the "distance of influence of the process" found by correlogram or variogram analysis (Section 13.1). The estimation procedure is repeated for each node of the grid. Several methods of interpolation may be used.

- Mean — Consider all the observed study sites found within the circle; assign the mean of these values to the grid node. This method does not produce smooth maps; discontinuities in neighbouring grid node values occur as observed points move in or out of the search circle.

- Inverse-distance weighting — Consider the observation sites found within the circle and calculate a weighted mean value, using the formula:

$$\hat{y}_{\text{Node}} = \sum_i w_i y_i \qquad \textbf{(13.19)}$$

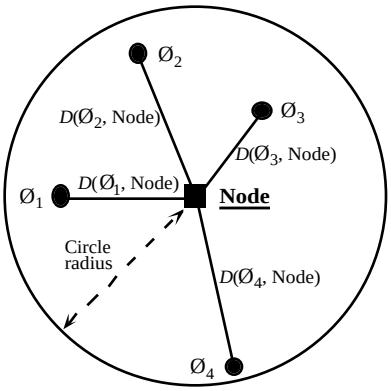


Figure 13.18 To estimate the value at a grid node (square), draw a search circle around it and consider the observed points ($Ø_i$) found within the circle. Observed points are separated from the node by distances $D(Ø_i, \text{Node})$.

where y_i is the value observed at point $Ø_i$ and weight w_i is the inverse of the distance (D) from point $Ø_i$ to the grid node to be estimated. The inverse distances, to some power k, are scaled by the sum of the weights for all points $Ø_i$ in the estimation, so as to produce values that are consistent with the values observed at points $Ø_i$ (unbiasedness condition):

$$w_i = \left(\frac{1}{D(Ø_i, \text{Node})^k} \right) \Big/ \sum_i \frac{1}{D(Ø_i, \text{Node})^k} \qquad \textbf{(13.20)}$$

A commonly-used exponent is $k = 2$. This corresponds, for instance, to the decrease in energy of waves dispersing across a two-dimensional surface. The greater the value of k, the less influence distant data points have on the value assigned to the grid node. This method produces smooth values over the grid of nodes. The range of estimated values is smaller than the range of observed data so that, contrary to trend-surface analysis (Fig. 13.13b), inverse-distance weighting does not produce meaningless values in the parts of the map beyond the area that was actually sampled. When the observation sites $Ø_i$ do not form a regular or nearly regular grid, however, this interpolation method may generate features in maps that have little to do with reality. As a consequence, inverse-distance weighting is not recommended in that situation.

• Weighted polynomial fitting — In this method, a trend-surface equation (Subsection 13.2.1) is adjusted to the observed data points within the search circle, weighting each observation $Ø_i$ by the inverse of its distance (using some appropriate power k) to the grid node to be estimated. A first or second-order polynomial equation is usually used. z_{Node} is taken to be the value estimated by the polynomial equation for the coordinates of the grid node. This method suffers from the same problem as inverse distance weighting with respect to observation sites $Ø_i$ that do not form a regular or nearly regular grid of points.

Kriging

• Kriging — This is the mapping tool in the toolbox of geostatisticians. The method was named by Matheron after the South African geostatistician D. G. Krige, who was the first to develop formal solutions to the problem of estimating ore reserves from sampling (core) data (Krige, 1952, 1966). Geostatistics was developed by Matheron (1962, 1965, 1970, 1971, 1973) and co-workers at the *Centre de morphologie mathématique* of the *École des Mines de Paris*. Geostatistics comprises the estimation of variograms (Subsection 13.1.6), kriging, validation methods for kriging estimates, and simulations methods for geographically distributed ("regionalized") data. Major textbooks have been written by former students of Matheron: David (1977) and Journel & Huijbregts (1978). Other useful references are Clark (1979), Rendu (1981), Verly *et al.* (1984), Armstrong (1989), Isaaks & Srivastava (1989), and Cressie (1991). Applications to environmental sciences and ecology have been discussed by Gilbert & Simpson (1985), Robertson (1987), Armstrong *et al.* (1989), Legendre & Fortin (1989), Soares *et al.* (1992), and Rossi *et al.* (1992). Geostatistical methods can be implemented using the software library of Deutsch & Journel (1992).

As in inverse-distance weighting (eq. 13.19), the estimated value for any grid node is computed as:

$$\hat{y}_{\text{Node}} = \sum_i w_i y_i$$

The chief difference with inverse-distance weighting is that, in kriging, the weights w_i applied to the points $Ø_i$ used in the estimation are not standardized inverses of the distances to some power k. Instead, the weights are based upon the covariances (semi-variances, eq. 13.9 and 13.10) read on a variogram model (Subsection 13.1.6). They are found by linear estimation, using the equation:

$$\mathbf{C} \cdot \mathbf{w} = \mathbf{d}$$

$$\begin{bmatrix} c_{11} & \cdots & c_{1n} & 1 \\ \cdot & \cdots & \cdot & 1 \\ \cdot & \cdots & \cdot & 1 \\ c_{n1} & \cdots & c_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ \cdot \\ \cdot \\ w_n \\ \lambda \end{bmatrix} = \begin{bmatrix} d_1 \\ \cdot \\ \cdot \\ d_n \\ 1 \end{bmatrix} \qquad \textbf{(13.21)}$$

where **C** is the covariance matrix among the n points $Ø_i$ used in the estimation, i.e. the semi-variances corresponding to the distances separating the various pair of points, as read on the variogram model; **w** is the vector of weights to be estimated (with the constraint that the sum of weights must be 1); and **d** is a vector containing the covariances between the various points $Ø_i$ and the grid node to be estimated. This is where a variogram model becomes essential; it provides the weighting function for the entire map and is used to construct matrix **C** and vector **d** for each grid node to be estimated. λ is a Lagrange parameter (as in Section 4.4) introduced to minimize the

variance of the estimates under the constraint $\Sigma w_i = 1$ (unbiasedness condition). The solution to this linear system is obtained by matrix inversion (Section 2.8):

$$\mathbf{w} = \mathbf{C}^{-1}\,\mathbf{d} \qquad \textbf{(13.22)}$$

Vector **d** plays a role similar to the weights in inverse-distance weighting since the covariances in vector **d** decrease with distance. Using covariances, the weights are statistical in nature instead of geometrical.

Kriging takes into account the grouping of observed points $Ø_i$ on the map. When two points $Ø_i$ are close to each other, the value of the corresponding coefficient c_{ij} in matrix **C** is high; this contributes to lowering their respective weights w_i. In this way, the redundancy of information introduced by dense groups of sampling sites is taken into account.

When anisotropy is present, kriging can use two, four, or more variogram models computed for different geographic directions and combine their estimates when calculating the covariances in matrix **C** and vector **d**. In the same way, when estimation is performed for sampling sites in a volume, a separate variogram can be used to describe the vertical spatial variation. Kriging is the best interpolation method for data that are not on a regular grid or display anisotropy. The price to pay is increased mathematical complexity during interpolation.

Among the interpolation methods, kriging is the only one that provides a measure of the error variance for each value estimated at a grid node. For each grid node, the error variance, called *ordinary kriging variance* (s^2_{OK}), is calculated as follows (Isaaks & Srivastava, 1989), using vectors **w** and **d** from eq. 13.21:

$$s^2_{OK} = \mathrm{Var}\,[y_i] - \mathbf{w}'\mathbf{d} \qquad \textbf{(13.23)}$$

where $\mathrm{Var}[y_i]$ is the maximum-likelihood estimate of the variance of the observed values y_i (eq. 13.14). Equation 13.23 shows that s^2_{OK} only depends on the variogram model and the local density of points, and not on the values observed at points $Ø_i$. The ordinary kriging variance may be used to construct confidence intervals around the grid node estimates at some significance level α, using eq. 13.4. It may also be mapped directly. Regions of the map with large values s^2_{OK} indicate that more observations should be made because sampling intensity was too low.

Kriging, as described above, provides point estimates at grid nodes. Each estimate actually applies to a "point" whose size is the same as the grain of the observed data. The geostatistical literature also describes how *block kriging* may be used to obtain estimates for blocks (i.e. surfaces or volumes) of various sizes. Blocks may be small, or cover the whole map if one wishes to estimate a resource over a whole area. As mentioned in the introductory remarks of the present Section, additive variables only may be used in block kriging. Block kriging programs always assume that the variable is *intensive*, e.g. the concentration of organisms (Subsection 1.4.2). For *extensive*

variables, such as the number of individual trees, one must multiply the block estimate by the ratio (block size / grain size of the original data).

3 — Measures of fit

Different measures of fit may be used to determine how well an interpolated map represents the observed data. With most methods, some measure may be constructed of the closeness of the estimated (i.e. interpolated) values $\hat{y}_i$ to the values y_i observed at sites $Ø_i$. Four easy-to-use measures are:

- The mean absolute error: $MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$
- The mean squared error: $MSE = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$
- The Euclidean distance: $D_1 = \sqrt{\sum_i (y_i - \hat{y}_i)^2}$
- The correlation coefficient (r) between values y_i and $\hat{y}_i$ (eq. 4.7). In the case of a trend-surface model, the square of this correlation coefficient is the coefficient of determination of the model.

In the case of kriging, the above measures of fit cannot be used because the estimated and observed values are equal, at all observed sites $Ø_i$. The technique of cross-validation may be used instead (Isaaks & Srivastava, 1989, Chapter 15). One observation, say $Ø_1$, is removed from the data set and its value is estimated using the remaining points $Ø_2$ to $Ø_n$. The procedure is repeated for $Ø_2$, $Ø_3$, …, $Ø_n$. One of the measures of fit described above may be used to measure the closeness of the estimated to the observed values. If replicated observations are available at each sampling site (a situation which does not often occur), the test of goodness-of-fit described in Subsection 1 can be used with all interpolation methods.

13.3 Patches and boundaries

Multivariate data may be condensed into spatially-constrained clusters. These may be displayed on maps, using different colours or shades. The present Section explains how clustering algorithms can be constrained to produce groups of spatially contiguous sites; study of the boundaries between homogeneous zones is also discussed. Prior to clustering, one must state unambiguously which sites are neighbours in space; the most common solutions to this are presented in Subsection 1.

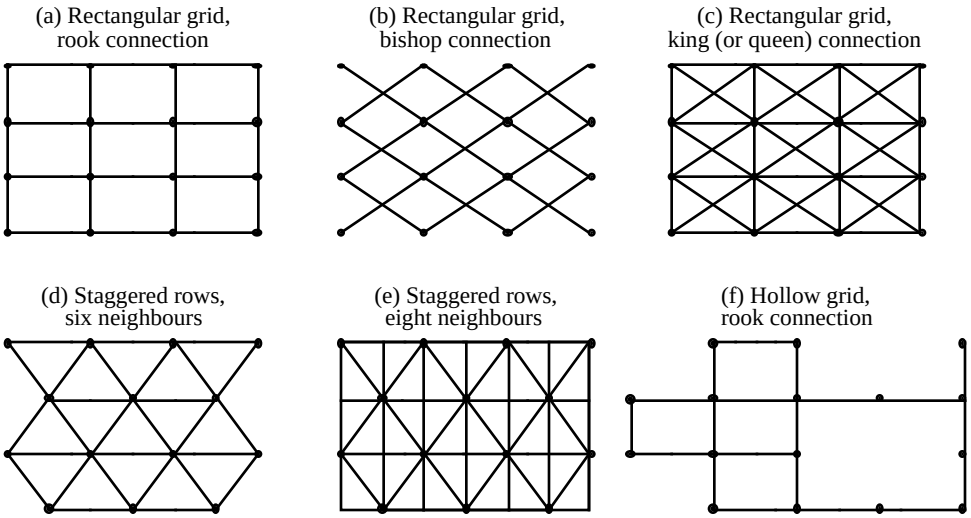


Figure 13.19 Connecting schemes for regular grids of points. See text.

1 — Connection networks

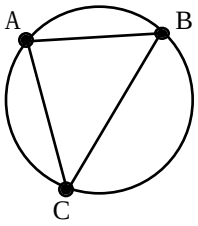
When sampling has been conducted on a regular rectangular grid, neighbouring points may be linked using simple connecting schemes whose names are derived from the game of chess (Cliff & Ord, 1981): rook's (rectangular: Fig. 13.19a), bishop's (diagonal: Fig. 13.19b), or king's connections (also called queen's: both rectangular and diagonal, Fig. 13.19c). Sampling in staggered rows leads to connecting each point (except borders) to six (Fig. 13.19d) or eight neighbours (Fig. 13.19e). Algorithms may allow the construction of regular grids with missing points (Fig. 13.19f). When the objects represent irregularly-shaped land units covering a geographic area (e.g. electoral units), parcels sharing a common boundary are regarded as contiguous.

When the localities are positioned in an irregular manner, geometric connecting schemes may be used, such as Delaunay triangulation, Gabriel graph, relative neighbourhood graph or minimum spanning tree. There exists an inclusion relationship among the four connecting schemes: all edges that are members of a minimum spanning tree also obey the relative neighbourhood graph criterion; these are all members of a Gabriel graph, which in turn are all included in a Delaunay triangulation (Toussaint, 1980; Matula & Sokal, 1980; Gordon, 1996c):

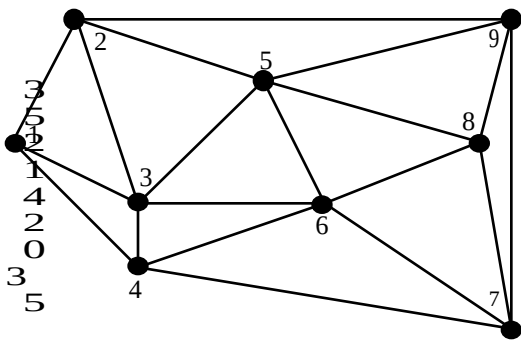
$$\text{Minimum spanning tree} \subseteq \text{Relative neighbourhood gr.} \subseteq \text{Gabriel gr.} \subseteq \text{Delaunay triangulation}$$

Delaunay triangulation

• Delaunay triangulation — The Delaunay triangulation criterion (Dirichlet, 1850; Upton & Fingleton, 1985) is illustrated in Fig. 13.20. For any triplet of points A, B and C, the three edges (i.e. lines) connecting these points are included in the triangulation



Point identifiers	Coordinates X	Y
1		0
2		1
3		2
4		2
5		4
6		5
7		8
8		7.5
9		8



19 edges form the Delaunay triangulation:

1–2	1–3	1–4	2–3	2–5	2–9
3–5	3–6	4–6	4–7	5–6	5–8
6–7	6–8	7–8	7–9	8–9	

Figure 13.20 Construction of a Delaunay triangulation.

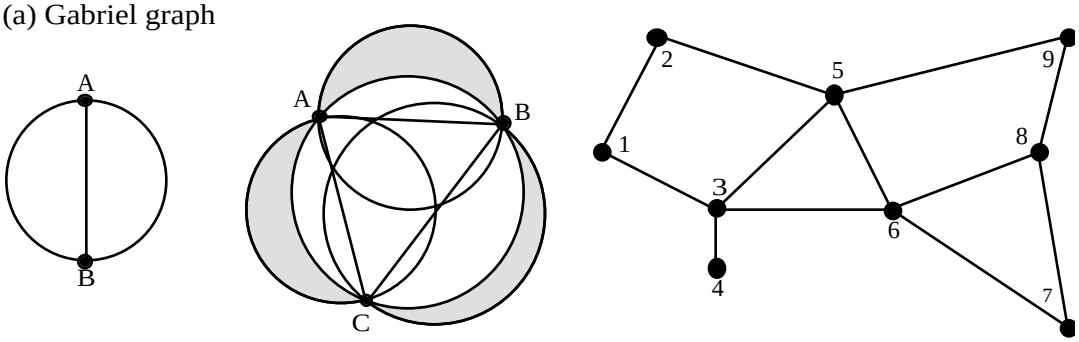
if and only if the circumscribed circle (i.e. the circle passing through the three points; on the left in the Figure) includes no other point. For example, the file of coordinates shown in the central part of the Figure gives rise to the triangulation on the right. The triangulation is fully described by a list of pairs of points corresponding to its edges; this is how the information can be passed on to a computer program for constrained clustering (Subsection 2).

Long edges may be created at the outskirts of a set of points, simply because there is no other point located farther away in the sampling design; this is called a *border effect*. For example, edges 2–9 and 7–9 might have been removed from the triangulation in Fig. 13.20 by the presence of other points in the circumscribed circles of triangles (2, 5, 9) and (7, 8, 9) had the sampling extent been broader. Long peripheral edges may be removed by hand or by the computer algorithm.

Gabriel graph

• Gabriel graph — The Gabriel graph criterion (Gabriel & Sokal, 1969) differs from that of the Delaunay triangulation (Fig. 13.21a). Draw a line between two points A and B. This line is part of the Gabriel graph if and only if no other point C lies inside the circle *whose diameter is that line.* In other words, the edge between A and B is part of the Gabriel graph if $D^2(A, B) < D^2(A, C) + D^2(B, C)$ for all other points C in the study, where $D^2(A, B)$ is the square of the geographic distance between points A and B. Another way of expressing this criterion is the following: if CENTRE represents the middle point between A and B, the edge connecting A to B is part of the Gabriel graph if $D(A, B)/2 < D(\text{CENTRE}, C)$ for any other point C in the study.

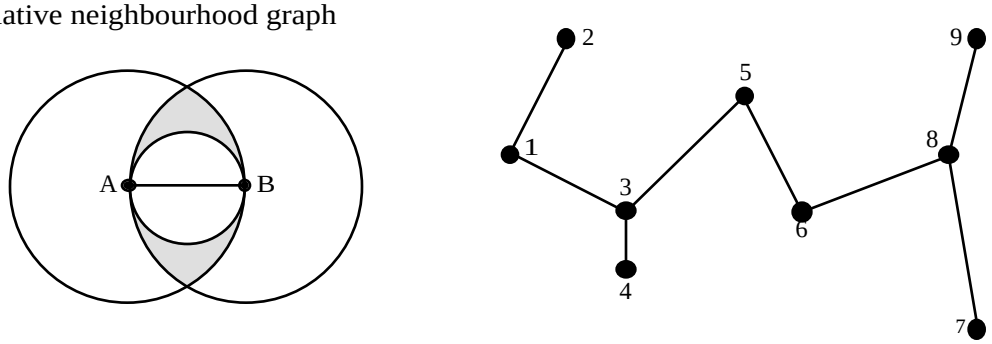
(a) Gabriel graph



12 edges form the Gabriel graph:

1–2 1–3 2–5 3–4
5–6 5–9 6–7 6–8

(b) Relative neighbourhood graph



8 edges form the relative neighbourhood graph

1–2 1–3 3–4 3–5 5–6
7–8 8–9

Figure 13.21 (a) Left: geometric criterion for constructing a Gabriel graph. Centre: the zone of exclusion of the Gabriel criterion, here for three points (grey zones + white inner circle), is larger than that of the Delaunay criterion (white inner circle). Right: Gabriel graph for the nine points of Fig. 13.20. (b) Left: geometric criterion for constructing a relative neighbourhood graph. The zone of exclusion of the relative neighbourhood criterion, here for two points (grey zones + white inner circle), is larger than that of the Gabriel criterion (white inner circle). Right: relative neighbourhood graph for the nine points of Fig. 13.20.

The Gabriel graph in Fig. 13.21a is constructed for the same points as the Delaunay triangulation in Fig. 13.20. The 12 edges forming the Gabriel graph are a subset of the 19 edges of the Delaunay triangulation. Indeed, as shown by the sketch in the centre of the Figure, the exclusion zone formed by the three circles corresponding to the Gabriel

criterion (which have for diameters the edges A–B, B–C and A–C) may contain, in the shadowed areas outside the Delaunay circle (white inner circle), some points that the Delaunay criterion circle does not exclude. This is why some edges that are authorized by the Delaunay criterion are excluded from the Gabriel graph.

Relative neighbourhood graph

• Relative neighbourhood graph — The relative neighbourhood criterion is as follows (Toussaint, 1980; Fig. 13.21b). Draw a line between two points A and B. Draw a first circle centred over A and a second one centred over B, each one having the line from A to B as its radius. This line is part of the graph if no other point C in the study lies inside *the intersection of the two circles*. In other words, the edge from A to B is part of the relative neighbourhood graph if and only if $D(A, B) \leq \max[D(A, C), D(B, C)]$ for all other points C in the study. For points forming an equilateral triangle, for instance, the three edges are included in the relative neighbourhood graph.

The relative neighbourhood graph in Fig. 13.21b is constructed for the same set of points as in Figs. 13.20 and 13.21a. The number of edges in a relative neighbourhood graph is $(n - 1)$. The 8 edges forming the relative neighbourhood graph are a subset of the 12 edges of the Gabriel graph. Indeed, as shown by the sketch on the left of the Figure, the exclusion zone at the intersection of the two circles corresponding to the relative neighbourhood criterion (which have for radius the edge A–B) may contain, in the shadowed zone outside the Gabriel circle (white inner circle), some points that the Gabriel criterion circle does not exclude. This is why some edges authorized by the Gabriel criterion are excluded from the relative neighbourhood graph.

Minimum spanning tree

• Minimum spanning tree — In this tree, which connects all points of a study, the sum of the edge lengths is minimum. Its construction is described at the end of Section 8.2; one way of obtaining it is to list the edges forming the primary connections of a single-linkage dendrogram. For points forming an equilateral triangle, for example, only two of the edges are included in the minimum spanning tree, whereas the three edges are included in a relative neighbourhood graph; the choice of the edge to leave out is arbitrary. The edges of a minimum spanning tree are either the same as, or a subset of, the edges of a relative neighbourhood graph of the same points. For the example data set, the edges that form the minimum spanning tree are the same as those of the relative neighbourhood graph of Fig. 13.21b.

Another approach is to select a distance threshold and connect all points that are within that distance of each other. One possible criterion to choose the distance threshold is to make it equal to the range of a variogram model (Fig. 13.7).

The list of connecting edges (Figs. 13.20 and 13.21) may be written out to a file. The file may be modified to take into account other information that researchers may have about the study area. For example, one may wish to eliminate edges that do not make sense in terms of gene flow because they cross unsuitable areas (e.g. a sea or a mountain range, in the case of terrestrial mammals). Or, one may wish to add connections that are potentially of interest although they do not imply first neighbours; for example, plants or animals may cross water bodies (lake, sea) and settle in non-

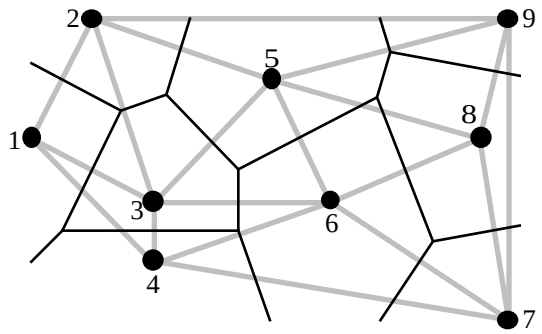


Figure 13.22 Delaunay triangulation (grey lines) and influence polygons (black lines) for the nine points of Fig. 13.20.

contiguous sites, which should nevertheless be considered contiguous because there is a direct path between them. Users of constrained clustering methods should not hesitate to modify lists of connections obtained from geometric criteria such as described above, to make the list of edges a better description of potential flow among sites, given the problem under study.

Influence polygon

It may be interesting to determine the geometric zone of influence of each point on a map. The zone of influence of a point A includes all the other points of the surface that are closer to A than to any other point in the study. The zones of influence so defined have the shape of polygons, also called tiles, tessellae, or tesserae (singular: tessella or tessera). The resulting picture is called a mosaic or tessellation (adjective: tessellated); it may be referred to as a Dirichlet tessellation (1850), Voronoï polygons (1909), or Thiessen polygons (1911), from the names of the authors who first described these mathematical structures.

Polygons are easily constructed from a Delaunay triangulation (Fig. 13.22). Draw the perpendicular bisector of each segment in the triangulation; the crossing points of the bisectors delimit the polygons (tiles). Computer algorithms may be used to calculate the surface area of each polygon, at least those that are closed; peripheral tiles may be open. Upton & Fingleton (1985) and Isaaks & Srivastava (1989) propose various applications of tessellations to spatial analysis.

2 — Constrained clustering

The delineation of clusters of contiguous objects has been discussed in Section 12.6 for time series and spatial transects. The method of chronological clustering, in particular, was described in Subsection 12.6.4; it proceeds by imposing to a clustering algorithm a constraint of contiguity along the time series. Constraints of contiguity have been applied to spatial clustering by several authors, including Lefkovitch (1978,

1980), Monestiez (1978), Lebart (1978), Roche (1978), Perruchet (1981) and Legendre & Legendre (1984c). In the present Subsection, it is generalized to two- or three-dimensional spatial data and to spatio-temporal data.

Constrained clustering differs from its unconstrained counterpart in the following way.

- Unconstrained clustering methods (Chapter 8) only use the information in the similarity or distance matrix computed among the objects. In hierarchical methods, a local criterion is optimized at each step; in all methods included in the Lance and Williams general model, for instance, the objects or groups clustered at each step are those with the largest fusion similarity or the smallest fusion distance. In partitioning methods, a global criterion is optimized; in *K*-means, for instance, the algorithm looks for *K* groups that feature the smallest sum of within-group sums-of-squares E_K.

- Constrained clustering methods take into account more information than the unconstrained approaches. In the case of spatial or temporal contiguity, the only admissible clusters are those that obey the contiguity relationship. Spatial contiguity may be described by one of the connecting schemes of Subsection 1. The criterion to be optimized during clustering is relaxed to give priority to the constraint of spatial contiguity. It is no surprise, then, that a constrained solution may be less optimal than its unconstrained counterpart in terms of the clustering criterion, e.g. E_K. This is balanced by the fact that the solution is likely to more readily interpretable.

It is fairly easy to modify clustering algorithms to incorporate a constraint of spatial contiguity (Fig. 13.23). As an example, consider the clustering methods included in the Lance and Williams general clustering model (Subsection 8.5.9). At the beginning of the clustering process, the vector of group membership has each object as a different group. Proceed as follows:

1. Compute a similarity matrix among objects, using the non-geographic information.

2. Choose a connecting scheme (Subsection 1) and produce a list of connection edges as in Figs. 13.20 and 13.21. Read in the file of edges and transform it into a *contiguity matrix* containing 1's for connected sites and 0's elsewhere.

3. Compute the Hadamard product of these two matrices. The Hadamard product of two matrices is the product element by element. The resulting matrix contains similarity values in the cells where the contiguity matrix contained 1's, and 0's elsewhere.

4. The largest similarity value in the matrix resulting from step 3 determines the next pair of objects or groups (*h* and *i*) to be clustered. Modify the vector of group membership (right of the Figure), giving the same group label to all members of former groups *h* and *i*.

5. Update the similarity matrix using eq. 8.11.

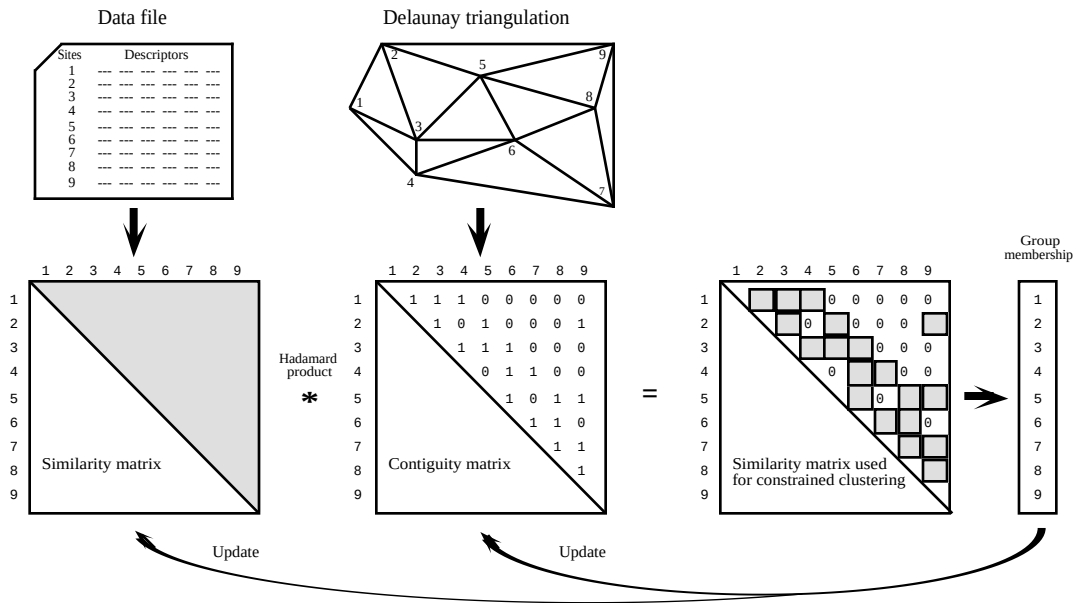


Figure 13.23 Summary of the spatially-constrained clustering procedure for methods included in the Lance and Williams general clustering model. The vector of group membership is represented at the start of the clustering iterations; see text. Locations of the points are the same as in Fig. 13.20.

6. Update also the contiguity matrix. All objects that were neighbours to *h* are now also neighbours to *i* and vice versa.

7. Go back to step 3. Iterate until all objects are members of a single group.

Ferligoj & Batagelj (1982) have shown, however, that the introduction of relational constraints (e.g. spatial contiguity) may occasionally produce reversals with any of the hierarchical clustering methods included in the Lance & Williams algorithm (Subsection 8.5.9], except complete linkage. Additional constraints may be added to the algorithm, for example to limit the size or composition of any group (Gordon, 1996c). *K*-means partitioning algorithms (Section 8.8) may also be constrained by the contiguity matrix shown in Fig. 13.23.

Spatially constrained clustering is useful in a variety of situations. Here are some examples.

- In many studies, there are compelling reasons to force the clusters to be composed of contiguous sites; for instance, when delineating ecological regions, political voting units, or resource distribution networks.

• One may wish to relate the results of clustering to geographically-located potential causal factors that are known to be spatially autocorrelated, e.g. geological data.

• One may wish to cluster sites based upon physical variables, using a constraint of spatial contiguity, in order to design a stratified biological sampling program to study community composition.

• To test the hypothesis that neighbouring sites are ecologically similar, one may compare unconstrained and constrained clustering solutions using the modified Rand index (Subsection 8.11.2). De Soete *et al.* (1987) give other examples where such comparisons may help test hypotheses in the fields of molecular evolution, psycholinguistics, cognitive psychology, and evolution of languages.

• Constrained solutions are less variable than unconstrained clustering results, which may differ in major ways among clustering methods. Indeed, the constraint of spatial contiguity reduces the number of possible solutions and forces different clustering algorithms to converge onto largely similar clusters (Legendre *et al.*, 1985).

Constrained clustering may also be used for three-dimensional or spatio-temporal sampling designs (e.g. Planes *et al.*, 1993). As long as the three-dimensional or spatio-temporal contiguity of the observations can be accurately described as a file of edges as in Figs. 13.20 and 13.21, constrained clustering programs have no difficulty in computing the solution; the only difficulty is the representation of the results as three-dimensional or spatio-temporal maps. Higher-dimensional extensions of the geometric connecting schemes presented in Subsection 1 are available if required.

Legendre (1987b) suggested a way of introducing spatial proximity into clustering algorithms which is less stringent than the methods described above. The method consists in weighting the values in the ecological similarity or distance matrix by some function of the geographic distances among points, before clustering. The idea was taken up by Bourgault *et al.* (1992) who proposed to use a multivariate variogram or covariogram as spatial weighting function prior to clustering. Large ecological distances between sites that are close in space are downweighted to some extent by this procedure. It is then easier for clustering algorithms to incorporate somewhat diverging sites into neighbourhood clusters. Oliver & Webster (1989) suggested to use a univariate variogram for the same purpose.

Constrained classification methods have recently been reviewed by Gordon (1996c). Formal aspects have been discussed by Ferligoj & Batagelj (1982, 1983). Algorithms have been surveyed by Murtagh (1985). Generalized forms of constrained clustering have been described by De Soete *et al.* (1987).

Numerical example. An artificial set of 16 sites was constructed to represent staggered-row sampling of a distribution with two peaks. From the geographic positions of the sites, a Delaunay triangulation (35 edges) was computed (Fig. 13.24a). A single variable was attributed to the sites. For three groups, the unconstrained *K*-means solution has a sum of within-group sums-of-squares $E_K = 53$ (Fig. 13.24b). The constrained *K*-means solution, for three groups,

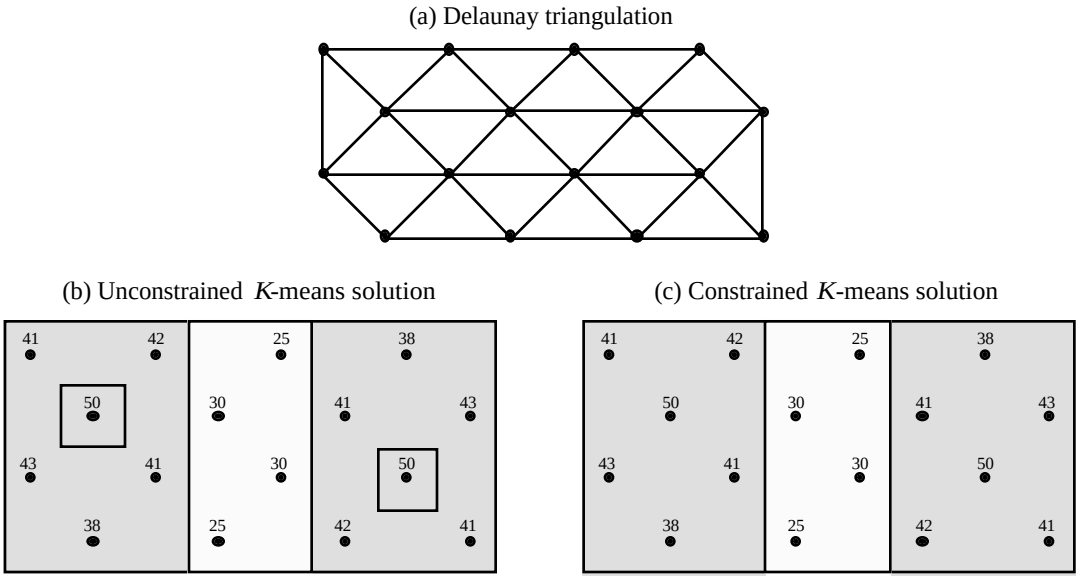


Figure 13.24 Numerical example showing the difference between the unconstrained (b) and constrained (c) clustering solutions. (a) Delaunay triangulation with 35 edges; they were used as constraint in (c). The values of the artificial variable are given in panels (b) and (c); the three groups obtained by unconstrained and constrained *K*-means are identified by shadings.

has a value E_K = 188 (Fig. 13.24c) which is higher than that of the unconstrained solution, for reasons explained above. The two partitions are interesting in different ways. The unconstrained solution identifies sites with similar values, whereas the constrained solution brings out the two peaks plus a region of lower values forming a valley between the peaks.

Spatially constrained clustering has been applied to a variety of ecological situations. Some applications to two-dimensional map data are: Legendre & Legendre (1984c), Legendre & Fortin (1989), Legendre *et al.* (1989), and Lapointe & Legendre (1994). Applications to transect data and paleoecology (stratigraphic data) have been listed in Subsection 12.6.4.

3 — Ecological boundaries

Detection of boundaries is a complementary problem to the detection of homogeneous regions of space. Boundaries appear on maps as a by-product of constrained clustering, for instance. Most methods of clustering delineate groups even in gradient situations; a boundary between groups does not have to correspond to a sharp discontinuity in the data. In any case, boundaries detected by clustering may be partly interpolated. Other