Chapter 13 *Spatial analysis*

13.0 Spatial patterns

Experiment
Gradient
Patch

The analysis of spatial patterns is of prime interest to ecologists because most ecological phenomena investigated by sampling geographic space are structured by forces that have spatial components. Spatial patterns are studied through surveys (called *mensurative experiments* by Hurlbert, 1984), whereas underlying processes can be studied by *manipulative experiments* (Subsection 10.2.3). Ecological processes may give rise to response data displaying recognizable spatial patterns, which may be the subject of spatial analysis. Most ecological patterns may be described as either patches (such as tree groves, phytoplankton patches, and animal herds) or gradients. The latter may be linear or not.

Ecologists investigate the spatial patterns of species assemblages in order to understand the mechanisms that control species distributions. Patchiness is found at all spatial scales — from micrometres to continental and ocean-wide scales. Displaying the spatial variation of an ecological variable in the form of a map shows whether the structure is smoothly continuous or marked by sharp discontinuities. Most field studies cover only a part of any variable's spatial structure. So, gradients or patches displayed by maps may only be interpreted with respect to the scale of the sampling programme, which should be compared to the scale of the phenomenon under study when it is known.

Historical events

It is now understood that species distributions result from the combined action of several forces, some of which are external whereas others are intrinsic to the community. According to the environmental control model (Whittaker, 1956; Bray & Curtis, 1957; Hutchinson, 1957), environmental characteristics are the external forces that control species distributions. The internal forces relate to population dynamics or to top-down or bottom-up biotic interactions within the community (Lindeman, 1942; Southwood, 1987). Both types of forces generate spatial patterns within species or communities. Historical events (Sousa, 1979; Pickett & White, 1985; Reynolds, 1987) are other possible sources of spatial patterns; examples are given in Subsection 14.1.4. The mechanisms that create spatial structures and, hence, spatial correlation in the data, have been discussed in Section 1.1.

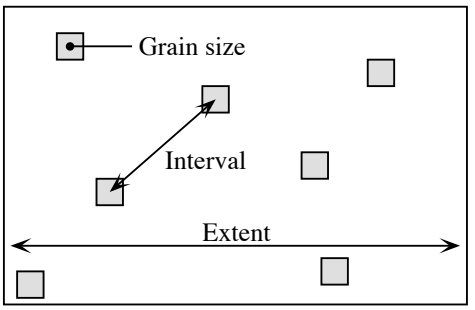


Figure 13.1 Components of a sampling design are the grain size, sampling interval, and extent. In the figure, the sampling units are represented by squares.

The present chapter is not a tutorial discussing all possible questions concerning the analysis of spatial structures. Its scope is more modest; it will describe some methods that allow the investigation of some of the questions of interest in spatial pattern analysis. A fundamental question will be left unanswered in this chapter: that of designing efficient sampling programmes for studying and analysing spatial patterns. The theory of spatial ecological sampling has to be re-written to provide meaningful answers to this question.

Scale

Scale is a key concept in both sampling design and the analysis of spatial (or temporal) patterns. It includes several spatial (or temporal) characteristics of random variables. Definition of these properties, which follows, depends on context.

In sampling theory (Fig. 13.1), spatial scale encompasses three elements of the sampling design (Wiens, 1989; Allen & Hoekstra, 1991; He *et al*., 1994; Dungan *et al*., 2002):

Grain size

- *Grain size* is the size of the elementary sampling units. It may be expressed as the diameter, surface or volume of matter supporting the measurements. In time series analysis, it is the duration over which measurements are integrated. The *resolution* of a study (Schneider, 1994) is equal to the grain size of its sampling design.

Sampling interval

- *Sampling interval* is the average distance between neighbouring sampling units. It is called *lag* in time series analysis (Section 12.0). For fixed extent, the sampling interval is a function of n, the number of sampling units. In turn, n is determined by the total effort that can be allocated to sampling.

Extent

- *Extent* is the total length, area or volume included in the study, or the total duration of the time series. It was called *range* by Schneider (1994) who also defined the *scope* as the ratio of the extent to the grain size. Since extent and grain size are expressed in the same units, scope is a dimensionless variable (Section 3.1).

It may happen that the data consist of contiguous sampling areas that completely cover the extent, instead of small sampling units distant from one another. This may occur in a variety of circumstances where a map is divided into contiguous "picture

Pixel

cells" or *pixels*. These include satellite data, video analysis of a transect, completely inventoried forest plots, and modelling. The linear measurement of grain size is equal to the sampling interval in such a case. Time series may also be entirely studied.

The spatial scale of patterns or processes is described as follows:

Ecological neighbourhood
Unit object
Unit process

- How large is a unit object, or how much space is disturbed by a unit process? This amount of space, which is equivalent to grain size, is called the ecological neighbourhood (Addicott *et al*., 1987) or the area of resolution of individuals (Wiens, 1989). *Unit objects* may be individual plants or animals, bacterial colonies, etc. Examples of measurable structures resulting from *unit processes* are: the neighbourhood occupied by a territorial animal, the width of the wetland zone along a stream or of a tidal sand flat, the size of the patch of soil modified by the root system of a plant, and the size of phytoplankton patches which result from the combined action of primary production and diffusion (see Ecological applications 3.2d and 3.3a).

- What is the average distance between unit objects or processes? This distance is equivalent to the sampling interval.

- Over how much space does this type of object, or this process, occur? This amount of space is equivalent to the extent. For some processes, the extent may be an ocean or the whole planet.

The same notions may be applied to temporally occurring patterns or processes. While they are readily applicable to patterns that concern the distribution of objects, they may sometimes be applied as well to processes.

Sampling design

The scale of the sampling design should follow from what is known (e.g. from a pilot study) about the scale of the pattern or process, and from the ecological question being addressed (Dungan *et al*., 2002). A well-focused question generally reduces the difficulty of choosing the type (simple random, systematic, stratified, etc.) as well as the scale components (grain, interval, extent) of the sampling design.

Sampling grain

- The sampling grain should be larger than a unit object (e.g. an individual organism) and the same as, or preferably smaller than, the structures resulting from a unit process (e.g. a patch), which should be detected by the sampling design.

Sampling interval

- The sampling interval should be smaller than the average distance between the structures resulting from a unit process to be detected by the sampling design.

Sampling extent

- The sampling extent may, in some cases, be the same as the total area covered by the type of objects or by the process under study. In other cases, it is limited to a smaller area, determined by the total allowable effort (n) and the maximum interval that one wishes to maintain between adjacent sampling units. For constant n, the sampling extent can be maximized by turning the sampling area into a transect (see Ecological application 13.1c).

The extent and grain define the observation window in spatial pattern analysis. No structure can be detected that is smaller than the grain or larger than the extent of a study. Wiens (1989) compares them to the overall size and mesh size of a sieve, respectively.

In quantitative ecology, the term "scale" is generally used in a sense opposite to that of cartography. For cartographers, the scale is the ratio between the linear size of an object on a map and its size in nature, so that a small-scale map (e.g. 1:100000) is less detailed than a large-scale map (e.g. 1:25000). For ecologists, scale generally refers to the unit of measurement, e.g. the kilometre sampling scale is bigger than the centimetre scale and weekly observations are broader-scaled than hourly observations.

Broad scale
Fine scale

Confusion is avoided by using "broad scale" for phenomena with large extents and "fine scale" for those with small extents (Wiens, 1989)*. In any case, these terms only have comparative values.

In many instances, not one but several scales may be pertinent for the study of a pattern or process. Different processes are often at work, depending on the scale, to determine spatial patterns. As a consequence, conclusions derived for a spatial scale often cannot be extrapolated to other scales. The scale chosen for any particular study may be considered as a variable-sized window through which one observes nature; see the notion of observational window in Section 12.0. He *et al*. (1994) have shown how species diversity, for example, changes as a function of different components of scale (grain size, sampling interval, and extent). The methods described in Section 13.1, in particular, allow researchers to depict how spatial correlation changes as a function of the sampling interval.

Scale is an important reference to help understand the difference between environmental management problems and the answers that may be found in ecological studies. Most studies are conducted at scales (extents) finer than those of natural or anthropogenic disturbances (Fig. 13.2). As a consequence, environmental problems usually involve scales broader than the information available from field studies — surveys or field experiments. Scaling up from studies to environmental problems is a challenge that ecologists are often facing. New concepts and modelling tools must be developed to do so (Thrush *et al*., 1997). Multiscale spatial analysis of the results of surveys conducted across several spatial scales is one means towards that end.

Heterogeneity

An important concept is that of *heterogeneity* (Kolasa & Rollo, 1991; Dutilleul & Legendre, 1993; Dutilleul, 2011). With reference to spatial patterns, heterogeneity is the opposite of *homogeneity* which means the absence of variation. In everyday's language, heterogeneous means "composed of unlike elements or parts". Pitard (1992) distinguishes *constitution heterogeneity*, which is a property of the objects under study, from *distribution heterogeneity* which can be altered by mixing. In spatial pattern

* Unfortunately these two terms are not antonymic. *Broad* scale refers to the extent; its antonym is *narrow*. *Fine* scale refers to the grain; its antonym is *coarse*.

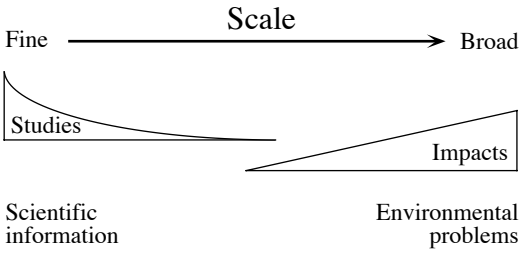


Figure 13.2 Scale differences between environmental management problems and the answers that may be found in ecological studies. From S. F. Thrush (pers. comm.).

analysis, heterogeneous refers to spatial variation in the measurements, in some general sense that applies to quantitative, semiquantitative, or qualitative variables (Subsection 1.4.1). The concept of heterogeneity may also be applied to the time dimension, considering repeated observations made at a single point in space. Heterogeneity can be measured in a univariate (e.g. the variance of a singe variable) or a multivariate way (e.g. the trace of a dispersion matrix). It can be decomposed into orthogonal components (as in PCA, Section 9.1) or with respect to spatial or temporal distance classes (e.g. correlograms for spatial survey data, Section 13.1, or for time series, Section 12.3). Kolasa & Rollo (1991) recognize that "measured heterogeneity", which reflects the observer's perspective, may be inadequate in that it may differ from "functional heterogeneity", which is the heterogeneity that influences the organisms. Functional heterogeneity may not be the same for different groups of organisms because the processes that are important for different groups may act at different temporal or spatial scales. In the sea, for instance, the doubling time of organisms is of the order of 1 day for phytoplankton, 10 to 40 days for zooplankton, 100 to 900 days for fish, and 120 to 500 days for mussels. Spatially, the horizontal scales that characterize patches are of the order of 0.1 to 1 km for phytoplankton and zooplankton, and 1 to 100 km for fish (Legendre *et al*., 1986). Measured heterogeneity converges towards functional heterogeneity as our knowledge of a system increases and, with it, our ability to use our measures to characterize important properties of the system (Kolasa & Rollo, 1991; Dutilleul & Legendre, 1993).

Point pattern

The analysis of spatial ecological patterns comprises two families of methods. *Point pattern analysis* is concerned with the distribution through space of individual objects — for instance individual plants or animals. Its chief purpose is to determine whether the geographic distribution of data points is random or not and to describe the type of pattern, in order to infer what kind of process may have generated it. In this family of methods, the quadrat-density and nearest-neighbour approaches have been widely used in vegetation science (Galiano, 1982; Carpenter & Chaney, 1983). Point pattern analysis will not be discussed further in this chapter. It has been authoritatively reviewed by a number of authors, including Pielou (1977), Cicéri *et al*. (1977), Getis

& Boots (1978), Ripley (1981, 1987), and Upton & Fingleton (1985), and by Dale (1999) and Fortin & Dale (2005) for ecological data.

Regionalized variable
Surface
Surface pattern

Values of a variable observed over a delimited geographic area form a *regionalized variable* (Matheron, 1965), also called a *surface* (Oden *et al*., 1993; Legendre & McArdle, 1997), if the sites where the variable has been observed may be viewed as a sample from an underlying continuous surface. The second family of methods to analyse spatial ecological patterns, called *surface pattern analysis*, deals with the study of spatially continuous phenomena. The spatial distributions of the variables are known, as usual, through sampling (or measurements on aerial photos or satellite maps) at discrete sampling sites. One or several variables are observed or measured at the observation sites, each site representing its surrounding portion of the geographic space. The analysis of continuous surfaces, where pixels cover the whole map (including data obtained by echolocation or remote sensing), is not specifically discussed here.

Surface pattern analysis includes a large number of methods developed to answer a variety of questions (Table 13.1). Several of these methods are discussed in the present chapter. General references are: Cliff & Ord (1981), Ripley (1981), Upton & Fingleton (1985, 1989), Griffith (1987), Legendre & Fortin (1989), and Rossi *et al*. (1992). The geostatistical literature is briefly reviewed in Subsection 13.2.2. The comparison of surfaces, i.e. univariate measures over the same area repeated at two or more sampling times, has been discussed by Legendre & McArdle (1997).

Multiscale modelling of univariate or multivariate ecological data can be done using spatial eigenfunction analysis, described in Chapter 14. Wavelet analysis, described in Subsection 12.5.4 for ecological data series, offers another method of multiscale modelling for *univariate spatial data on a regular grid*, e.g. remotely sensed data or entirely inventoried map areas.

The book of Fortin & Dale (2005) describes point pattern as well as surface pattern methods of spatial analysis. Section 13.6 provides a list of computer programs and functions for surface pattern analysis; most of these methods are not available in the major statistical packages.

Line pattern

Geographers have also developed *line pattern analysis* which is a topological approach to the study of networks of connections among points. Examples are: roads, electric or telephone lines, and river networks.

For a point pattern, heterogeneity refers to the distribution of individuals across space; one often compares the observed density variation of organisms to that expected for randomly distributed objects. For a surface pattern, heterogeneity refers to the variability of quantitative or qualitative descriptors across space. Dutilleul & Legendre (1993) provide a summary of the main statistical tools available to ecologists to quantify spatial heterogeneity in both the point and the surface pattern cases. Dutilleul (1993) describes in more detail how experimental designs can be accommodated to the

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 correlation (Descriptions using structure functions should always be complemented by maps.)	Univariate structure functions: correlogram, variogram, etc. (Section 13.1) Multivariate structure functions: multivariate variogram, Mantel correlogram (Section 13.1) Testing for a gradient in multivariate data: canonical ordination between the multivariate response data and the geographic coordinates of the sites (Section 13.4).
2) Mapping; estimation of values at given locations	Univariate data: local interpolation map, kriging; trend-surface map (global statistical model) (Section 13.2) Multivariate data: space-constrained clustering, search for boundaries (Section 13.3); interpolated map of the 1st (2nd, etc.) ordination axis (Section 13.4); multivariate trend-surface map obtained by canonical analysis (Section 13.4)
3) Modelling species-environment relationships while taking spatial structures into account	Spatial modelling through canonical analysis (Section 13.5); Multiscale analysis: spatial eigenfunction modelling (Chapter 14).
4) Performing valid statistical tests on autocorrelated data	Subsections 1.1.2 and 14.5.3.

spatial heterogeneity found in nature; spatial heterogeneity may be a nuisance for the experimenter, or a characteristic of interest. The analysis of spatial patterns is the study of organized arrangements of [ecological] heterogeneity across space.

These concepts, and more, are discussed in the book of Dutilleul (2011) on spatio-temporal heterogeneity. The book describes the study designs (field surveys and experiments) as well as the methods of analysis to interpret point and surface patterns in data collected to answer questions about the spatial, temporal, and spatio-temporal heterogeneity of ecosystems.

13.1 Structure functions

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

- either support the null hypothesis that no significant spatial correlation is present in a data set, or that none remains after detrending (Subsection 13.2.1) or after 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 correlation is present in the data, in order to use it in conceptual or statistical models.

Tests of spatial correlation coefficients may only support or reject the null hypothesis of the absence of significant spatial structure. When a significant spatial structure is found, it may correspond to induced spatial dependence (Subsection 1.1.1, model 1) or true spatial autocorrelation (model 2).

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 that description is usually supported by maps of the univariate or multivariate data (Sections 13.2 to 13.4). The most commonly used spatial structure functions are correlograms, variograms, and periodograms.

Spatial correlogram

A *correlogram* is a graph in which spatial correlation 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, Subsection 13.1.1, or the spatial correlation function, Subsection 13.1.5) or for multivariate data (multivariate variogram, Subsection 13.1.4, and Mantel correlogram, Subsection 13.1.6). In all cases, a test of significance is available for each individual spatial correlation 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 13.1.3). 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 applicable to regular grids of points; it is described Priestley (1964), Ripley (1981), Renshaw and Ford (1984) and Legendre & Fortin (1989). It is not discussed further in this book. Spatial eigenfunction analysis, described in Chapter 14, carries out a similar form of analysis and is more general since it can be used on irregularly-spaced points.

1 — Spatial correlograms

For quantitative variables (univariate data), spatial correlation can be estimated by Moran's *I* (1950) or Geary's *c* (1954) spatial correlation 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*:
$$c(d) = \frac{\frac{1}{2W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - y_i)^2}{\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{for } h \neq i \qquad \textbf{(13.2)}$$

y_h and y_i are the values of the observed variable at sites h and i, and d is the distance class considered in the calculation. Before computing spatial correlation coefficients, a matrix of geographic distances $\mathbf{D} = [D_{hi}]$ among observation sites must be calculated. Statistical details about these coefficients are available in Cliff & Ord (1981) and d'Aubigny (2006).

In the presence of explanatory variables generating spatial structure in the variable of interest, true spatial autocorrelation must be estimated on the *residuals* of a model that takes these explanatory variables into account. This is in agreement with the definition of spatial autocorrelation (Section 1.1), which is the spatial dependence *among the error components* of the observed data (eq. 1.2).

* These statistics are often called spatial *autocorrelation* coefficients. This terminology is misleading since the coefficients measure any type of spatial structure, be it due to induced spatial dependence (eq. 1.1) or true spatial autocorrelation (eq. 1.2).

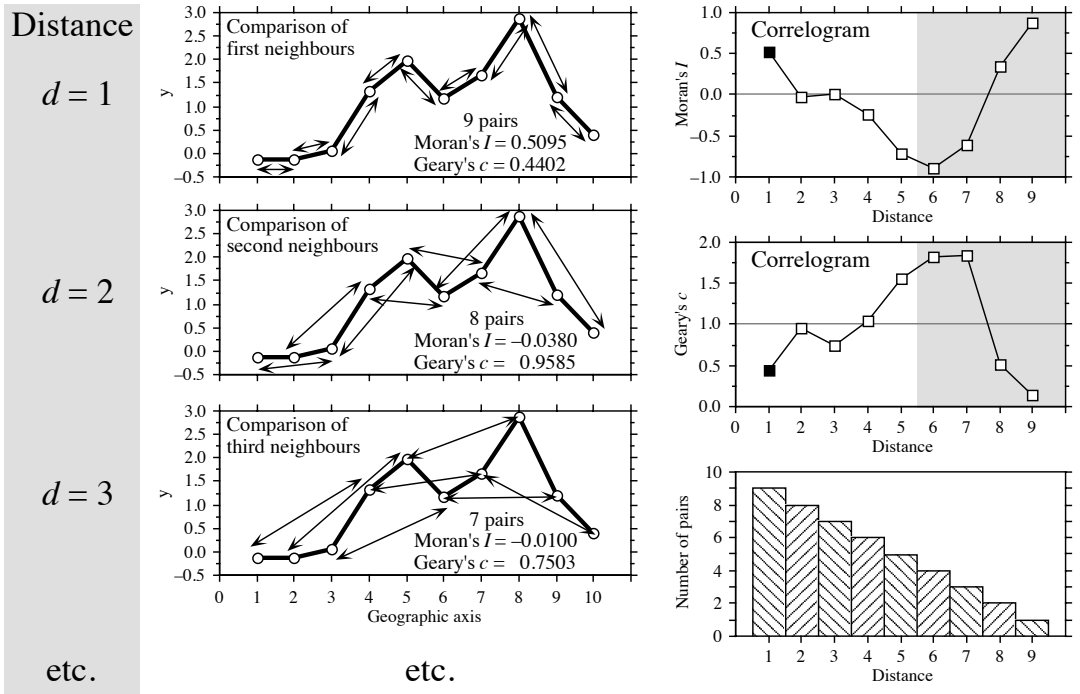


Figure 13.3 Construction of correlograms. Left: data series observed along a single geographic axis (10 equispaced observations). Moran's I and Geary's c statistics are computed from pairs of observations found at preselected distances ($d = 1$, $d = 2$, $d = 3$, etc.). Right: correlograms are graphs of the spatial correlation statistics plotted against distance. Dark squares: significant correlation 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 exclude some points found in the centre of the series or surface.

In the construction of a correlogram, spatial correlation coefficients are computed, in turn, for the various distance classes d. The weights w_{hi} are Kronecker deltas (as in eq. 7.21); the binary 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 $(n \times n)$ spatial weighting matrix **W**; an example of a binary spatial weighting matrix is matrix **X**(1) of Fig. 13.14. 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 or by $w_{hi} = 0$. In most applications, however, the geographic distance matrix among sites is symmetric and the coefficients can be computed from the half-matrix of distances; the formulae remain the same, because W and the sum in the numerator are half the values computed over the whole distance matrix **D**.

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. 13.21). For very broad-scale studies, involving a whole ocean or continent, "great-circle distances", i.e. distances along the earth's curved surface, should be used instead of straight-line distances through the earth crust.

Moran's I formula is related to the Pearson 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 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 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 spatial correlation in the data translates into positive values of I; negative correlation produces negative values.

Readers who are familiar with correlograms in time series analysis (Section 12.3) 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 (eq. 13.1) for the different distance classes is nearly the same as computing the autocorrelation coefficient of time series analysis (Fig. 12.5, eq. 12.7).

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 spatial correlation produces high values of I and low values of c (Fig. 13.3). Positive spatial correlation translates in values of c between 0 and 1 whereas negative correlation produces values larger than 1. Hence, the reference 'no correlation' value is $c = 1$ in Geary's correlograms.

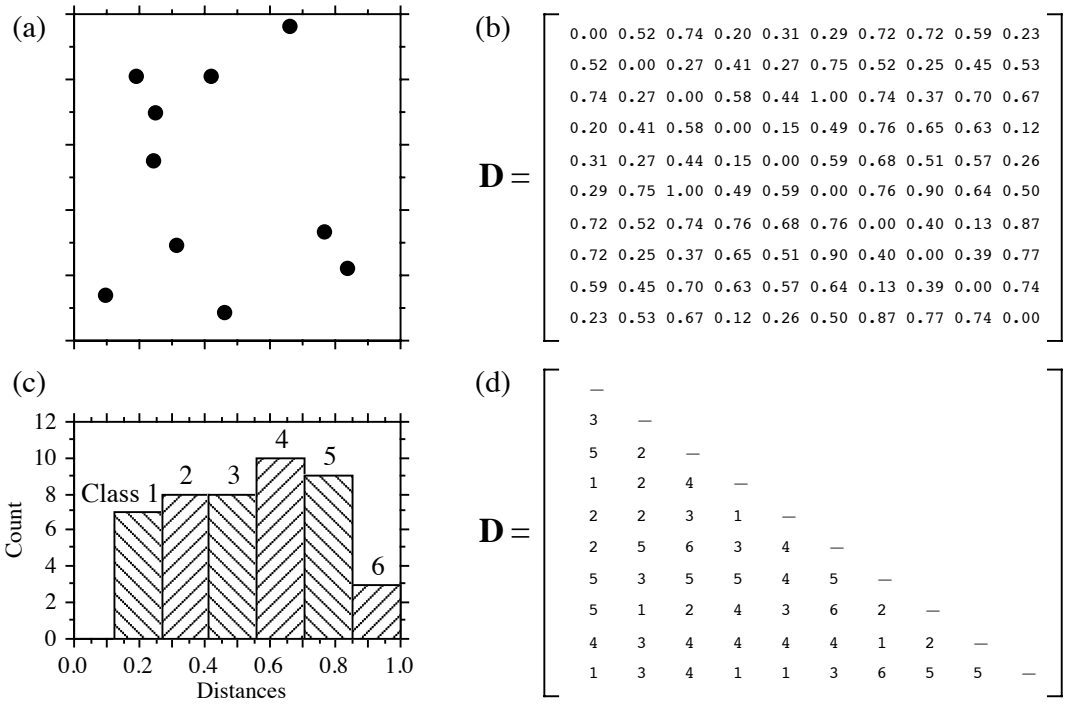


Figure 13.4 Calculation of distance classes, artificial data. (a) Map of 10 sites in a 1-km^2 sampling area. (b) Geographic distance matrix (**D**, in km). (c) Frequency histogram of distances (classes 1 to 6) for the upper (or lower) triangular portion of **D**. (d) Distances recoded into 6 classes.

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 correlation 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). Sturges' (1926) 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 correlation coefficients can be tested for significance and confidence intervals can be computed. With proper correction for multiple testing, one can determine if a significant spatial structure is present in the data and what are the distance classes showing significant positive or negative correlation. Tests of significance require, however, that certain conditions specified below be fulfilled.

Second-order stationarity

The tests require that the condition of *second-order stationarity* be satisfied. Second-order stationarity refers to the vectors separating pairs of values in the study area. This rather strong condition states that the mean of the variable is constant over the study area, and the spatial covariance (numerator of eq. 13.1) depends only on the length and orientation of the vector between any two points, not on its position in the study area (David, 1977). The variance (denominator of eq. 13.1) must be the same for all points in the study area (homogeneity of the variance; Dutilleul, 2011).

Intrinsic stationarity

A relaxed form of stationarity, called *intrinsic stationarity*, states that the differences $(y_h - y_i)$ for any distance d (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 spatial correlation function is adequate to describe the entire surface under study. An example where intrinsic stationarity 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 spatial correlation functions. Second-order stationarity implies intrinsic stationarity, but the reciprocal is not true. Intrinsic stationarity is a weaker form of stationarity compatible with a broader range of models. 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 correlation coefficients. For any normally distributed statistic *Stat*, a confidence interval at significance level α is obtained as follows:

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

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

$$Stat_\alpha = z_\alpha \sqrt{\mathrm{Var}\,(Stat)} + \text{Expected value of } Stat \text{ under } H_0 \qquad \textbf{(13.5)}$$

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:

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

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 (second-order stationarity assumption) 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**;
- $b_2 = n\sum_{i=1}^{n}(y_i-\bar{y})^4 \Big/ \left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^2$ 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 correlation are one-tailed because the sign of the correlation is stated in the ecological hypothesis; for instance, contagious biological processes such as growth, reproduction, and dispersal, all suggest that ecological variables are positively correlated 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 spatial correlation (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 spatial correlation), 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 spatial correlation), 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 permutation (Subsection 1.2.2).

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 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 spatial correlation 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 correlation statistic that is significant at the Bonferroni-corrected significance level (Box 1.3). Simulations by Oden (1984) showed 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 spatial correlation statistics that are significant, in order to reach an interpretation (Subsection 13.1.2). One could rely on Bonferroni-corrected tests for all individual correlation 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 spatial correlation is to be expected in the smallest distance classes and the purpose of the analysis is to determine the extent of the spatial correlation (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. In addition, 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.

Spatial correlation coefficients and tests of significance also exist for qualitative (nominal) variables (Cliff & Ord, 1981); they have been used, for example, to analyse spatial patterns of sexes in plants (Sakai & Oden, 1983; Sokal & Thomson, 1987). Special types of spatial correlation 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 spatial correlation function is the same for all geographic directions considered, the phenomenon is *isotropic*. The opposite of isotropy is *anisotropy*. When a variable is isotropic, a single correlogram can be computed over all directions of the study area. The correlogram is said to be *all-directional* or *omnidirectional*. Directional correlograms, which are computed for a single spatial direction, are discussed together with anisotropy and directional variograms in Subsection 13.1.3.

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 be misleading and should be supported by examination of 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.

1. 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 Sturges' rule (eq. 13.3), and correlograms were computed. According to Oden's test, the correlograms were globally significant at the $\alpha = 5\%$ level since several individual values were 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 correlation 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 correlation 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 correlation 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, for which 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 correlation coefficient, which is above 0 in Moran's correlogram and below 1 in Geary's, indicates positive spatial correlation 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 correlation 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 were larger than the patch size, the first neighbours would be dissimilar in values and the correlation would be negative for the first distance class. The next peaking positive correlation 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 spatial correlation 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 spatial correlation 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 correlation 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 correlation coefficients for larger distance classes less significant and interpretable.

2. 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

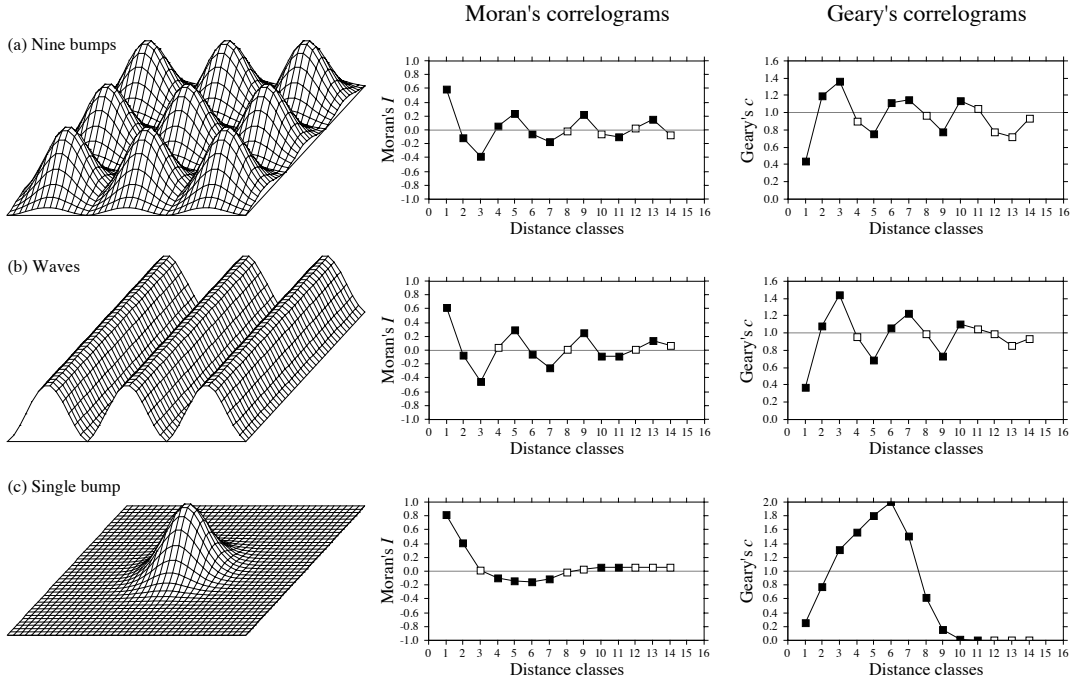


Figure 13.5 Spatial correlation analysis of artificial spatial structures shown on the left: (a) nine bumps; (b) waves; (c) a single bump. Centre and right: all-directional correlograms. Dark squares: correlation statistics that remain significant after progressive Bonferroni correction ($\alpha = 0.05$);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

correlograms alone cannot tell apart regular bumps from regular waves; directional correlograms or maps are required.

3. 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 spatial correlation, 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).

4. Linear gradient (Fig. 13.5d). — The correlogram is monotonic decreasing. Nearly all spatial correlation 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. Model 1 of Subsection 1.1.1 (induced spatial dependence, eq. 1.1) can generate a true gradient; see Fig. 1.5, case 4. That gradient can be modelled using trend-surface analysis (Subsection 13.2.1). The observed values have independent error terms,

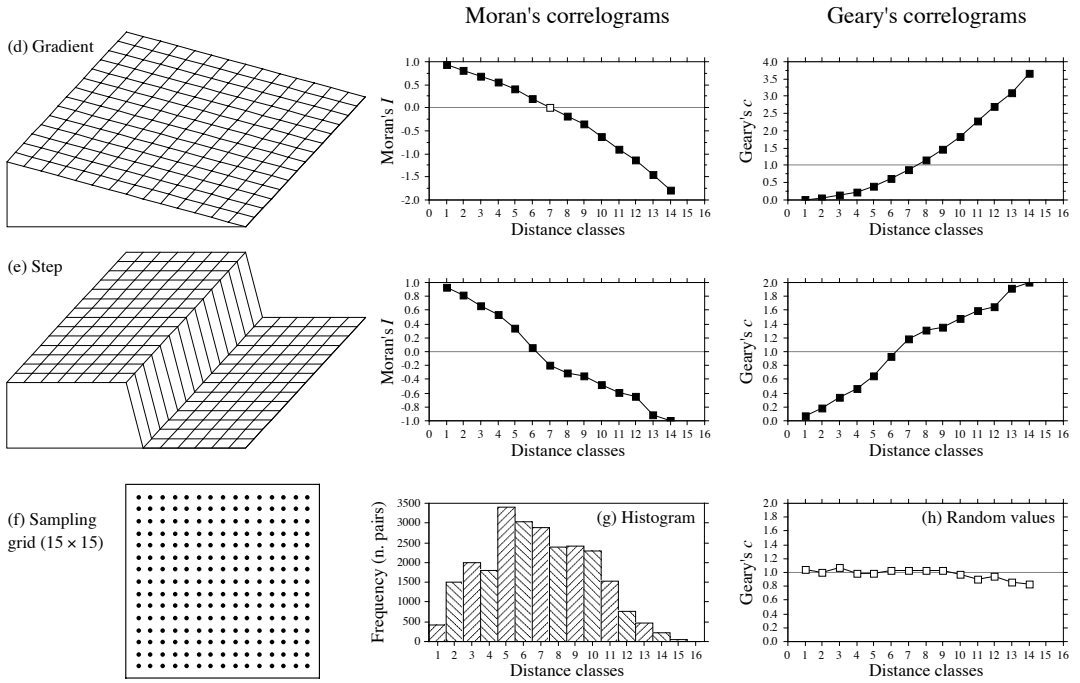


Figure 13.5 **(continued)** Spatial correlation analysis of artificial spatial structures shown on the left: (d) gradient; (e) step. (h) All-directional correlogram of random values. (f) Sampling grid used on each of the artificial spatial structures to obtain 225 "observed values" for spatial correlation 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 correlation statistics (I or c) are not shown for distance classes 15 and 16; see text.

i.e. error terms that are not autocorrelated. *False gradients*, on the other hand, are structures that look like gradients, but actually correspond to spatial correlation generated by some spatial process. Model 2 of Subsection 1.1.1 (spatial autocorrelation, eq. 1.2) can generate a false gradient, especially when the sampling area is small relative to the range of influence of the generating process; see Fig. 1.5, case 3.

In the case of "true gradients", spatial correlation coefficients should not be tested for significance because the condition of second-order stationarity is not satisfied (definition in Subsection 13.1.1); 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 intrinsic stationarity is satisfied. One may also choose to extract a "true gradient" using trend-surface analysis, compute residuals, and look for spatial correlation 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 correlation 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 spatial correlation at broader spatial scale ("false gradient").

5. Step (Fig. 13.5e). — A step between two flat surfaces is enough to produce a correlogram that 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 correlation analysis.

6. 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 spatial correlation coefficients were small and non-significant at the 5% level. Only the Geary correlogram is presented.

Sokal (1979) and Cliff & Ord (1981) described, 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.

- The absence may indicate that the surface under study is free of spatial correlation at the study scale. This conclusion is subject to *type II 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 spatial correlation) one wants to detect, (3) the number of observations (n), and (4) the variance of the sample of data (Cohen, 1988):

$$\text{Power} = (1 - \beta) = f(\alpha, \text{size of effect}, n, s_x^2)$$

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.

- The absence may also indicate that the sampling design is inadequate to detect the spatial correlation that may exist in the system. Are the grain size, extent and sampling interval (Section 13.0) adequate to detect the type of spatial correlation 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 the process generating the spatial correlation, not the exact geographic position of each resulting value. Correlograms may be used to support or reject a 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 *et al*. (1999) and Bjørnstad & Falck (2001) 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 a confidence envelope around the entire correlogram. Confidence envelopes allow one to test the similarity between correlograms of real or simulated data. See package NCF in Section 13.6.

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. Pictures of plates were taken during the course of the study. The present ecological application uses data obtained from a picture of a plate taken 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 application (Fig. 13.6a; not published in *op. cit*.). The hypothesis to be tested was that barnacles had 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-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 correlation 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 spatial correlation was

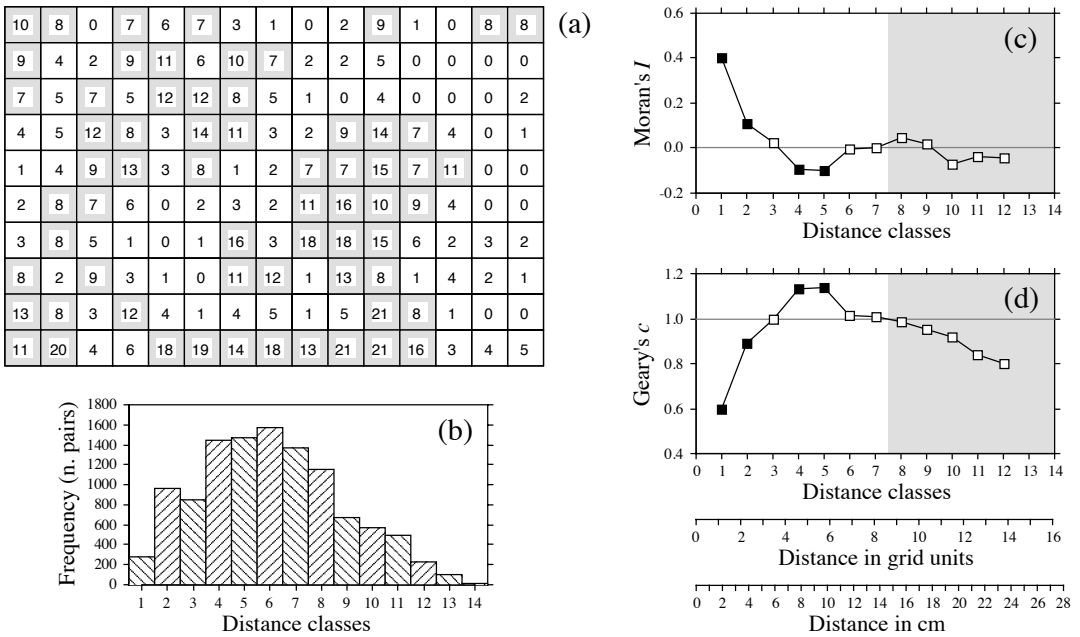


Figure 13.6 (a) Counts of adult barnacles in 150 (1.7 × 1.7 cm) pixels on a plate of artificial substrate (17 × 25.5 cm). The mean concentration is 6.17 animals per pixel; pixels with counts ≥ 7 are shaded to display the aggregates. (b) Histogram of the number of pairs in each distance class. (c) Moran's correlogram. (d) Geary's correlogram. Dark squares: spatial correlation statistics that remain significant after progressive Bonferroni correction ($\alpha = 0.05$); white squares: non-significant values. Grey zones: coefficients that should not be interpreted because they exclude some points in the centre of the study area. 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 in cm.

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 Moran's *I* correlation value is found. This occurs in classes 4 and 5, which correspond 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).

A spatial correlogram is an overall function of spatial correlation across a study area. It is not meant to display details of the structure across the area. Anselin (1995) proposed to decompose the global spatial correlation coefficients into *Local Indicators of Spatial Association* (LISA), producing a local statistic for each sampling unit compared to its surrounding units. LISA can be computed using Moran's *I* or Geary's *c* formulas (eqs. 13.1 and 13.2), and the resulting values can be plotted on maps. Fortin & Dale (2005) give examples of such maps of LISA computed for

LISA

simulated data. Readers can also run the example provided in the documentation file of function ***lisa()*** of package NCF in R.

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 13.1.3 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 was proposed by Oden & Sokal (1986); Legendre & Fortin (1989) gave 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 correlation 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. A spatial correlation 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:

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

or, for symmetric distance and weight matrices, which is the most common case:

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

$\gamma(d)$ 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 equation forms produce 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.

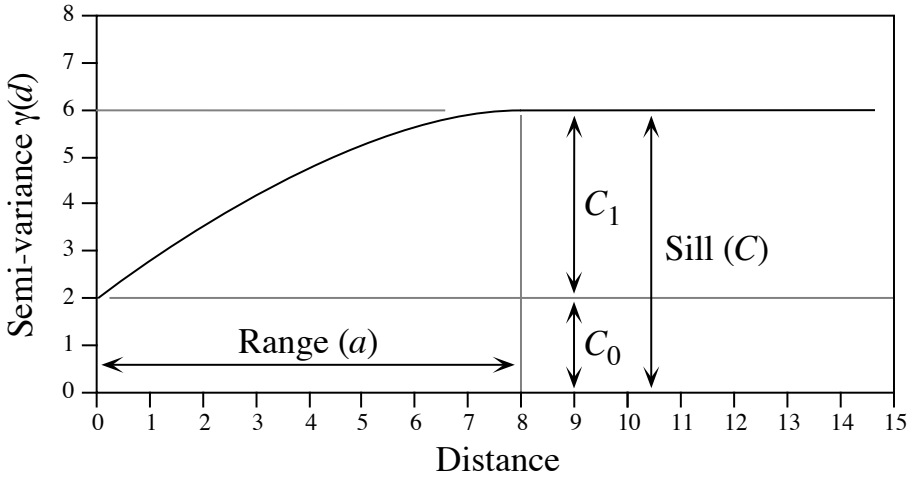


Figure 13.7 Spherical variogram model showing characteristic features: nugget effect ($C_0 = 2$ in this example), spatially structured component ($C_1 = 4$), sill ($C = C_0 + C_1 = 6$), and range ($a = 8$).

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

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

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 spatial correlation function applies to the entire surface under study (intrinsic stationarity, 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

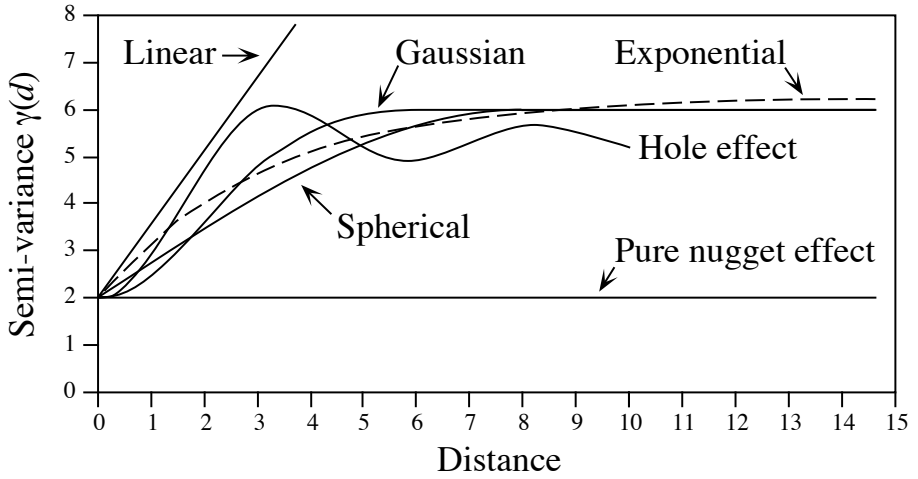


Figure 13.8 Commonly used variogram models.

by the error term ε_{ij} in spatial structure model 2 (eq. 1.2) 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 $C_0/(C_0 + C_1)$.

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 latter 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 variogram 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 may be fitted by weighted least squares; the weights are functions of the distance and the number of pairs in each distance class (Cressie, 1991); in practice, variograms are often fitted by visual estimation. Fitting a variogram model requires that the hypothesis of intrinsic stationarity be satisfied (Subsection 13.1.1).

Anisotropy

As mentioned at the beginning of Subsection 13.1.2, anisotropy is present in data when the spatial correlation 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.

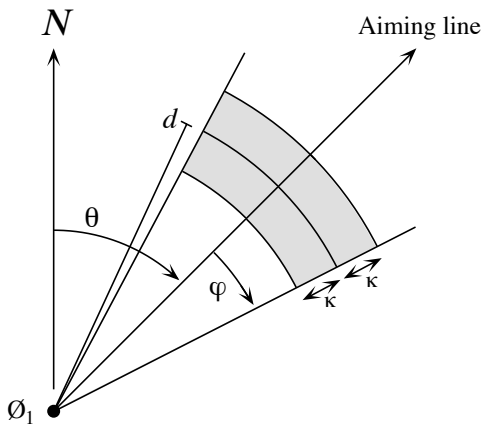


Figure 13.9 Search parameters for pairs of points in directional variograms and correlograms. From an observed study site $Ø_1$, an aiming line is drawn in the direction determined by angle θ (usually by reference to the geographic north, indicated by *N* in the figure). The angular tolerance parameter 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 $\gamma(d)$.

Directional variogram and correlogram

Directional variograms and correlograms may be used to determine whether anisotropy (defined in Subsection 13.1.2) is present in 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 within the search window are paired with the reference point $Ø_h$ and included in the calculation of semi-variance or spatial correlation 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, their Chapter 7) propose a way to assemble directional measures of semi-variance into a single table and 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.

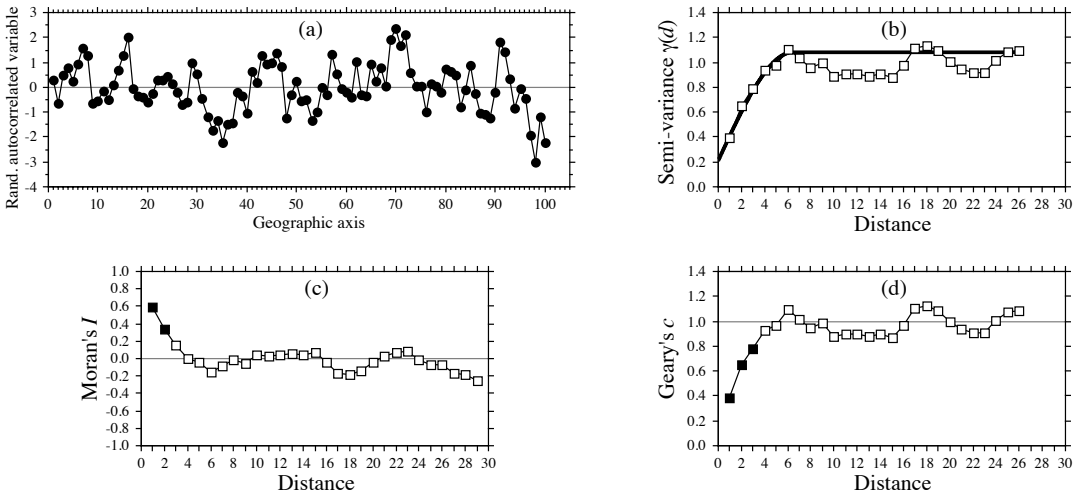


Figure 13.10 (a) Series of 100 equispaced random, spatially autocorrelated data. (b) Sample variogram, with spherical model superimposed (heavy line). Abscissa: distances between points along the geographic axis in (a). (c-d) Spatial correlograms. Dark squares: spatial correlation statistics that remain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

Numerical example. An artificial data set was produced containing random autocorrelated data (Fig. 13.10a). 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. The sample variogram (without test of significance) and spatial correlograms (with tests) are shown in Fig. 13.10b-d. In this example, the data were standardized during data generation, so that the denominator of eq. 13.2 was 1; therefore, the sample variogram and Geary's correlogram were 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 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 spatial correlation and variance that would be observed if the grain size of the sampling design were different from that actually used in a field study (Bellehumeur *et al*., 1997).

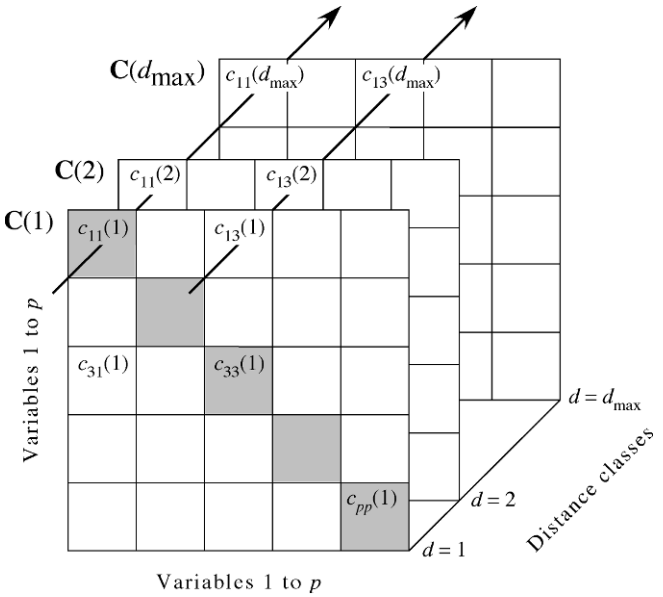


Figure 13.11 Representation of a variogram matrix **C** containing the information from all variograms and cross-variograms. **C** is composed of separate variance-covariance matrices **C**(d), each of size ($p \times p$), corresponding to one of the distance classes d. Redrawn from Wagner (2003).

4 — Multivariate variogram

Consider a multivariate matrix **Y** with n rows (sites) and p columns, e.g. species presence-absence or abundance data. A variogram $\gamma_j(d)$ for a single variable j is computed using eq. 13.10. The cross-variogram $\gamma_{jk}(d)$ between two variables j and k is now defined as follows (Isaaks & Srivastava, 1989):

$$\gamma_{jk}(d) = \frac{1}{2W(d)} \sum_{(h,i)\,|\,d_{hi} \approx d}^{W(d)} (y_{jh} - y_{ji})\,(y_{kh} - y_{ki}) \qquad \textbf{(13.11)}$$

It partitions the covariance between two variables among the distance classes d.

Each variogram and cross-variogram can be seen as a vector containing values computed for different distance classes; the largest distance class is labelled $d_{\max}$. For a multivariate response matrix **Y** of size ($n \times p$), a variogram is produced for each of the p variables and there is a cross-variogram for each of the $p(p-1)/2$ pairs of variables. These vectors can be assembled in a distance-dependent cubic symmetric variance-covariance matrix called the *variogram matrix* **C** (Myers, 1997; Fig. 13.11) with elements $c_{ij}(d) = \gamma_{jk}(d)$ (eq. 13.11). The arrows in the figure show the values

Variogram matrix

$c_{11}(d)$ used to draw the variogram $\gamma_{11}(d)$ of variable 1 and the values $c_{13}(d)$ used to draw the cross-variogram $\gamma_{13}(d)$ crossing variables 1 and 3.

Matrix **C** contains a series of square variance-covariance matrices **C**(d). Each matrix **C**(d) is of size ($p \times p$) because it is computed among the p descriptors; it contains the information for one of the distance classes d of each variogram and cross-variogram. The variance-covariance matrix $\mathbf{S_Y}$ of the p-dimensional matrix **Y** is the weighted sum of the **C**(d) matrices, showing that the set of **C**(d) matrices represents an additive decomposition of the total variance-covariance matrix $\mathbf{S_Y}$ among the distance classes d. The weights in that sum are the number of pairs of points used to compute the values in each distance class divided by the total number of pairs of points.

In order for the variances of the variables in data matrix **Y** to be additive, these must be in the same physical dimensions or standardized. This question was discussed in the first paragraph of Subsection 9.1.5. The variogram matrix can be used to plot several graphs (Wagner, 2003):

- The empirical variogram of variable j is obtained by plotting the diagonal elements $c_{jj}(d)$ (e.g. the values along the left-hand arrow in Fig. 13.11) against distances d.

- The empirical cross-variogram of variables j and k is obtained by plotting the non-diagonal elements $c_{jk}(d)$ (e.g. the values along the right-hand arrow in Fig. 13.11) against distances d.

Multivariate variogram

- Sum the diagonal elements (gray squares in Fig. 13.11) in each matrix **C**(d). Since the sum of the diagonal elements of **S** is the total variance in **Y** and the matrices **C**(d) decompose **S**, a plot of these sums against distances d is the *multivariate variogram* decomposing the total variance in **S**. An example is given in Ecological application 13.1b. Furthermore, Wagner (2003) showed that for species presence-absence data, a plot of these sums against distances d is an empirical *variogram of complementarity*, meaning the variogram of the dissimilarity in species composition. These sums are direct measures of species turnover between sites located at distances d; a higher sum of variances indicates larger differences among the sites separated by that distance than for other distances where the among-site sum of variances is lower.

- As shown in Section 4.1, the sum of all values in matrix **S** is equal to the variance of a new variable, **y**, computed as the sum by rows of all variables in **Y**. Because the matrices **C**(d) represent a decomposition of **S** among the distance classes d, one can sum all elements of each matrix **C**(d) and plot these sums against distances d to obtain a variogram of **y**. If **Y** contains species abundance data, the graph is a variogram of the total number of individuals at the sites, which can in some cases be interpreted as the total yield or the carrying capacity of the sites. If **Y** comprises species presence-absence data, a variogram of species richness is obtained (Wagner, 2003).

The statistics in multivariate variograms can be tested for significance using Mantel tests (Wagner, 2004). The tests used in function ***mso()*** of VEGAN in R, which

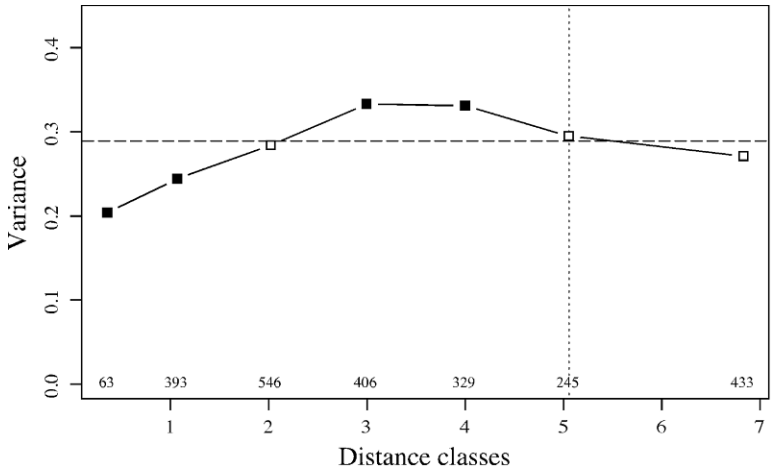


Figure 13.12 Multivariate variogram of the Hellinger-transformed and detrended mite data, computed using function ***mso()***. Dark squares: variances with p-values significant at the 5% level, after Bonferroni correction for 7 simultaneous tests. Dashed horizontal line: total variance in the data. Vertical dotted line: half the maximum number of classes; the last point, to the right of that line, includes all remaining pairs of sites and should not be interpreted. Values written above the abscissa: number of pairs involved in the calculation of each statistic.

are based on the matrix of squared distances, are identical to those used in the Mantel correlogram (Subsection 13.1.6).

Ecological application 13.1b

The oribatid mite data of Borcard & Legendre (1994), analysed in Ecological application 11.5, are used here to compute a multivariate variogram. Prior to analysis, the mite data were Hellinger-transformed (eq. 7.69) and detrended along the north-south axis of the study area to meet the stationarity assumption. Function ***mso()*** of the R VEGAN package was used to compute the variogram; see Section 13.6.

The results are shown in Fig. 13.12. The interval size of the distance classes was the distance that kept all points connected in a dbMEM analysis; this is the threshold distance (*thresh*) of Section 14.1, 1.01119 m. The horizontal line in Fig. 13.12 is the total variance in the data. It is also the weighted sum of the variances (sums of the diagonal elements) of the $\mathbf{C}(d)$ matrices over the different distance classes. Because the sum of the weights is 1, as explained in the description of the method, the dashed line is located at the weighted mean of the multivariate variogram values and can be used as a reference for their visual assessment.

The p-values were Bonferroni-corrected for 7 simultaneous tests. The variogram displays significant spatial correlation; it may correspond to a spherical or a hole model. These data will be further analysed by multiscale ordination in Section 14.4.

5 — 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. The assumption of second-order stationarity (Subsection 13.1.1) may be restated as follows:

- The first moment (mean of values i) of the variable has a constant and finite value:

$$\mathrm{E}[y_i] = \frac{1}{n}\sum_{i=1}^{n} y_i = m_i \qquad \textbf{(13.12)}$$

and its value does not depend on the position in the study area.

- The second moment (spatial covariance, numerator of eq. 13.1) of the variable exists (i.e. the variogram has a finite sill value):

$$C(d) = \left[\frac{1}{W(d)}\sum_{(h,i)\,|\,d_{hi}\approx d}^{W(d)} y_h y_i\right] - m_h m_i \qquad \textbf{(13.13)}$$

$$C(d) = \mathrm{E}[y_h y_i] - m^2 \quad \text{for } h, i \,|\, d_{hi} \approx d \qquad \textbf{(13.14)}$$

$h, i \,|\, d_{hi} \approx d$ means that the pairs of points h and i used to compute covariance $C(d)$ are at distances d_{hi} that are approximately equal to d. The values of $C(d)$ depend only on d and on the orientation of the distance vectors, not on their positions in the study area.

To understand eq. 13.13 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):

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

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

$$\mathrm{Var}[y_i] = \mathrm{E}[y_i - m_i]^2 = C(0) \qquad \textbf{(13.15)}$$

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

$$\gamma(d) = 0.5\ \mathrm{E}[y_h - y_i]^2 \quad \text{for } h, i \,|\, d_{hi} \approx d \qquad \textbf{(13.16)}$$

A few lines of algebra obtain the following formula:

$$\gamma(d) = \frac{\sum y_i^2 - \sum y_h y_i}{W(d)} = C(0) - C(d) \quad \text{for } h, i \mid d_{hi} \approx d \qquad \textbf{(13.17)}$$

Two properties are used in the derivation of eq. 13.17 from eq. 3.16: (1) $\sum y_h = \sum y_i$, and (2) the variance (Var[y_i], eq. 13.15) 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. For a spatial process, the (auto)correlation is written as follows:

$$r(d) = \frac{C(d)}{s_h s_i} = \frac{C(d)}{\mathrm{Var}[y_i]} = \frac{C(d)}{C(0)} \qquad \textbf{(13.18)}$$

The right-hand formula is Moran's I (eq. 13.1). Consider the formula for Geary's c (eq. 13.2), which is the semi-variance divided by the overall variance (ignoring the fact that the variance in eq. 13.2 is computed with division by $n-1$ instead of n). The following derivation

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

shows that Geary's c is one minus the coefficient of spatial (auto)correlation (ignoring again the division by $n-1$ instead of n). In a graph, the semi-variance and Geary's c coefficient have exactly the same shape (e.g. Fig. 13.10, b and d); only the ordinate scales may differ if Var[y_i] is not 1. An autocorrelogram plotted using $r(d)$ has the exact reverse shape as a Geary correlogram. The 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. 13.13, using values of *two different variables* observed at locations distant by d (Isaaks & Srivastava, 1989). Equation 13.18 leads to a formula for cross-correlation that may be used to plot cross-correlograms; the construction of the cross-correlation statistic is the same as for time series (eq. 12.9). With transect data, the result is similar to that of eq. 12.9. 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 13.1c

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

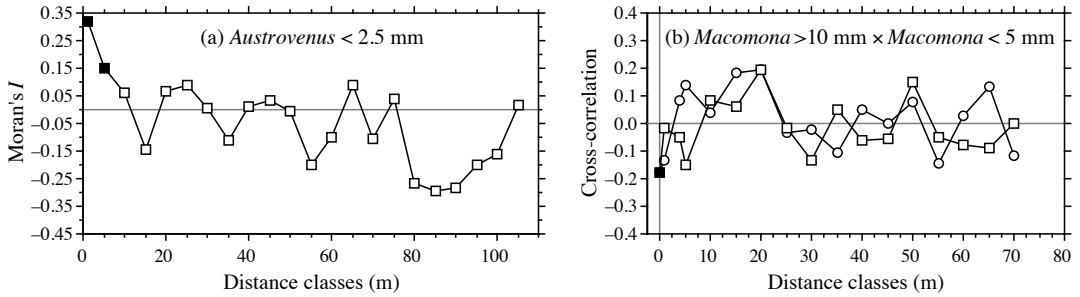


Figure 13.13 (a) Spatial autocorrelogram for juvenile *Austrovenus* densities. (b) Cross-correlogram for adult-juvenile *Macomona* interactions, folded about the ordinate: circles = positive lags, squares = negative lags. Dark symbols: correlation statistics that are significant after progressive Bonferroni correction ($\alpha = 0.05$). Redrawn from Hewitt *et al*. (1997).

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. This way, there were transects perpendicular to the direction of tidal flow, and others parallel. 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 comparisons 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 correlation at distances of 1 and 5 m (Fig. 13.13a). The same pattern was found in the transects parallel to tidal flow. Figure 13.13a also indicates that the range of influence of spatial correlation 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, their Figs. 3 and 4).

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. 13.13b); 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. 13.13b, 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.

6 — 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 **Y** of multivariate similarities or distances (using, for instance, coefficients S_{17} or D_{14} in the case of species abundance data), and compare **Y** to a *model matrix* **X** (Subsection 10.5.1), which is different for each geographic distance class (Fig. 13.14).

• 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 **X**(1) contains zeros. A first Mantel statistic (r_{M1}) is calculated between **Y** and **X**(1).

• The process is repeated for the other distance classes d, building each time a model-matrix **X**(d) and recomputing the normalized Mantel statistic. Matrix **X**(d) may contain 1's for pairs that are in the given distance class, or the code value for that distance class (d) (as in Fig. 13.14), 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. Borcard & Legendre (2012) have shown that for univariate data, the tests of significance in a Mantel correlogram computed on the matrix of *squared* Euclidean distances was equivalent to the tests in a Geary's c correlogram. Using numerical simulations, they also showed that the power of the test in Mantel correlograms was high for multivariate data.

A multivariate correlogram can be computed with function ***mantel.correlog()*** in R; see Section 13.6. If the calculation is based upon a *squared* Euclidean distance matrix, the Mantel test results in the multivariate correlogram are identical to the Mantel test results computed by the multivariate variogram function ***mso()***, provided that the distance classes are the same (Borcard & Legendre, 2012). As in the case of univariate correlograms (above), one is advised to use some form of correction for multiple testing (Box 1.3) before interpreting multivariate correlograms and variograms.

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. 13.14. Matrix **S** played here the role of $\mathbf{D_Y}$ 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 **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 **X**(d) matrices in turn; positive and significant Mantel

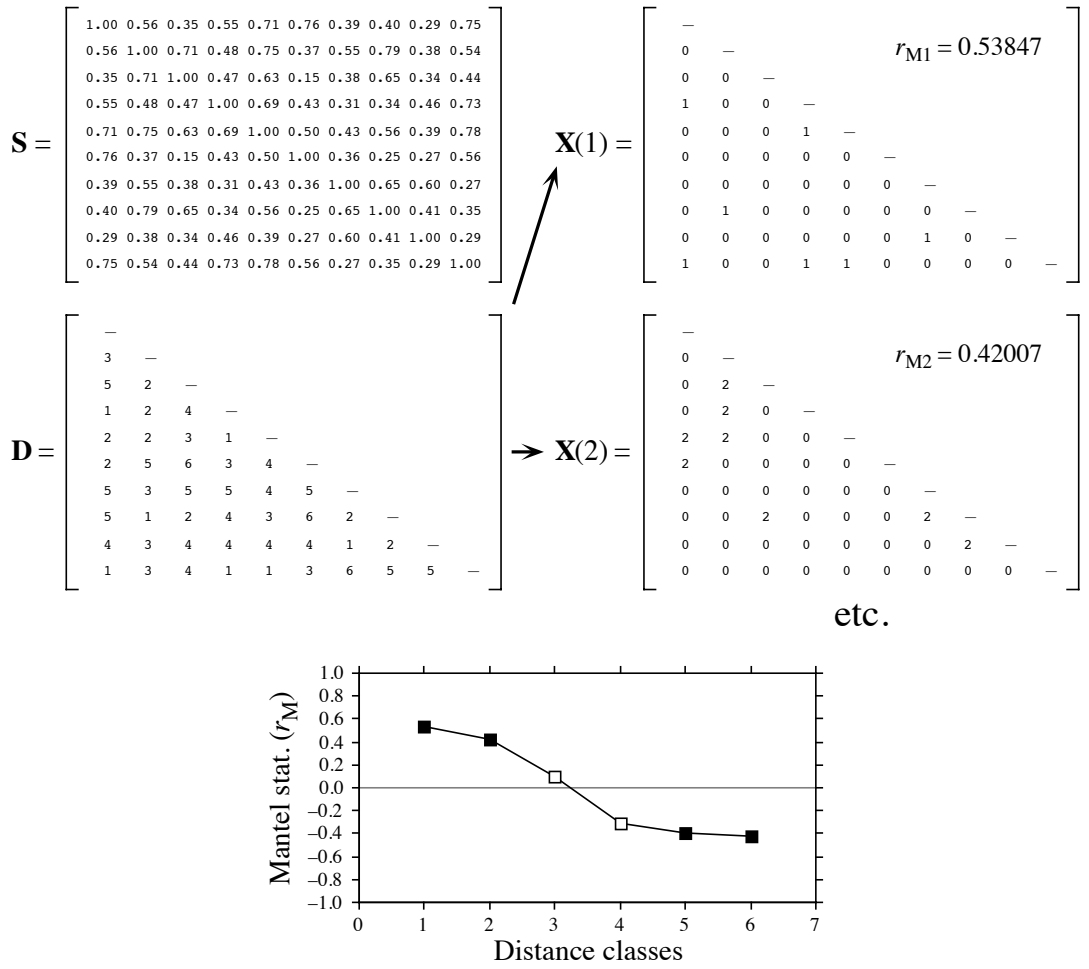


Figure 13.14 Construction of a Mantel correlogram for a similarity matrix **S** ($n = 10$ sites). The matrix of geographic distance classes **D**, from Fig. 13.4d, gives rise to model matrices **X**(1), **X**(2), etc. for the various distance classes d. These are compared, in turn, to matrix **S** using standardized Mantel statistics (r_{Md}). Dark squares in the correlogram: Mantel statistics that are significant after progressive Bonferroni correction ($\alpha = 0.05$).

statistics indicate positive spatial correlation 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 spatial correlation 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, variance and covariance at broad scale. The correlogram illustrated in Fig. 13.14 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 13.1.2.

When $\mathbf{D_Y}$ is a similarity matrix and distance classes are coded as described above, positive Mantel statistics correspond to positive spatial correlation; this is the case in the numerical example. When the values in $\mathbf{D_Y}$ are distances instead of similarities, or if the 1's and 0's are interchanged in matrix $\mathbf{X}$, the signs of all Mantel statistics are changed. One should always specify whether positive spatial correlation is expressed by positive or negative values of the Mantel statistics when presenting Mantel correlograms. Mantel correlograms have been computed for real data by Legendre & Fortin (1989), Le Boulengé *et al*. (1996), and Fortin & Dale (2005).

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. 13.15b and 13.18b), 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 ways mapped values are obtained. Spatial interpolation methods have been reviewed by Lam (1983).

Geographic information systems (GIS) are widely used nowadays, especially by geographers and increasingly by ecologists, 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 $\mathbf{Y}$) 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 because the interpolated values of such variables only make sense by reference to sampling units of the same size as those used in the original sampling design. Block kriging (Subsection 13.2.2) for blocks representing surfaces or volumes that differ from the grain of the observed data 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 13.2.2, but through a regression model calibrated over the entire study area.

In 1914, W. S. Gosset, writing under the pseudonym 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 many 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):

$$\hat{y} = f(\mathrm{X}, \mathrm{Y}) = b_0 + b_1\mathrm{X} + b_2\mathrm{Y} + b_3\mathrm{X}^2 + b_4\mathrm{XY} + b_5\mathrm{Y}^2 + b_6\mathrm{X}^3 + b_7\mathrm{X}^2\mathrm{Y} + b_8\mathrm{XY}^2 + b_9\mathrm{Y}^3 \quad \textbf{(13.19)}$$

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 that is linear in its parameters, although the response of y to the explanatory variables in matrix $\mathbf{X} = [\mathrm{X},\mathrm{Y}]$ 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; hence parameter b_0 does not have to be fitted and it can be removed from the model.

Numerical example. The data from Table 10.6 are used here to illustrate the method of trend-surface analysis. The dependent variable of the analysis, y, is Ma, which was the log-transformed ($\log_e(x + 1)$) 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. 13.15a). 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

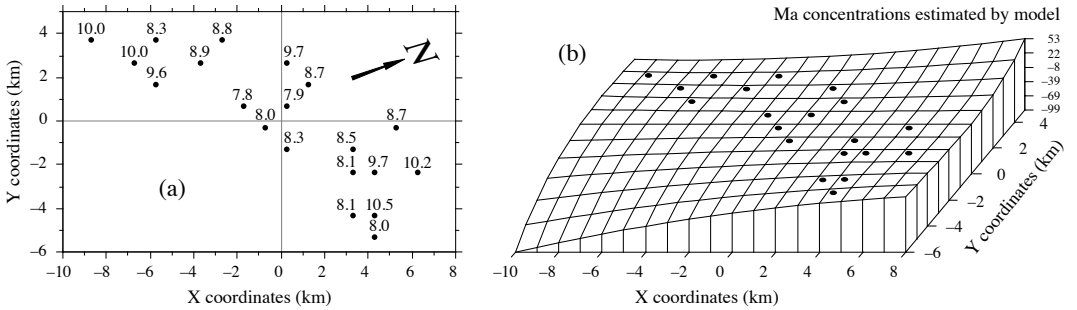


Figure 13.15 Variable Ma (log-transformed concentrations of aerobic heterotrophic bacteria growing on marine agar at salinity of 34 psu) at 20 sites in the Thau coastal lagoon, France, on 25 October 1988. (a) Map of the sampling sites with respect to arbitrary geographic coordinates X and Y. The observed values of Ma, from Table 10.6, 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.

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.

• Using the values of coordinates X and Y, calculate the terms of the third-degree polynomial, by combining variables X and Y as follows: X^2, X×Y, Y^2, X^3, X^2×Y, X×Y^2, Y^3. Alternatively, one could compute a third degree orthogonal polynomial of the geographic coordinates. Ordinary and orthogonal polynomials can both be computed by function ***poly()*** in R (Section 13.6).

• 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, 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, with the consequence that nonsignificant relationships are declared significant too often (Subsection 1.1.2).

• 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. 13.15b) using the file with (X', Y', and $\hat{y}'$). Values estimated by a trend-surface equation at the study sites do not coincide with the values observed at these sites; regression is not an exact interpolator, contrary to kriging (Subsection 13.2.2).

Different features could be displayed by rotating the figure. The orientation chosen in Fig. 13.15b 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 that 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.

The degree of the polynomial that is appropriate to model a phenomenon is partly predictable. 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 that 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, as 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 that case, is expected to be stronger near the impacted site, tapering off as one gets farther away from it.

Assuming that data were collected along a transect (a single geographic coordinate X) and that the impacted site was near the centre of the transect, a polynomial equation would not be appropriate to model an inverse-squared-distance diffusion process (Fig. 13.16a). 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. 13.16b). The minimum of that equation is b_0; this value occurs 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:

$$\hat{y} = b_0 + \frac{b_1 X^2 + b_2 Y^2}{b_3 X^2 + b_4 Y^2 + 1}$$

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, but it does not produce accurate fine-grained maps of the spatial variation of a variable. Other methods described in Chapter 14 allow researchers 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 situations, and especially in studies that cover large geographic expanses, the broad-scale trend may be already known and understood; researchers interested in geographic variation

Detrending

patterns may want to conduct analyses on detrended data, i.e. data from which the

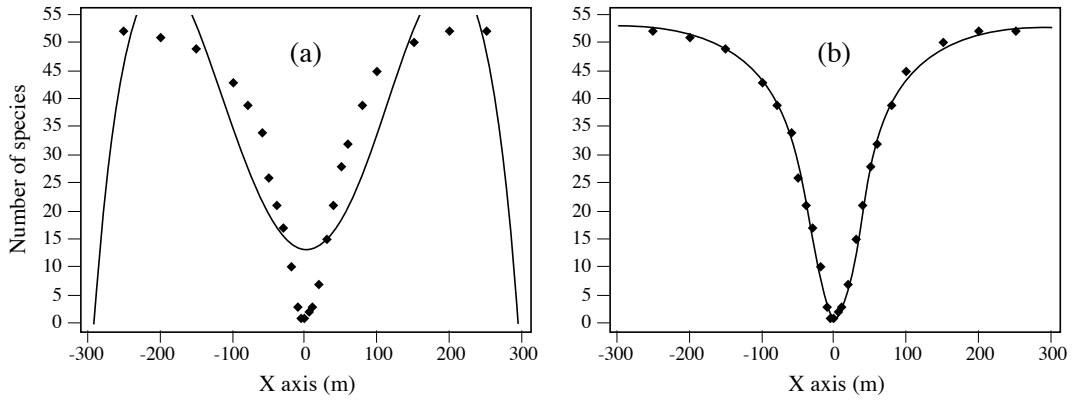


Figure 13.16 (a) Artificial data representing the number of species around the site of an environmental disturbance (located at X = 0) are not well-fitted by a 4th-order polynomial equation of the X coordinates ($R^2 = 0.7801$). (b) They are well-fitted by the following inverse-squared-distance diffusion equation: $\hat{y} = 1 + [0.0213X^2 / (0.0004X^2 + 1)]$ ($R^2 = 0.9975$).

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 geographical observation 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 latter variation is computed from the deviations from the means at the various locations; it is the residual mean square of an ANOVA testing for differences among locations. When the trend surface goes through the expected values at the various locations, these two error mean squares are not much different, and their F-ratio does not significantly differ from 1. 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 larger than the residual within-location variation. 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. 13.17. 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 of data into 6 sites as the classification

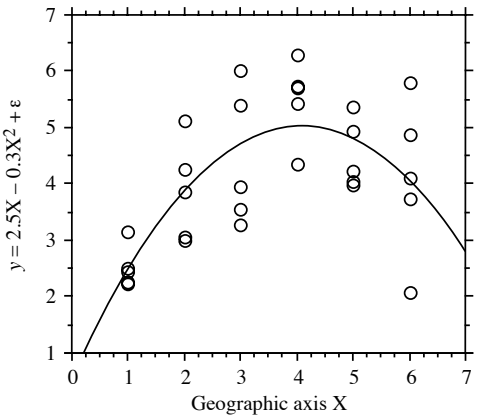


Figure 13.17 Artificial data representing sampling along a geographic axis X with 5 replicates at each site; $n = 30$. The F-test of goodness-of-fit indicates that the trend-surface equation $\hat{y} = 0.562 + 2.184X - 0.267X^2$ ($R^2 = 0.4899$) fits the data properly.

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 ($p = 0.530$), 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, $p = 0.165$, 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 $-0.3X^2$), 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**.

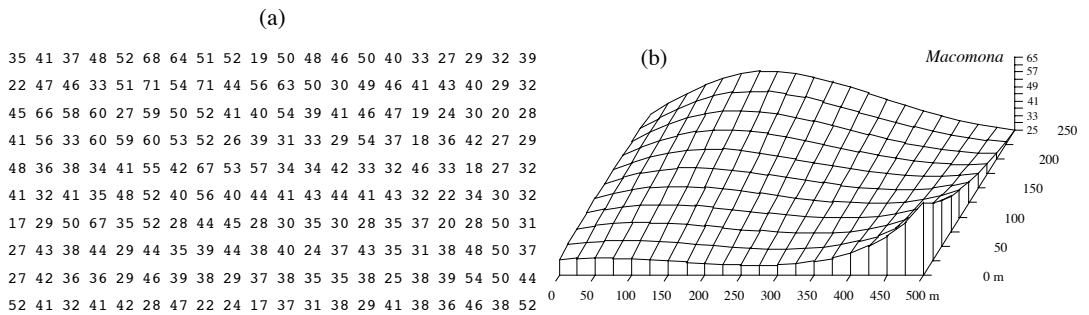


Figure 13.18 *Macomona* > 15 mm at 200 sites in Manukau Harbour, New Zealand, on 22 January 1994. (a) Actual counts at sampling sites in 200 regular grid cells; in the field, sites were not perfectly equispaced. (b) Map of the trend-surface equation explaining 32% of the spatial variation in the data. 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).

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, respectively.

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 controlled 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 were reported by 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 (R^2 = 0.10-0.20) than for larger animals (R^2 = 0.30-0.55). The best models, i.e. those with the highest 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

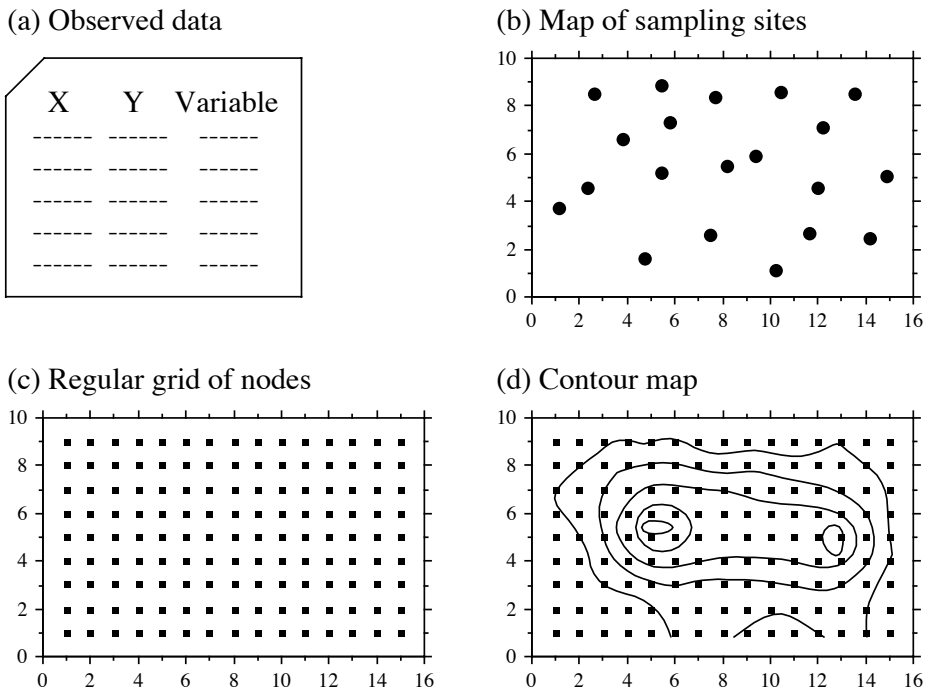


Figure 13.19 Summary of the interpolation procedure.

numerous than *Austrovenus*. A map illustrating the trend-surface equation is presented for the largest *Macomona* size class (Fig. 13.18); the field counts are also shown for comparison.

2 — Interpolated maps

In the family of interpolated map 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 13.2.1), where estimates of the variable at given locations were not obtained by interpolation, as in the present subsection, but through a statistical model whose parameters were estimated from all observations in the study area. Figure 13.19 illustrates the principle of interpolation mapping. A regular grid of nodes (Fig. 13.19c) is defined over the area that contains the study sites $Ø_i$ (Fig. 13.19a, b). Interpolation assigns a value to each point of the grid. This is the single most important step in mapping. Following that, results may be represented in the form of contours (e.g. Fig. 13.19d) with or without colours or shades, or three-dimensional constructs such as Fig. 13.18b.

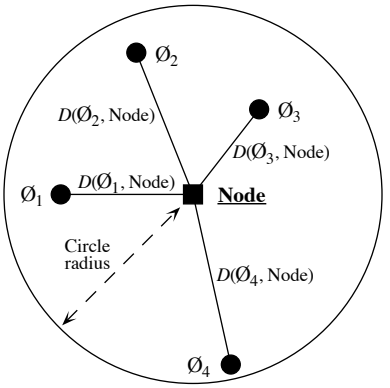


Figure 13.20 To estimate the value at a grid node (square), draw a search circle around it and consider the observed points ($Ø_i$) found within the circle. Observed points are separated from the node by distances $D(Ø_i, \text{Node})$.

Estimating the value corresponding 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 vertices (corners) of a Delaunay triangle, provides interpolated values for all points inside the triangle. Maps obtained using these solutions are shown in Chapter 11 of Isaaks & Srivastava (1989).

Alternatively, one may draw a "search circle" (or an ellipsoid for anisotropic data) around each grid node (Fig. 13.20). 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:

$$\hat{y}_{\text{Node}} = \sum_i w_i y_i \qquad \textbf{(13.20)}$$

where y_i is the value observed at point $\emptyset_i$ and weight w_i is the inverse of the distance (D) from point $\emptyset_i$ 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 $\emptyset_i$ in the estimation, so as to produce values that are consistent with the values observed at points $\emptyset_i$ (unbiasedness condition):

$$w_i = \left(\frac{1}{D(\emptyset_i, \text{Node})^k} \right) \Big/ \sum_i \frac{1}{D(\emptyset_i, \text{Node})^k} \qquad \textbf{(13.21)}$$

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. 13.15b), 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 $\emptyset_i$ 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 $\emptyset_i$ 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. The value estimated by the polynomial equation for the coordinates of a grid node is denoted z_{Node}. This method suffers from the same problem as inverse distance weighting with respect to observation sites $\emptyset_i$ 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 13.1.3), 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. 13.20), the estimated value for any grid node is computed as:

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

The chief difference between kriging and inverse-distance weighting is that, in kriging, the weights w_i applied to the points $\emptyset_i$ 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, eqs. 13.9 and 13.10) read on a variogram model (Subsection 13.1.3). They are found by linear estimation, using the equation:

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

$$\begin{bmatrix} c_{11} & \dots & c_{1n} & 1 \\ . & \dots & . & 1 \\ . & \dots & . & 1 \\ c_{n1} & \dots & c_{nn} & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ . \\ . \\ w_n \\ \lambda \end{bmatrix} = \begin{bmatrix} d_1 \\ . \\ . \\ d_n \\ 1 \end{bmatrix} \qquad \textbf{(13.22)}$$

where **C** is the covariance matrix among the n points $\emptyset_i$ used in the estimation, i.e. the semi-variances corresponding to the distances separating the various pair of points, provided by 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 $\emptyset_i$ 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. Element λ in vector **w** 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):

$$\mathbf{w} = \mathbf{C}^{-1}\,\mathbf{d} \qquad \textbf{(13.23)}$$

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. 13.22:

$$s^2_{OK} = \mathrm{Var}[y_i] - \mathbf{w}'\mathbf{d} \qquad \textbf{(13.24)}$$

where $\mathrm{Var}[y_i]$ is the maximum-likelihood estimate of the variance of the observed values y_i (eq. 13.15). Equation 13.24 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 a single block may 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, only additive variables can be used in block kriging. Block kriging programs always assume that the variable is *intensive*, e.g. 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 can be used instead (Isaaks & Srivastava, 1989, their 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 that does not often occur), the F-test of goodness-of-fit described in Subsection 13.2.1 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 problem are presented in Subsection 13.3.1.

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. 13.21a), bishop's (diagonal: Fig. 13.21b), or king's connections (also called queen's: both rectangular and diagonal, Fig. 13.21c). Sampling in staggered rows leads to connecting each point

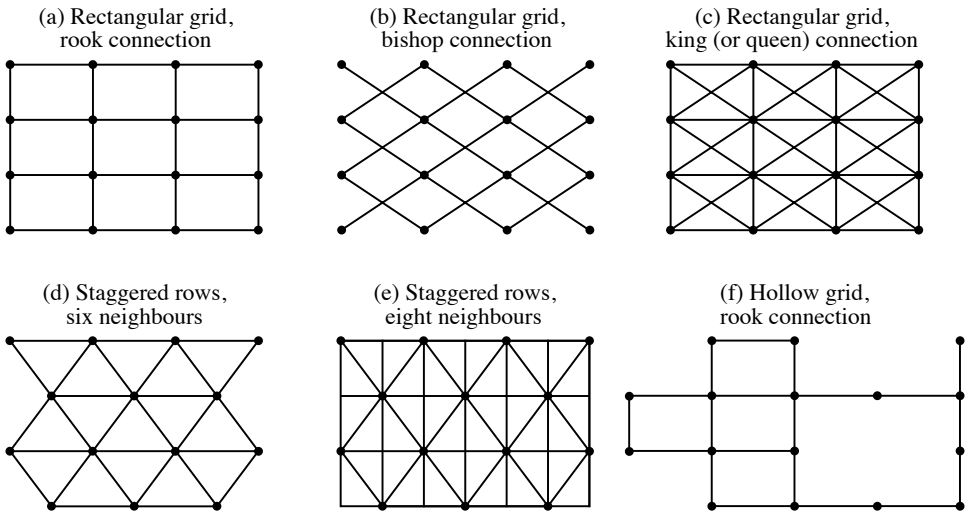


Figure 13.21 Connecting schemes for regular grids of points. See text.

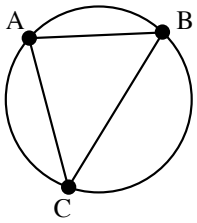
(except borders) to six (Fig. 13.21d) or eight neighbours (Fig. 13.21e). Algorithms may allow the construction of regular grids with missing points (Fig. 13.21f). When the objects represent irregularly-shaped land units covering a geographic area (e.g. types of ecosystems in a nature reserve), parcels sharing a common boundary are regarded as contiguous.

When the sites are positioned in an irregular manner, one can use geometric connecting schemes such as Delaunay triangulation, Gabriel graph, relative neighbourhood graph or minimum spanning tree, described below. There exists an inclusion relationship among these four connecting schemes: all edges that are members of a minimum spanning tree (MST) 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):

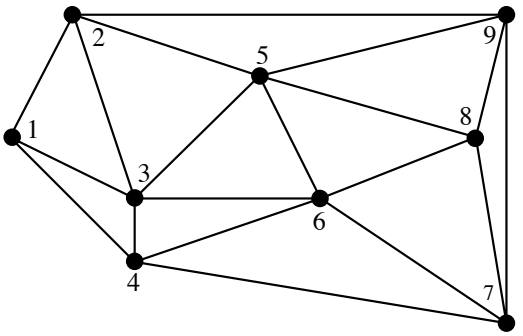
$$\text{MST} \subseteq \text{Relative neighbourhood graph} \subseteq \text{Gabriel graph} \subseteq \text{Delaunay triangulation}$$

Delaunay triangulation

• Delaunay triangulation. — The Delaunay triangulation criterion (Dirichlet, 1850; Upton & Fingleton, 1985) is illustrated in Fig. 13.22. For any triplet of points A, B and C, the three edges (i.e. lines) connecting these points are included in the 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 space-constrained clustering (Subsection 13.3.2).



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



19 edges form the Delaunay triangulation:

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

Figure 13.22 Construction of a Delaunay triangulation for 10 points.

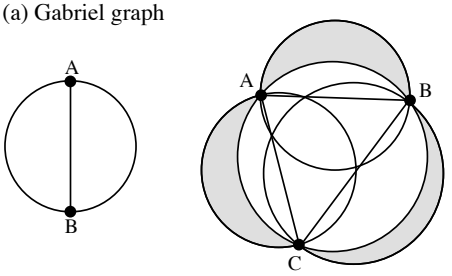
Long edges may be created at the outskirts of a set of points, simply because there is no other point located farther away in the sampling design; this is called a *border effect*. For example, edges 2–9 and 7–9 could have been removed from the triangulation in Fig. 13.22 by the presence of other points in the circumscribed circles of triangles (2, 5, 9) and (7, 8, 9) had the sampling extent been broader. Long peripheral edges can be removed by hand from the list, or by the computer algorithm.

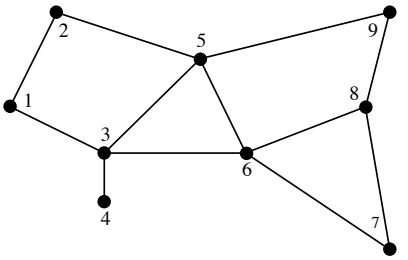
Gabriel graph

• Gabriel graph. — The Gabriel graph criterion (Gabriel & Sokal, 1969) differs from that of the Delaunay triangulation (Fig. 13.23a). 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.

The Gabriel graph in Fig. 13.23a is constructed for the same points as the Delaunay triangulation in Fig. 13.22. 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

(a) Gabriel graph

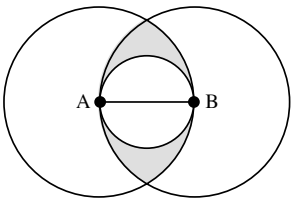


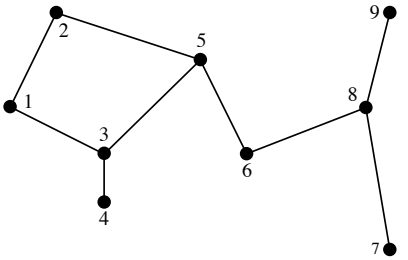


12 edges form the Gabriel graph:

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

(b) Relative neighbourhood graph





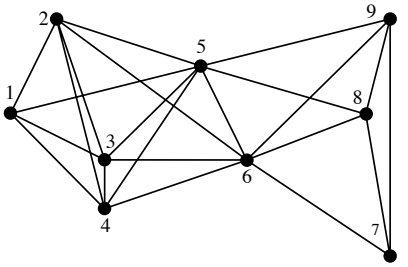
9 edges form the relative neighbourhood graph:

1–2 1–3 2–5 3–4 3–5 5–6
6–8 7–8 8–9

(c) Maximum distance graph

Criterion: $D \leq$ threshold

In this example, $D \leq 5$



22 edges form the maximum distance graph:

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

Figure 13.23 (a) Left: geometric criterion for the Gabriel graph. Centre: the zone of exclusion of the criterion, here for three points (grey zones + white inner circle), is larger than that of the Delaunay criterion (white inner circle). Right: graph for the example data, containing 12 edges. (b) Left: geometric criterion of the relative neighbourhood graph. The zone of exclusion of the criterion, here for two points (grey zones + white inner circle), is larger than that of the Gabriel criterion (white inner circle). Right: graph for the example data, containing 9 edges. (c) Left: criterion of the maximum distance graph. Right: graph for the example data with $D \leq 5$, with 22 edges.

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. 13.23b). 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*. Points that fall on the circumference of one of the circles in the intersection zone do not count. In algebraic terms, 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. 13.23b is constructed for the same set of points as in Figs. 13.22 and 13.23a. The 9 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.

Maximum distance graph

• Maximum distance graph. — Another strategy is to select a distance threshold and connect all points that are within that distance of each other. The result is called a maximum distance graph or an influence circle graph (Fig. 13.23c). One possible criterion to choose the distance threshold is to make it equal to the range of a variogram model (Fig. 13.7) computed for univariate (Subsection 13.1.3) or multivariate response data (Subsection 13.1.4).

Minimum spanning tree (MST)

• Minimum spanning tree (MST). — This tree connects the n points in the study with $(n - 1)$ edges. The sum of the weights (i.e. distances) of the edges used in the tree is minimum, meaning that it is smaller than or equal to the sum of the edge weights of any other tree connecting these n objects. 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. The minimum spanning tree for the example data set is shown in Fig. 14.3.

The list of connecting edges (Figs. 13.22 and 13.23) 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

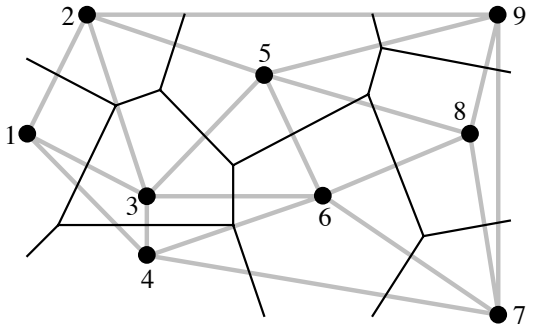


Figure 13.24 Delaunay triangulation (grey lines) and influence polygons (black lines) for the nine points of Fig. 13.22.

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 be able to cross water bodies (lake, sea) and settle in non-contiguous sites, which should be considered contiguous for the analysis 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 polygons

It is sometimes 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 is also referred to as a Dirichlet tessellation (1850), Voronoï polygons (1909), or Thiessen polygons (1911), from the names of the authors who described these mathematical structures.

Polygons are easily constructed from a Delaunay triangulation (Fig. 13.24). 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 — Space-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 as follows from its unconstrained counterpart:

- 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 smallest fusion distance or the largest fusion similarity. 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^2;

- 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 13.3.1. 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^2. This is balanced by the fact that the resulting clusters are likely to more readily interpretable.

It is fairly easy to modify clustering algorithms to incorporate a constraint of spatial contiguity (Fig. 13.25). As an example, consider the clustering methods included in the Lance and Williams general agglomerative model (Subsection 8.5.9). At the beginning of the clustering process, the vector of group membership has each object in a different group (Fig. 13.25, right). Proceed as follows:

1. Compute a distance matrix (**D**) among objects using the non-geographic information. Turn it into a similarity matrix **S** using one of the equations of Subsection 7.2.1. This transformation will make step 3 of the procedure possible.

2. Choose a connecting scheme (Subsection 13.3.1) and produce a list of connection edges as in Figs. 13.22 and 13.23. 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, i.e. their product element by element (Section 2.5). 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

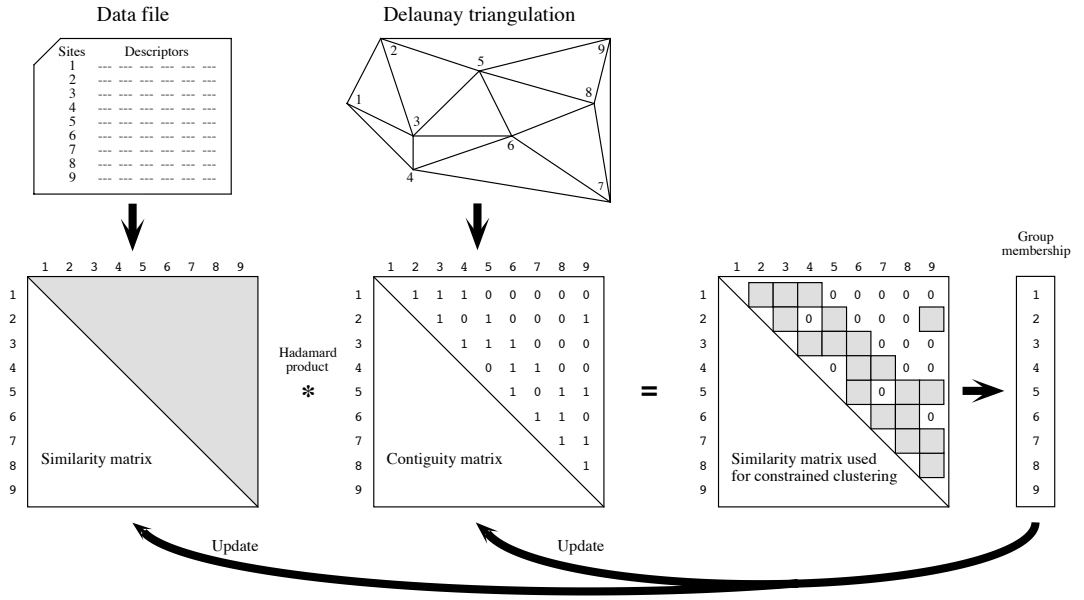


Figure 13.25 Summary of the spatially-constrained clustering procedure for methods included in the Lance and Williams general clustering model. The vector of group membership is represented on the right; at the start of the clustering process, each object is in a different group (numbers 1 to 9 in the example). Locations of the points are the same as in Fig. 13.22.

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. 8.12.

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.

8. Determine the most informative number of clusters, either by visual inspection of space-constrained clustering maps, or after calculating one of the indices mentioned at the end of Section 8.8 (available in R function ***clustIndex()*** of package CCLUST). Among those indices, the Calinski-Harabasz criterion was recommended by Gordon (1999) for constrained clustering. Pawitan & Huang (2003) proposed a permutation procedure to test the significance of successive partition levels in constrained clustering. Cross-validation seems another promising way of identifying the most informative partition in constrained clustering.

Ferligoj & Batagelj (1982) showed 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) can also be constrained by the contiguity matrix shown in Fig. 13.25.

Space-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, administrative 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 environmental 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 8.12.2). 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 can 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. 13.22 and 13.23, 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 13.3.1 are available in the literature. In addition, space-constrained clustering can be used to detect discontinuities in spatial transects or time series, a topic that has been discussed in Section 12.6.

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

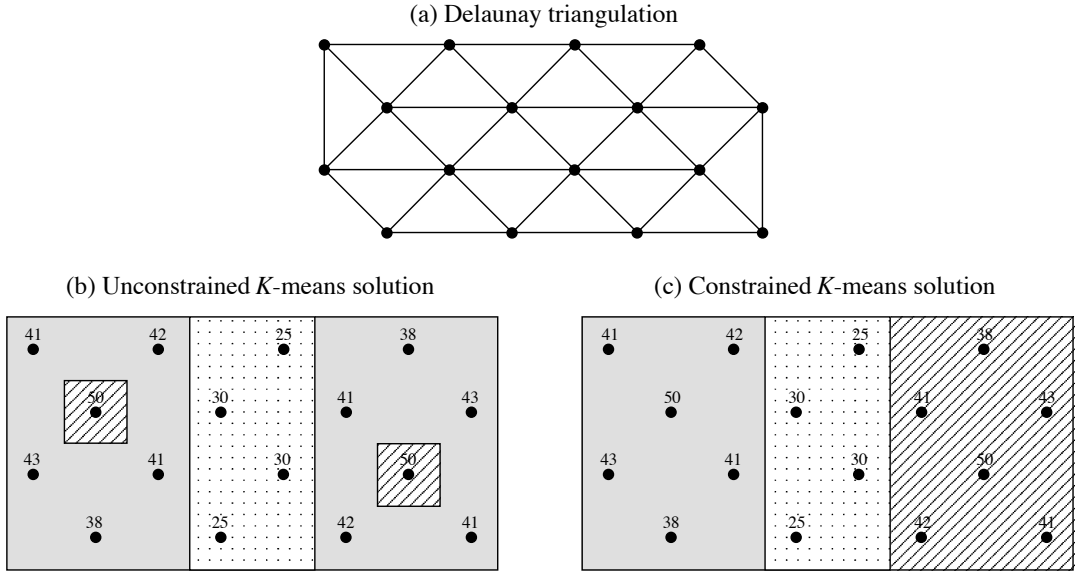


Figure 13.26 Numerical example showing the difference between (b) unconstrained and (c) constrained clustering solutions. (a) Delaunay triangulation with 35 edges, which were used as constraint in (c). The values of the artificial variable are given in panels (b) and (c); the three groups obtained by unconstrained and constrained *K*-means are identified by shadings.

implemented 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 were reviewed by Gordon (1996c, 1999) and algorithms were surveyed by Murtagh (1985). Formal aspects were discussed by Ferligoj & Batagelj (1982, 1983). Generalized forms of constrained clustering were 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. 13.26a). 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 $E_K^2 = 53$ (Fig. 13.26b). The constrained *K*-means solution, for three groups, has a value $E_K^2 = 188$ (Fig. 13.26c) 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 regions with high values plus a region with lower values forming a valley between the peaks.

Space-constrained clustering has been applied to a variety of ecological situations. Applications to two-dimensional map data are found in Legendre & Legendre (1984c), Legendre & Fortin (1989), Legendre *et al*. (1989), Lapointe & Legendre (1994), and Fortin & Dale (2005, their Section 4.1.2). Two examples of application of space-constrained clustering to community composition data surveyed on a geographic surface and along a transect, respectively, are available in the documentation file of function ***constrained.clust()*** of the R package CONST.CLUST; see Section 13.6. Users are invited to run these examples. A space-constrained clustering map of the data of Table 13.2 (bacterial data from the Thau coastal lagoon, France) is shown in Fig. 13.28b. Applications to transect and stratigraphic data (sediment cores) were mentioned in Subsection 12.6.5 and 12.8.

3 — Ecological boundaries

Detection of boundaries is the complementary problem to the detection of homogeneous regions of space. Boundaries appear on maps as a by-product of constrained clustering, for example. 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. Other methods have been developed that focus on boundary elements; these methods do not aim at completely isolating regions of space.

For univariate or multivariate transect data, boundaries can be detected using the methods described in Section 12.6. Detection of boundaries of various sorts on maps is more complex. This is a well-studied topic in the field of image analysis; it has been reviewed by Davis (1975), Peli & Malah (1982) and Huang & Tseng (1988); see also Hobbs & Mooney (1990). The present section briefly summarizes the efforts made to detect boundaries in ecological data sets, using a technique called *wombling*, and to statistically assess their significance. Readers are referred to Section 4.2 (*Boundary delineation*) of the book of Fortin & Dale (2005) for details; several examples are also presented in that book.

Wombling

Wombling is a technique for detecting zones of rapid spatial change in a set of regionalized variables. It was developed by Womble (1951) and Barbujani *et al*. (1989) for gene frequencies and morphological measurements, and refined by Fortin and co-authors (Oden *et al*., 1993; Fortin, 1994, 1997; Fortin & Drapeau, 1995; Fortin *et al*., 1996) with emphasis on ecological data. The original form of wombling (*lattice wombling*) could only be applied to quantitative variables observed at sites forming a regular, rectangular grid of points. Recent developments include *categorical wombling* for qualitative variables (Oden *et al*., 1993) and *triangulation wombling* for sites linked by a Delaunay triangulation which do not necessarily correspond to a regular sampling grid (Fortin, 1994). The latter is a frequent situation in ecology.

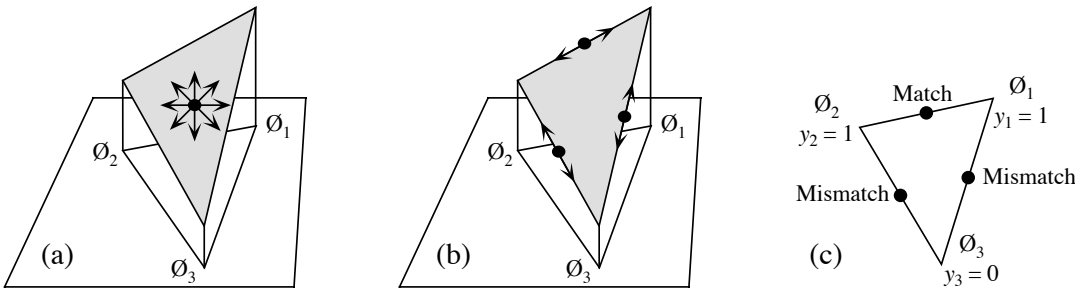


Figure 13.27 Examples of initial calculations on spatial elements (triplets of points) for three different boundary detection methods. Consider three points $Ø_i$, with coordinates (X_i, Y_i), forming a Delaunay triangle. (a, b) Variable y, measured at sites $Ø_i$, is quantitative (values shown as heights). (a) Method 1: find the direction of maximum slope of the triangle; allocate this slope value to the triangle centroid (dot). (b) Method 2: compute slopes along the edges connecting adjacent sites; allocate the values to the edge mid-points (dots). (c) Method 3: for a qualitative variable y (with 2 states in this example), adjacent sites are compared in terms of matches (0-0 or 1-1) or mismatches (0-1); allocate the matches and mismatches to the edge mid-points (dots).

A boundary is delineated on a map by linking adjacent points where the variable shows high rates of change (Fortin, 1994). Triangulation wombling (Fig. 13.27a) proceeds as follows:

- Link the observed sites by a Delaunay triangulation (Subsection 13.3.1).

- Consider a quantitative variable measured at three sites $Ø_i$ forming a Delaunay triangle. Each site has geographic coordinates (X_i, Y_i) and an observed value y_i. The plane to be fitted to these points is a linear function $y = f(X, Y) = b_0 + b_1X + b_2Y$ whose parameters can be computed by matrix inversion (Section 2.8):

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1 & X_1 & Y_1 \\ 1 & X_2 & Y_2 \\ 1 & X_3 & Y_3 \end{bmatrix}^{-1} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

- Find the direction of maximum slope of the triangle. The slope varies with the direction considered (arrows in Fig. 13.27a). Using the b coefficients calculated above, the *maximum slope* of the triangle is:

$$m = \sqrt{\left[\frac{\partial f(X, Y)}{\partial X}\right]^2 + \left[\frac{\partial f(X, Y)}{\partial Y}\right]^2} = \sqrt{b_1^2 + b_2^2} \qquad \textbf{(13.25)}$$

and the *angle* from the X coordinate axis is given by $\tan^{-1}(b_2/b_1)$. Note that angles are in radians in R. Allocate this value of slope (m) to the centroid of the triangle, which is the point with coordinates:

$$[X, Y]_{centroid} = \left[\frac{X_1 + X_2 + X_3}{3}, \frac{Y_1 + Y_2 + Y_3}{3}\right] \quad \textbf{(13.26)}$$

• If several variables are considered (i.e. several species), calculate the mean slope ($\bar{m}$) of the variables at the centroid of each Delaunay triangle.

• Create an ordered list of the slope values. Starting at the top of the list (highest slopes), mark the corresponding triangle centroids on the map; they become *boundary elements*. Going down the list, mark a predetermined proportion of the slopes (e.g. 10%), or go down to a preselected value of slope. Other strategies are possible, e.g. going down the list to the value of the mean plus one or two standard deviations.

Boundary

• A *boundary* is delineated by linking contiguous boundary elements. A single boundary element unlinked to other elements may be seen as a small boundary.

An alternative would be to compute the slopes of the edges between adjacent sites (Fig. 13.27b). For univariate data, the rate of change would simply be the absolute value of the difference between values at sites $Ø_h$ and $Ø_i$: $|y_h - y_i|$. For multivariate data, any of the distance functions of Chapter 7 could be used. The disadvantage of this method is that slopes calculated along the edges of the Delaunay triangle do not have the same value as the *maximum* slope of the triangle, computed by eq. 13.25. To alleviate this problem, Dufrêne & Legendre (1991) calculated multivariate distances in four directions between pixels of a map; for each pixel, they used the largest of the distances to delineate boundaries.

Computation of statistics along the edges between adjacent sites is the option used in categorical wombling, which is appropriate for species presence-absence data. The basic statistic is to record a match or a mismatch between adjacent observed sites (Fig. 13.27c). For multivariate qualitative data, one may count both the positive and negative matches and embed this number into one of the symmetrical binary coefficients of Subsection 7.3.1 (e.g. the simple matching coefficient); for species presence-absence data, one may count the positive matches only and embed this number into one of the asymmetrical binary coefficients of Subsection 7.3.2 (e.g. the Jaccard coefficient).

Tests of significance, based on permutations (Subsection 1.2.2), have been proposed by Fortin and co-authors (Oden *et al*., 1993; Fortin, 1994, 1997; Fortin & Drapeau, 1995; Fortin *et al*., 1996) to answer the following questions:

• Are the boundaries found by this analysis similar to random boundaries in terms of the number of separate boundaries, their maximum or mean lengths, or other boundary or graph-theoretic statistics?

• Are the boundaries found by wombling the same as borders stated by hypothesis, or found by clustering methods, or obtained using different data for the same locations?

These papers also present applications of the method to real and simulated data. A computer program for wombling is commercially available (BOUNDARYSEER, Table 13.3). An R package is also available (Section 13.6).

4 — Dispersal

Individuals, populations, and communities often cross ecological boundaries; such crossings occur on different time scales. The routes taken by species when they invade a territory after a perturbation event (long-term, e.g. glaciation; short term, e.g. pollution) is a question of interest in biogeographic analysis. Dispersal routes may be easier to identify if, as a first step in the analysis, one delineates regions that are largely homogeneous in species composition. Regions may be delimited using prior hypotheses, by unconstrained or constrained cluster analysis, or using boundary detection methods.

Legendre & Legendre (1984c) developed coefficients to measure the likelihood of species dispersal between geographically contiguous regions, for species presence-absence or abundance data. The assumptions of these coefficients are that the species arrived by migration, and the past dispersal has left traces in present-day distributions. For presence-absence data, adjacent regions $\mathbf{x}_1$ and $\mathbf{x}_2$ can be compared using the same quantities a, b, and c as in the similarity coefficients of Subsection 7.3.1 and 7.3.2: a is the number of species that two regions have in common; b is the number of species found in $\mathbf{x}_1$ but not in $\mathbf{x}_2$; c is the number of species found in $\mathbf{x}_2$ but not in $\mathbf{x}_1$. The *combination* of the following indications is evidence for species dispersal from region $\mathbf{x}_1$ to $\mathbf{x}_2$:

• the number of species common to the two zones is high, i.e. a is large;

• b is substantially larger than c. Conversely, c larger than b would support the hypothesis of dispersal from $\mathbf{x}_2$ to $\mathbf{x}_1$.

The basic form of the *coefficient of species dispersal direction* (*DD*) is thus $a(b - c)$. To make the values of the coefficient comparable for faunas with different richness, each term is standardized by dividing it by the richness of the fauna or flora of the two regions combined:

$$DD_1(\mathbf{x}_1 \rightarrow \mathbf{x}_2) = \frac{a}{(a+b+c)} \frac{(b-c)}{(a+b+c)} \qquad \textbf{(13.27)}$$

When this coefficient is positive, it measures the likelihood that species dispersed from $\mathbf{x}_1$ to $\mathbf{x}_2$. A negative sign indicates that, if dispersal occurred, species migrated from $\mathbf{x}_2$ to $\mathbf{x}_1$ instead.

The asymmetric portion of this coefficient may be tested for significance using a McNemar test. Under the null hypothesis of no asymmetry (H_0: $b = c$), the test statistic

$$X^2_P = \frac{(|b-c|-1)^2}{(b+c)} \qquad \textbf{(13.28)}$$

is distributed as χ^2 with one degree of freedom. The value –1 subtracted in the numerator is Edwards' (1948) correction for continuity; this is the correction used by function ***mcnemar.test()*** in R. The test may be one-tailed if one has specific hypotheses about the direction of dispersal; otherwise, a two-tailed test is used.

The log-linear form of the McNemar statistic is (Sokal & Rohlf, 1995):

$$X^2_W = 2\,(b\,\log_e b + c\,\log_e c - (b+c)\log_e[\,(b+c)/2]\,)/q \qquad \textbf{(13.29)}$$

where q is the Williams (1976) correction for continuity:

$$q = 1 + \frac{1}{2\,(b+c)}$$

Equation 13.29 provides a more powerful test than the classical McNemar equation (eq. 13.28). If any of the values b or c is 0, the corresponding term ($x \log_e x$) is 0 since $\lim_{x \to 0} (x \log_e x) = 0$ (Section 6.5).

The first part of the DD_1 equation is easily recognized as the Jaccard coefficient of similarity (eq. 7.10). One may prefer to give double weight to the number of common species, a, as in the coefficient of Sørensen (eq. 7.11):

$$DD_2(\mathbf{x}_1 \rightarrow \mathbf{x}_2) = \frac{2a}{(2a+b+c)}\,\frac{(b-c)}{(a+b+c)} \qquad \textbf{(13.30)}$$

Two other forms of the coefficient use species abundance data instead of presence-absence:

$$DD_3(\mathbf{x}_1 \rightarrow \mathbf{x}_2) = \frac{W\,(A-B)}{(A+B-W)^2} \qquad \textbf{(13.31)}$$

and

$$DD_4(\mathbf{x}_1 \rightarrow \mathbf{x}_2) = \frac{2W}{(A+B)}\,\frac{(A-B)}{(A+B-W)} \qquad \textbf{(13.32)}$$

where W, A, and B are as in the Steinhaus similarity coefficient (eq. 7.24). Coefficient DD_4 gives double weight to the abundances of the species in common and is thus the counterpart of DD_2, whereas DD_3 gives these species single weight, as in DD_1. These two coefficients take the following indications as evidence for dispersal from $\mathbf{x}_1$ to $\mathbf{x}_2$:

- the number of species common to the two zones and their abundances are high, i.e. W is large;

• *A* is substantially larger than *B*; *B* larger than *A* would produce a negative coefficient, indicating possible dispersal from $\mathbf{x}_2$ to $\mathbf{x}_1$.

Legendre & Legendre (1984c) used *DD* coefficients and tests of significance to identify plausible routes taken by freshwater fishes to reinvade the Québec peninsula after the last glaciation. Borcard *et al*. (1995) used the same method in a finer-scale study, showing possible patterns of migration of Oribatid mites between zones of an exploited peat bog in the Swiss Jura. At broader scale, Bachraty *et al*. (2009) used *DD* coefficients to identify possible faunal dispersal pathways between adjacent deep-sea hydrothermal provinces of the world ocean.

13.4 Unconstrained and constrained ordination maps

Subsection 13.3.2 has shown how maps for multivariate data can be produced by clustering methods; these maps display discontinuous zones. For continuous representation of quantitative variables, however, the techniques of Section 13.2 can only produce maps for single variables. The present section shows how continuously-varying maps can be obtained for multivariate data sets through ordination methods. The relationship between univariate or multivariate structure functions (Section 13.1) and maps has been stressed in the introductory paragraph of Section 13.2.

The simplest method consists in analysing a data table with one of the ordination methods of Chapter 9 and map the first, or the first few ordination axes. For example:

• Decompose the variation of a (sites × species) presence-absence or abundance table into successive ordination axes, using PCA, CA, PCoA, or nMDS (Chapter 9).

• Consider the ordination of sites along the first axis. This axis is a new, synthetic quantitative variable describing the variation among sites. Associate it to the (X, Y) geographic coordinates of the sites. Produce a map using one of the methods described in Section 13.2. An example of such a map is given in Fig. 9.15 for correspondence analysis axis I of a vegetation data table.

• Repeat the operation, producing maps for ordination axes II, III, etc. as long as interesting or significant spatial variation can be detected. Univariate correlograms of the successive ordination axes (Subsection 13.1.1), with tests of significance, may be used as criterion for deciding which of the ordination axes should be mapped.

Simple ordination analysis leaves it to chance to find spatially-structured components of variation. One may decide instead to look directly for such components, by forcing the analysis to bring out axes of variation that are related to the X and Y coordinates, or combinations of X and Y into a spatial polynomial equation. The spatial polynomial is constructed as in Subsection 13.2.1. (In Chapter 14, the X and Y coordinates will be used to derive spatial eigenfunctions that will replace the

Table 13.2 Data from Table 10.6. There are two bacterial response variables (Bna and Ma, forming matrix **Y**) and three environmental variables (NH_4, phaeopigments, and bacterial production, forming matrix **X**). Five spatial variables (X^2, X^3, X^2Y, XY^2, and Y^3, included in matrix **W**) were derived from the X and Y coordinates, reported in the table, obtained by PCA rotation of the geographic coordinates of Table 10.6. The variables are described in more detail in Numerical example 1 of Subsection 10.3.5.

Station No.	Bna y_1	Ma y_2	NH_4 x_1	Phaeo. *a* x_2	Prod. x_3	X after PCA rotation	Y after PCA rotation
1	4.615	10.003	0.307	0.184	0.274	–9.4173	–1.2516
2	5.226	9.999	0.207	0.212	0.213	–7.1865	–1.0985
3	5.081	9.636	0.140	0.229	0.134	–5.8174	–1.4528
4	5.278	8.331	1.371	0.287	0.177	–6.8322	0.2706
5	5.756	8.929	1.447	0.242	0.091	–4.6014	0.4238
6	5.328	8.839	0.668	0.531	0.272	–4.2471	1.7929
7	4.263	7.784	0.300	0.948	0.460	–1.8632	–0.2848
8	5.442	8.023	0.329	1.389	0.253	–0.4940	–0.6391
9	5.328	8.294	0.207	0.765	0.235	0.8751	–0.9934
10	4.663	7.883	0.223	0.737	0.362	–0.1398	0.7300
11	6.775	9.741	0.788	0.454	0.824	–1.1546	2.4534
12	5.442	8.657	1.112	0.395	0.419	0.2145	2.0992
13	5.421	8.117	1.273	0.247	0.398	4.9824	–2.0562
14	5.602	8.117	0.956	0.449	0.172	3.9676	–0.3328
15	5.442	8.487	0.708	0.457	0.141	3.4602	0.5289
16	5.303	7.955	0.637	0.386	0.360	6.3515	–2.4105
17	5.602	10.545	0.519	0.481	0.261	5.8441	–1.5488
18	5.505	9.687	0.247	0.468	0.450	4.8293	0.1746
19	6.019	8.700	1.664	0.321	0.287	4.6762	2.4054
20	5.464	10.240	0.182	0.380	0.510	6.5527	1.1894

spatial polynomial in the same type of analysis as described here.) Ordination analysis of a species data table, constrained to be related to a spatial polynomial, can be carried out by canonical analysis (Chapter 11). Canonical analysis then becomes an extension to multivariate data tables of the method of trend surface analysis. The method will be described with the help of a numerical example. Another example (vegetation data) is found in Legendre (1990).

Numerical example. Bacterial data from the Thau coastal lagoon are used again here (Tables 10.6 and 13.2). The response data include Bna and Ma bacteria in the present example. No significant linear trend was present in the response data. To facilitate mapping, the X and Y geographic coordinates of the sites were rotated by principal component analysis (PCA using the covariance matrix; scaling type 1 was used); Table 13.2 gives the rotated coordinates. A third-degree polynomial of these new X and Y coordinates was created (Subsection 13.2.1) and subjected to forward selection in order to select the spatial monomials that significantly contributed to the explanation of the Bna and Ma bacterial variables. The following five terms of

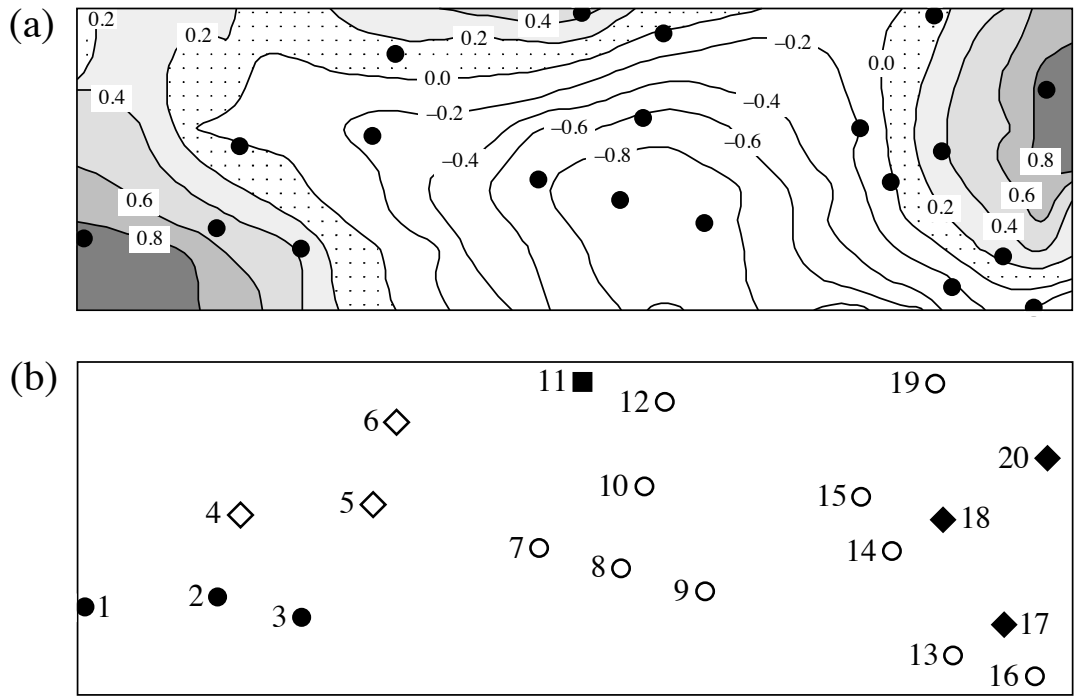


Figure 13.28 (a) Fitted site scores map of the first canonical axis of the bacterial variables in the Thau coastal lagoon constrained by a spatial polynomial. (b) Space-constrained clustering map of the bacterial data, 5 groups (symbols). Dots are the 20 sampling sites, with identification numbers in map b. The north-south direction is nearly parallel to the vertical axis of the maps; compare with the point positions in Fig. 13.15a.

the spatial polynomial were retained by the selection procedure: X^2, X^3, X^2Y, XY^2 and Y^3. Spatial eigenfunction analysis, described in Chapter 14, would not have produced a good spatial model with these data because the 20 sites are too irregularly spaced on the map (Fig. 13.28).

Redundancy analysis (RDA, Section 11.1) produced two canonical axes ($\lambda_1 = 0.622$, $\lambda_2 = 0.111$) because there were two dependent variables only (Bna, Ma) (see Table 11.1). The canonical relationship accounted for $R^2_a = 0.639$ of the variation in the bacterial data; it was globally significant ($p = 0.001$); so was the first canonical eigenvalue. 81% of the variance of Ma was expressed along axis I, but only 7% of the variance of Bna. The second canonical eigenvalue was not significant at $\alpha = 0.05$ ($p = 0.123$) although 42% of the variance of Bna was expressed on this axis. There are two non-canonical axes representing the non-spatially-structured variation of the response variables; they represent 16% and 10% of the variation of matrix **Y**, respectively. Canonical axis I differs from the first principal component of matrix **Y**: that axis would express the variation in the response variables (Bna, Ma) without the constraint of being a linear combination of the spatial monomials.

For axis I, the fitted site scores from matrix **Z** (eq. 11.18) were mapped (Fig. 13.28a) by kriging (Subsection 13.2.2) using program OKB2D of the GSLIB library (Deutsch & Journel, 1992); an all-directional spherical variogram models was fitted to the empirical variograms prior to kriging (Subsection 13.1.3). The trend surface equation that produced the fitted site scores for the 20 sites is:

$$\widehat{\text{Axis I}} = 1.0526X^2 + 1.1881X^3 - 0.5225X^2Y - 0.7674XY^2 + 0.7167Y^3$$

The spatial monomials were standardized before computing this equation.

Interpretation of the map is rather simple in this example: examination of Table 13.2 shows that the sites with the highest scores along canonical axis I (i.e. sites 1-3, 11, 17, 18 and 20, located in grey areas of Fig. 13.28a) possessed the highest concentrations of aerobic heterotrophic bacteria growing on marine agar (variable Ma), which was the variable dominating axis I. These sites also formed three separate and clearly identified groups in the space-constrained clustering map (Fig. 13.28b) produced using function ***constrained.clust()*** using the Ward.D2 algorithm (Subsection 8.5.8); see Section 13.6. The clustering level, 5 groups, displayed in the map was selected by cross-validation.

Thioulouse *et al*. (1995) proposed a different approach to mapping, which combines connection networks, decomposition of the variation into local and global components, eigenvalue decomposition, and mapping. The neighbouring relationships among sites are represented by a connection network (e.g. Delaunay triangulation for a homogeneous two-dimensional sampling area, or neighbouring relationships for sites along a river network) which is translated into a contiguity matrix **M** (Fig. 13.25). **M** is standardized to **P** by division by the total number of pairs of neighbours. A diagonal matrix **D** describes the degree of connectedness of the sites. Using matrices **P** and **D**, the authors define principal component and correspondence analysis for the total, local, and global components of variation; each fraction is decomposed into orthogonal axes, which may be mapped to facilitate interpretation. Their paper presents applications to simulated and real ecological data (bird survey).

13.5 Spatial modelling through canonical analysis

The significance of spatial heterogeneity for the functioning of ecosystems was discussed in Section 1.1. Models of ecosystem processes may fail to correctly model the spatial variation of communities (i.e. beta diversity, Subsection 6.5.3) if they do not include the spatial organization of the populations and communities among the models' predictor variables. This can be achieved by explicitly incorporating spatial predictors in ecological models of community composition data (or other response data matrices) using canonical analysis. The method consists in modelling the variation of the variables of interest across the study area as a linear combination of the environmental variables *and* some function of the geographic coordinates of the sites.

In this approach, one is interested in explicitly identifying the effect of spatial structures, singling it out from other environmental effects, through methods discussed in previous chapters and sections: partial regression analysis (univariate, Subsection 10.3.5) or partial canonical analysis (multivariate, Section 11.1.6) on the one hand; trend surface analysis (univariate, Subsection 13.2.1) and spatially constrained ordination maps (multivariate, Section 13.4) on the other hand. In its basic form, the analysis considers three data sets: **Y** contains the response variables; **X** is the set of explanatory environmental variables; **W** is the set of explanatory spatial variables. In the present section, which introduces the method, **W** contains a polynomial of the geographic coordinates of the sites. In Chapter 14, spatial eigenfunctions will replace the spatial polynomial. Variation partitioning (univariate, Section 10.3.5, or multivariate, Section 11.1.11) produces synthetic presentations of the results. Variation partitioning can actually accommodate more than one environmental and one spatial matrix — up to four with the presently available version of function ***varpart()*** in R.

There are two motivations, described in more detail in Section 14.1.4, to carry out spatial modelling of response matrix **Y** by variation partitioning. The first focuses on fraction [a] of Fig. 10.10, in cases where one wishes to control for spatial correlation in the analysis of species-environment relationships. The second corresponds to situations where both the spatial and non-spatial structures of the explanatory variables **X** are of interest to explain the variation of **Y**, in which case fractions [a], [b] and [c] can all be interpreted. In this approach, any structure identified in the response data is considered to indicate the presence of some process generating it. Mapping fraction [c] of the variation may help generate hypotheses about the spatial process or processes responsible for the observed residual spatial pattern.

The variation partitioning approach described in the present section is essentially correlative. It differs from the analysis of variance, which estimates the variation associated with well-defined effects in structured sampling or manipulative (i.e. controlled) experiments. In the initial stages of ecological research, correlative methods are routinely used to sort out hypotheses centring on broad correlative patterns among groups of variables, before specific hypotheses can be experimentally tested. In particular, the analysis presented in this section allows researchers to consider different groups of explanatory variables (environmental, spatial, or temporal) and examine their capacity to explain patterns in the multivariate response variables (species or others) that are of interest in a study; it further allows one to measure the degree of overlap that exists among these groups of explanatory variables with regard to that capacity (Anderson & Gribble, 1998). The correlations brought out by the analyses are only interpretable insofar as hypotheses can be formulated about the processes that may have generated the observed patterns. This approach is related to regression (Section 10.3) and path analysis (Section 10.4), in which a large number of plausible relationships may be hypothesized and sorted out by statistical analysis. The method is illustrated by a numerical example.

Numerical example. The data of Table 13.2 (Thau coastal lagoon) are reanalysed here. In the numerical example of Section 13.4, the variable selection procedure retained the following terms of the spatial polynomial: X^2, X^3, X^2Y, XY^2, and Y^3; the same terms are used in the present example. The following variation partitioning table (R^2 and R^2_a columns) was obtained using function ***varpart()*** of VEGAN. The values in the three columns to the right were obtained by RDA and partial RDA.

Fractions of variation	Proportion of var. of Y (R^2)	Adjusted R^2	Probability (999 perm.)	Canonical λ_1	Probability (999 perm.)
[a + b]	0.450	0.347	0.005*	0.359	0.025*
[b + c]	0.734	0.639	0.001*	0.622	0.001*
[a + b + c]	0.784	0.628	0.001*	0.632	0.001*
[a]		–0.011≈0	0.549	0.042	0.561
[b]		0.358	-----	-----	-----
[c]		0.281	0.011*	0.304	0.004*
Residuals = [d]		0.372	-----	-----	-----
[a + b + c + d]		1.0000			

In the table, significant fractions at level $\alpha = 0.05$ are identified by asterisks. Eigenvalues of the first canonical axis (canonical λ_1) are reported as fractions of the total variance of **Y**. None of the second canonical axes were significant; since each analysis only produced two canonical eigenvalues in this example, the portions of variation corresponding to λ_2 are the differences between columns 2 and 5 from the left. Fraction [b] is not an independently-calculated component of the variation; hence, it cannot be tested for significance nor decomposed into canonical axes (see Méot *et al*., 1998, for alternative solutions).

Variation partitioning decomposed the total explained variation [a + b + c], expressed as R^2_a, into a significant environmental component [a + b] and a significant component [c] that estimated the spatially-structured variation of **Y** not explained by the environmental variables. The table shows that [a + b], which is the variation of **Y** explained by the environmental variables, was entirely spatially structured; [a] is a small negative value that must be interpreted as a zero. The partitioning results are represented by a Venn diagram in Fig. 13.29.

Figure 13.30 shows maps of the fitted site scores from matrix **Z** (eq. 11.18) of the first canonical axis of fraction [a + b + c] and of its two components, [a + b] and [c]. These maps were obtained by kriging (Subsection 13.2.2) using program OKB2D of the GSLIB library (Deutsch & Journel, 1992); all-directional spherical variogram models were fitted to the empirical variograms prior to kriging (Subsection 13.1.3). While the adjusted proportions of variation of [a + b] and [c] add up to that of [a + b + c] (0.347 + 0.280 = 0.627), this is not the case for the proportions of variation represented by the first canonical axes: $\lambda_{1[a+b]} + \lambda_{1[c]} \neq \lambda_{1[a+b+c]}$. This is because the partition of fraction [a + b + c] into canonical axes is done independently of the partitions of [a + b] or [c]. As a consequence, maps of a given axis of variation (e.g. axis I of the various fractions, mapped in Fig. 13.30) do not exactly add up with this method; they only add up approximately.

The adjusted fraction [b + c] (0.639 of the bacterial data variation) is the one extracted by canonical analysis in the numerical example of Section 13.4 (same data); Fig. 13.28a maps this

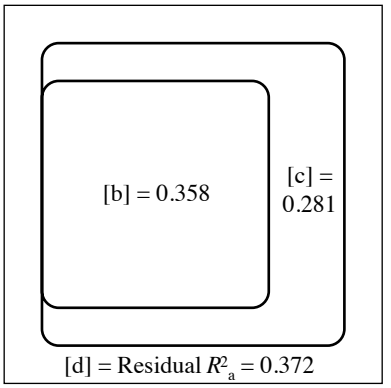


Figure 13.29 Venn diagram illustrating the variation partitioning results for the numerical example. The rounded rectangle surface areas approximate the relative fraction sizes with respect to the size of the outer rectangle, which represents the total variation in the response data. The fractions are identified by letters [b] to [d]; [a] is not shown because it is approximately zero. The values next to the identifiers are adjusted R^2 (R^2_a).

fraction of the variation. In this example, the map of axis I of fraction [b + c] (Fig. 13.28a) is very similar to the map of axis I of [a + b + c] (Fig. 13.30) because [a] is close to zero.

The maps of axis I of fraction [a + b + c] (63% of the variation in the bacterial variables) and [a + b] (36%) are quite similar, whereas the map of axis I of fraction [c] (33% of the variation) is quite different. The trend surface equation that produced the fitted site scores for the 20 sites is:

$$\widehat{\text{Axis I of [c]}} = 1.8017X^2 + 2.2817X^3 - 1.0809X^2Y - 1.3064XY^2 + 1.5563Y^3$$

In this equation, the spatial variables are residuals of the standardized terms of the spatial polynomial after controlling for the effect of the three environmental variables. Examination of the map of fraction [c] suggests a hypothesis for the origin of this fraction of variation, i.e. a marine influence, which had not been included among the explanatory variables in the analysis. Indeed, the negative values on the map form a plume originating at the connections of the Thau lagoon with the sea and extending westwards. To "explain away" fraction [c], i.e. to make it become non-significant, another analysis could be conducted that would include variables quantifying the marine influence on the stations of the lagoon among the environmental variables. Such variables could, for example, be obtained from a hydrodynamic model of the lagoon.

Interpretation of the fractions is described in Subsection 14.1.4. Applications of this method cover a wide range of ecological problems. Here is a selected list of fields and papers: palaeoecology (Zeeb *et al*., 1994; see also Ecological application 10.3b), stream monitoring (Passy, 2007), vegetation (Heikkinen & Birks, 1996; Bjorholm *et al*., 2005), periphyton (Cattaneo *et al*., 1993), protozoa (Buttler *et al*., 1996), zooplankton (Pinel-Alloul, 1995), aquatic macroinvertebrates (Pinel-Alloul *et al*., 1996), fish (Rodríguez & Magnan, 1995), birds (Bersier & Meyer, 1994; Gordo *et al*., 2007), and mammal conservation (Burbidge *et al*., 2008).

Variation partitioning has been applied to more than two explanatory data sets. (1) Pinel-Alloul *et al*. (1995) tested the hypothesis that biotic and abiotic factors, as well as spatial structuring, explained together the broad-scale spatial heterogeneity of zooplankton assemblages among lakes. The explanatory variables comprised abiotic

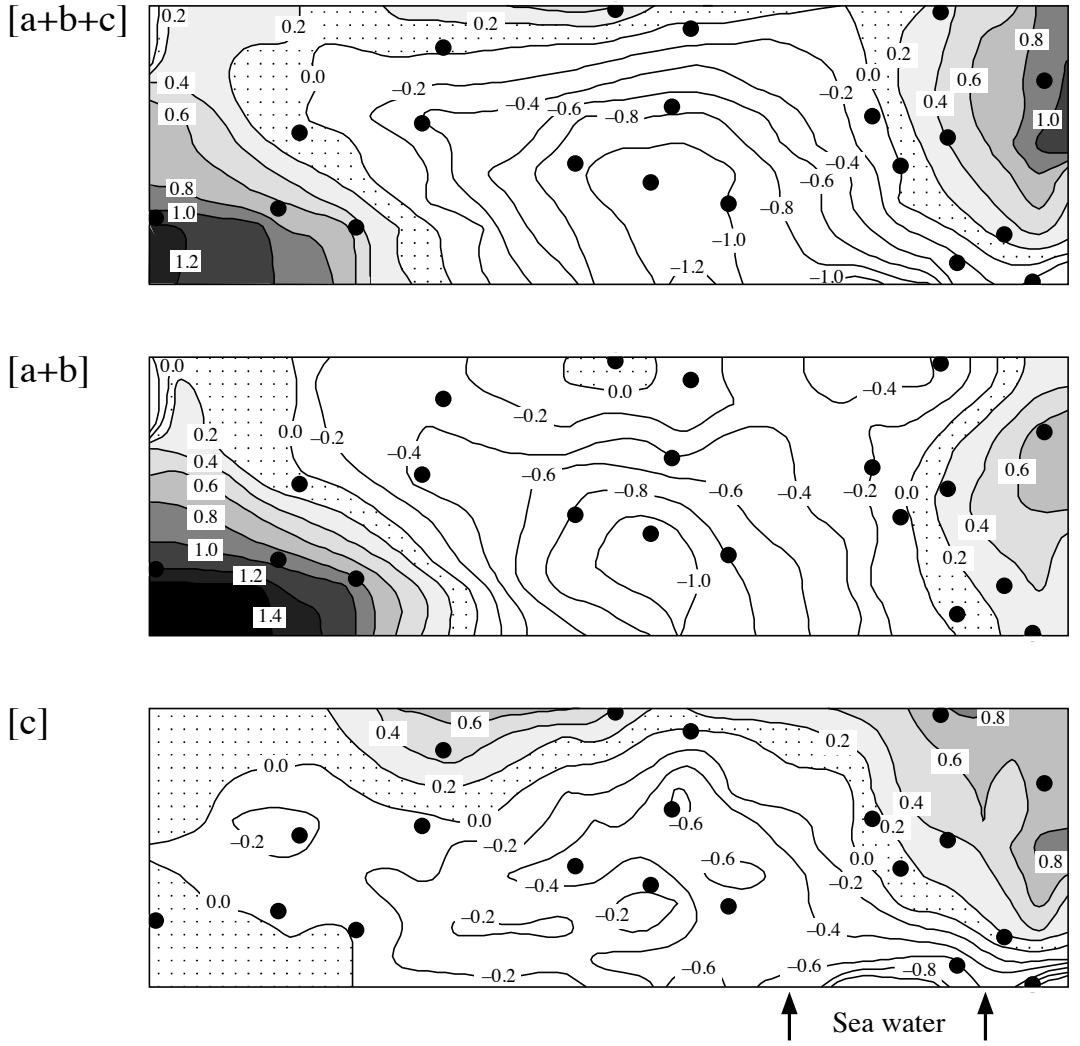


Figure 13.30 Bacterial variables: map of the fitted site scores of the first canonical axes of three fractions of the variation: top [a + b + c], middle [a + b], bottom [c]. Dots represent the 20 sampling sites. North is nearly parallel to the vertical axis of the maps. Compare with Fig. 13.28a, which represents fraction [b + c]. Arrows at the base of map [c], labelled "sea water", indicate the positions of connections of the Thau coastal lagoon with the adjacent Mediterranean Sea.

(physics and chemistry, morphometry) and biotic factors (phytoplankton and fish assemblages); these factors were analysed separately and together, using four approaches described in the paper. (2) Quinghong & Bråkenhielm (1995) explained the spatial patterns of epiphytic green algae and lichens using climatic, pollution, and

Table 13.3 Computer programs implementing the methods of surface pattern analysis discussed in Chapter 13. The list is not exhaustive.

BOUNDARYSEER Detection of boundaries (wombling) and space-constrained clustering

BioMedware Inc., 3526 W. Liberty, Suite 100, Ann Arbor, Michigan 48103, USA. http://www.biomedware.com/

GEOEAS Variogram, kriging; contour mapping

Developed by US Environmental Protection Agency. Available from ACOGS, P.O. Box 44247, Tucson, Arizona 85733-4247, USA. ftp://math.arizona.edu/incoming/unix.geoeas/

GSLIB Geostatistical software library

Geostatistical methods described in the Deutsch & Journel (1992) book. http://www.gslib.com/

ISATIS Variogram, kriging; contour mapping.

Geovariances, 49bis avenue Franklin-Roosevelt, F-77212 Avon Cedex, France. http://www.geovariances.com/en/software-ru2

SAAP Spatial autocorrelograms (Moran's *I* and Geary's *c*)

Written by D. Wartenberg, Department of Environmental and Community Medicine, Robert Wood Johnson Medical School, Piscataway, New Jersey, USA. Distributed by Exter Software: http://www.exetersoftware.com

SPACESTAT Directional and omnidirectional spatial variograms; several types of kriging; interpolated maps; global and local spatial correlation statistics; geographically weighted regression

BioMedware Inc., 3526 W. Liberty, Suite 100, Ann Arbor, Michigan 48103, USA; http://www.biomedware.com.

SURFER Kriging; other interpolation methods; contour mapping

RockWare Inc., 2221 East St. #101, Golden, Colorado 80401, USA. http://www.rockware.com/

geographic variables. They showed how to isolate the seven components of variation resulting from crossing three sets of explanatory variables. (3) Anderson & Gribble (1998) applied the variation partitioning method to three matrices of explanatory variables representing the environmental, spatial and temporal components, respectively. Using this approach, they were able to resolve the confounding of space and time that is often encountered when sampling is conducted over a long time period because of the large size of the surveyed area.

13.6 Software

Most of the methods described in this chapter cannot be implemented using the major statistical packages. Table 13.3 lists some of the software available commercially or from researchers. The list is not exhaustive.

The R language offers functions for all methods described in this chapter:

1. Structure functions. — Function ***correlog()*** of NCF: Moran's I spatial correlograms computed as described in Subsection 13.1.1 (two-tailed tests). ***sp.correlogram()*** of SPDEP: spatial correlograms based on Moran's I, Geary's c, or the spatial correlation function (eq. 13.18), based on a list of connection edges (two-tailed tests).

In package NCF, functions ***Sncf()***, ***Sncf.srf()***, ***Sncf2D()***, ***spline.correlog()*** and ***spline.correlog2D()*** estimate spline correlograms in various ways for univariate or multivariate spatial data, including directional correlograms for anisotropic data. Confidence envelopes are computed through bootstrapping. Function ***lisa()*** computes LISA statistics (local Moran) for single response variables. A randomization test is used to test for significance of the LISA statistics.

Functions ***variog()*** of GEOR, ***Variogram()*** of NLME, ***vario()*** of PASTECS, and ***est.variogram()*** of SGEOSTAT compute empirical variograms. A multivariate variogram of $\mathbf{Y}$ with permutation test is computed by ***mso()*** of VEGAN, after a PCA of $\mathbf{Y}$ by ***rda()***. A multivariate correlogram is computed by ***mantel.correlog()*** of VEGAN.

2. Maps. — Function ***lm()*** of STATS can be used to compute trend-surface analysis of univariate response data. ***eyefit()*** of GEOR adjusts variogram models to data. ***krige.conv()*** in GEOR carries out conventional kriging. Among the functions available with the Borcard *et al*. (2011) book, ***sr.value()*** draws bubble maps; values of a variable are represented either by circles of different sizes or by circles with shades of grey. Function ***s.value()*** of ADE4 also draws bubble maps using squares instead of circles.

3. Patches and boundaries. — Function ***constrained.clust()*** of package CONST.CLUST* carries out constrained hierarchical clustering along a time or spatial series or on a geographic surface, with cross-validation of the results. Wombling analysis for the estimation of boundaries (Subsection 13.3.3) is computed by package WOMBSOFT. Coefficients of dispersal direction, described in Subsection 13.3.4, are available in function ***bgdispersal()*** of VEGAN.

4. Unconstrained and constrained ordination maps. — R functions are listed in Section 9.5 for unconstrained and in Section 11.7 for constrained ordination.

5. Spatial modelling through canonical analysis. — Functions for canonical analysis and variation partitioning are described in Section 11.7

6. Miscellaneous methods. — Function ***geoXY()*** in SODA transforms latitude-longitude (LatLon) data to flat Cartesian coordinates. ***lm()*** in STATS can be used to carry out spatial detrending of univariate or multivariate data. ***poly()*** in STATS computes ordinary or orthogonal polynomials.

* Available on the Web page http://numericalecology.com/rcode.